Sensitivity Analysis of Model Output

Proceedings of the 4th International Conference on Sensitivity Analysis of Model Output (SAMO 2004)

Santa Fe, New Mexico; March 8-11, 2004

Kenneth M. Hanson and François M. Hemez, editors

Published March 2005

Forward

The 4th International Conference on Sensitivity Analysis of Model Output (SAMO 2004) was held in Santa Fe, New Mexico, USA, March 8-11, 2004. This was the first time that the SAMO meeting had been held in the United States of America. The meeting took place in the Inn at Loretto, which proved to be a wonderful venue for holding a relaxed meeting.

The first International Symposium on Sensitivity Analysis of Model Output (SAMO) was held in 1995 in Belgirate, Italy, under the auspices of the Joint Research Centre (JRC) of the European Commission. It was organized by Andrea Saltelli, who brought together a small, informal group of researchers dedicated to the advancement of methodologies for understanding uncertainty in mathematical simulations (models) of physical systems. The same group organized a second meeting in 1998 at the University Ca' Foscari in Venice. The third SAMO meeting was held in 2001 in Madrid, Spain. It was organized by a group from Spain's Department of Environmental Impact of Energy (DIAE) of CIEMAT, together with two universities, Politechnical University of Madrid (UPM) and the University Rey Juan Carlos (URJC).

The theme of the SAMO series has been the study of the variation in the output of a model caused by variations in its inputs variables, parameters, and factors related to numerical solution methods. Especially highlighted at this meeting was the importance of the quantification of the total uncertainty in model prediction. The following techniques were discussed:

- Innovative methods of sensitivity and uncertainty-importance analysis
- Designs and sampling plans for computer experiments
- Model calibration
- Model evaluation and validation
- Reliability analysis and robustness analysis
- Probabilistic and non-probabilistic analysis of uncertainty and sensitivity
- Modeling knowledge and judgment
- Decision-making under uncertainty

The application areas included economics, engineering, environment, nuclear safety, and physics.

The Organizing Committee was comprised of Scott Doebling, Ken Hanson, François Hemez, Rudy Henninger, Michael McKay, and Kathie Womack, all from the Los Alamos National Laboratory. Kathie Womack's diligent attention to the organizational details contributed greatly to the smooth functioning of the meeting. Vivian Romero of the Statistical Sciences group developed and maintained the SAMO 2004 web site.

To compose the final program, the Organization Committee relied heavily on the recommendations from the Scientific Committee, which included the following people:

- James Cavendish, Research and Development Center, General Motors Corporation, U.S.A.
- Kenneth M. Hanson (committee chair), Continuum Dynamics, Los Alamos National Laboratory, U.S.A.
- Toshimitsu Homma, Department of Reactor Safety Research, Japan Atomic Energy Research Institute, Japan
- Michiel Jansen, Biometris, Wageningen University and Research Centre, The Netherlands
- Hyoung-Man Kim, Structural Dynamics, The Boeing Company, U.S.A.
- Jack P.C. Kleijnen, Department of Information Management, Center for Economic Research, Tilburg University, The Netherlands
- Pedro Padro Herrero, Department of Environmental Impact of Energy, Research Centre for Energy, Environment and Technology, Spain
- Stefano Tarantola, Institute for the Protection and Security of the Citizen, Joint Research Centre of the European Commission, Italy

This committee was tasked with reviewing the 78 abstracts that were submitted for consideration.

The final program consisted of 35 oral presentations, including eight invited tutorials. In addition, 24 posters were presented. By avoiding parallel tracks, it was possible for everyone to hear everything and to provide ample time for questions and comments throughout the four-day conference. There were numerous lively discussions. Additionally, an hour-long open discussion was held on Model Calibration and Validation, with Michiel Jansen, Michael McKay, Anthony O'Hagan, and Timothy Trucano as panelists and Ken Hanson as moderator.

Tutorials were presented on the basic elements of sensitivity analysis by the well-known experts Max Morris, Andrea Saltelli, Michael McKay, Anthony O'Hagan, Katherine Campbell, Timothy Trucano, Roger Cooke, and Jon Helton.

SAMO 2004 drew to Santa Fe 106 registrants from 13 countries. Attendees had ample opportunity for informal technical discussions. Evenings were spent socializing and exploring the unique attractions of Santa Fe. The banquet was held at the Inn at Loretto. Guest speaker Andrew White, of the Los Alamos Computing Project, spoke about the development of computing at LANL in a talk entitled “A History of Predicting the Future.”

The conference was generously supported by the Los Alamos National Laboratory (LANL). LANL, which is operated by the University of California for the National Nuclear Security Administration of the U. S. Department of Energy. LANL’s contribution facilitated the participation of numerous internationally recognized leaders in the fields of sensitivity analysis and simulation science, many of whom presented the invited tutorials. LANL’s sponsorship made it possible for many graduate students to attend. The specific organizations at LANL that provided substantial financial and logistic support include the Weapons Response group (ESA-WR), the Continuum Dynamics group (CCS-2), the Statistical

Sciences group (D-1), and the Research Library. Additional sponsors include the Joint Research Centre of the European Union, the American Statistical Association, and the Society for Industrial and Applied Mathematics.

The SAMO 2004 proceedings include 50 contributed papers. This collection is archived on the web by the LANL Research Library at http://library.lanl.gov/. Mark Martinez, of the Library-Without-Walls team, helped create our web pages, and maintain the archive. The conference web pages, which include photos taken during the conference, are maintained by the Statistical Sciences Group at http://www.stat.lanl.gov/SAMO2004/ .

We suggest that citations to papers from these proceedings include the following information: Author list, "Paper Title," *Sensitivity Analysis of Model Output*, K. M. Hanson and F. M. Hemez, eds., pp. page numbers (Los Alamos National Laboratory, Los Alamos, 2005) (http://library.lanl.gov/ccw/samo2004/).

Kenneth M. Hanson and François M. Hemez, editors
Los Alamos, New Mexico
March 2005

Table of Contents

Forward

Page

Input Screening: Finding the Important Inputs on a Budget (Tutorial)
Max D. Morris 1

The Evaluation of Combustion Mechanisms Using Local and Global Sensitivity and Uncertainty Methods
Alison S. Tomlin 7

Application of MCMC - Global Sensitivity Analysis for Model Calibration to Urban Runoff Modeling
Assem Kanso; B. Tassin, G. Chebbo 17

Global Sensitivity Analysis: An Introduction (Tutorial)
Andrea Saltelli 27

Sensitivity Analysis of the e-Business Readiness Composite Indicator
S. Tarantola, M. Nardo, A. Saltelli, I. Kioutsioukis, R. Liska 44

Sensitivity Indices for Imprecise Probability Distributions
Jim W. Hall 54

Accelerated Estimation of Sensitivity Indices Using State Dependent Parameter Models
Marco Ratto, Stefano Tarantola, Andrea Saltelli and Peter C. Young 61

Sensitivity Analysis of Monte Carlo Estimates from Computer Models in the Presence of Epistemic and Aleatory Uncertainties
Bernard Krzykacz-Hausmann 71

Sensitivity Analysis When Model Outputs Are Functions (Tutorial)
Michael D. McKay, Katherine S. Campbell, Brian J. Williams 81

Decision-theoretic Sensitivity Analysis Using Value of Information
Jeremy Oakley 90

A Second-order Differential Importance Measure for Reliability and Risk Applications
Enrico Zio, Luca Podofillini 96

A Response-Modeling Alternative to Surrogate Models for Support in Computational Analyses
Brian Rutherford 106

Local Analysis of Parameter Covariances Resulting From the Calibration of an Over-parameterized Water Quality Model
Ulrich Callies, M. Scharfe 116

Inference About the Plastic Behavior of Materials From Experimental Data
Kenneth M. Hanson 126

Comments on GLUE
Renata Romanowicz, Keith Beven 137

A Bayesian Analysis of Complex Dynamic Computer Models
Stefano Conti, Clive W. Anderson, Marc C. Kennedy, Anthony O'Hagan 147

Validation and Error Estimation of Computational Models
Ramesh Rebba, Sankaran Mahadevan, Shuping Huang 157

Probabilistic Inversion for Chicken Processing Lines (Tutorial)
Roger M. Cooke, Maarten Nauta, Arie H. Havelaar, Ine van der Fels 170

The Good, the Bad, and the Ugly of Predictive Science
François M. Hemez; Yakov Ben-Haim 181

Our Calibrated Model Has no Predictive Value: An Example from the Petroleum Industry
Jonathan N. Carter, P. J. Ballester, Z. Tavassoli, P. R. King 194

Towards Guidance in Assessing and Communicating Uncertainties
Peter H.M. Janssen, Arthur C. Petersen; Jeroen van der Sluijs;James S. Risbey; Jerome R. Ravetz 201

Probability Bounds Analysis is a Global Sensitivity Analysis
Scott Ferson, J. Hajagos, W.T. Tucker 211

Sampling-Based Methods for Uncertainty and Sensitivity Analysis (Tutorial)
Jon C. Helton 221

Estimation Procedures and Error Analysis for Inferring the Total Plutonium Produced by a Graphite-moderated Reactor
Patrick G. Heasler, Bruce Reid, Christopher Gesh;Tom Burr; Charles Bayne 230

Hydrocarbon Exploration Risk Evaluation Through Uncertainty and Sensitivity Analyses Techniques
Paolo Ruffo, Livia Bazzana, Alberto Consonni, Anna Corradi, Andrea Saltelli, Stefano Tarantola 240

Model Validation in Chemical Process with Multiple Steady States
Carlos M. Villa, Jean-Paul Masy, Flor A. Castillo, Leigh H. Thompson, John W. Weston 250

Response Surfaces and Sensitivity Analyses for an Environmental Model of Dose Calculations
Bertrand Iooss, Nicolas Devictor, François Van Dorpe 260

Simulation Error Models for Improved Reservoir Prediction
Alannah O'Sullivan, Mike Christie 270

Local and Global Uncertainty Analysis of Complex Chemical Kinetic Systems
Judit Zádor, István Gyula Zsély, Tamás Turányi 288

Techniques for Quality Assurance of Models in a Multi-Run Simulation Environment
Michael Flechsig, U. Böhm, Th. Nocke, C. Rachimow 297

Global Sensitivity Analysis for Calculating the Contribution of Genetic Parameters to the Variance of Crop Model Prediction
David Makowski, Cédric Naud, Hervé Monod, Marie-Hélène Jeuffroy, Aude Barbottin 307

Sensitivity Analysis in Presence of Model Uncertainty and Correlated Inputs
Julien Jacques; Christian Lavergne; Nicolas Devictor 317

Varying Correlation Coefficients Cannot Account for Uncertainty About Dependence, but There Are Comprehensive Methods to Do So
Scott Ferson 324

Application of a Markov Chain Monte Carlo Calibration and Uncertainty Framework to a Process-based Integrated Nitrogen Model (INCA)
Bethanna M. Jackson, H. S. Wheater, N. McIntyre; P. Whitehead 329

Validating Mathematical Models of Biological Systems: Application of the Concordance Correlation Coefficient
Normand R. St-Pierre 341

On a Strategy of Reduction of the Lack of Knowledge in Model Validation
Pierre Ladevèze; G. Puel; T. Romeuf 351

Modeling Uncertainty in Population Biology: How the Model is Written Does Matter
Janos G. Hajagos 363

Enhancing the Morris Method
Francesca Campolongo, J. Cariboni, A. Saltelli; W. Schoutens 369

Initial Evaluation of Pure and "Latinized" Centroidal Voronoi Tessellation for Non-Uniform Statistical Sampling
Vicente J. Romero; John V. Burkardt, Max D. Gunzburger, Janet S. Peterson 380

MCMC-based Sensitivity Estimations
Carlos Javier Pérez Sánchez, Jacinto Martín Jiménez 402

Sensitivity Analysis in Conjunction with Evidence Theory Representations of Epistemic Uncertainty
Jon C. Helton, J. D. Johnson, W. L. Oberkampf 410

Combined Array Experiment Design
Leslie M. Moore, Michael D. McKay, Katherine S. Campbell 421

Halftoning and Quasi-Monte Carlo
Kenneth M. Hanson 430

Model Validation Methods for Phase Diagram Determination
Brian J. Reardon, Marius Stan 443

Approximate Representations of Random Intervals for Hybrid Uncertainty Quantification in Engineering Modeling
Cliff Joslyn; Scott Ferson 453

Designing Computer Experiments to Determine Robust Control Variables
Jeffrey S. Lehman; Thomas J. Santner, William I. Notz; Donald Bartel 470

ASSA: Algorithms for Stochastic Sensitivity Analysis
Michiel J.W. Jansen 473

Case Studies in Gaussian Process Modelling of Computer Codes
Marc C. Kennedy, Clive W. Anderson, Stefano Conti, Anthony O'Hagan 476

Nonparametric Prior Elicitation
Jeremy Oakley, Anthony O'Hagan 486

Stochastic Sensitivity Analysis for Computing Greeks
M. Kodal K. Ohmori, D. Yokomatsu, T. Amemiya 492

Participant list

Input Screening: Finding the Important Inputs on a Budget

Max D. Morris

Departments of Statistics, and Industrial and Manufacturing Systems Engineering
Iowa State University, Ames, Iowa, 50011, USA
E-mail: mmorris@iastate.edu

Abstract: One general goal of sensitivity or uncertainty analysis is the determination of which inputs most influence model outputs of interest. Simple methodologies based on randomly sampled input values are attractive because they require few assumptions about the nature of the model. However, when the number of inputs is large and the computational effort required per model evaluation is significant, techniques based on more complex assumptions, analysis techniques, and/or sampling plans are needed. This talk will review some approaches that have been proposed for input screening, with an emphasis on the balance between assumptions and economy, including a brief description of recent work in economical sampling plans.

Keywords: Computer experiment, sensitivity analysis, uncertainty analysis

1. INTRODUCTION

Especially in the early stages of work with a computer model, it is important to determine which inputs are *important* and which are not. The precise definition of "important" is not always the same (and in some cases is never carefully addressed) but is generally related to how much or what kind of influence each input has on outputs of interest. For very simple computer models, such questions may be addressed directly through analysis of the underlying equations. But more complex models require an empirical approach, or *computer experiment* designed to allow determination of the importance of inputs through analysis of numerical output values. The approaches we shall discuss are described as entirely empirical (i.e. "black box"), even though it is understood that in many applications these can be tailored to take advantage of specific knowledge about a model.

In order to be specific, let $y = m(\mathbf{x}), x \in \Delta$, represent what we mean by a "computer model", a deterministic function mapping a vector $\mathbf{x}$ of k input arguments from a defined domain Δ to a scalar-valued output y. In most real problems y would also be vector-valued, but we shall not address complications that this may created here. A particular input x_i may be deemed important if (1.) $\partial y/\partial x_i$ is large in at least some regions of Δ, (2.) y is relatively complex (in some sense) as a function of x_i, or (3.) y varies substantially as the value of x_i changes. These three concepts of "importance" are relatively vague, certainly related, and certainly not exclusive, but one or more of them have been found to be useful in a large variety of problems.

Two characteristics of this problem that make identifying important inputs practically difficult are (1.) the dimension of $\mathbf{x}$ (typically not small), and (2.) the effort required

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

to evaluate m (typically not trivial). The difficulty is easy to understand; if k is large, the number of "points" needed to "fill" it sufficiently to allow characterization of y as a function of $\mathbf{x}$, without extensive knowlege or assumptions about the nature of m, will also be large. But computer experiments requiring a large number of model executions will be prohibitive if each execution is expensive.

Methodologies for the input screening problem have been proposed by several authors, and vary in the assumptions required, the sense in which importance is measured, and the number of model executions required for satisfactory performance. The four approaches reviewed in this paper are representatives of a large collection of ideas introduced as *uncertainty analysis* or *sensitivity analysis*. Our intent here is to point out the spectrum of compromises they offer between required assumptions and required evaluations.

2. ASSUMPTIONS, INPUT IMPORTANCE, AND MODEL RUNS

2.1. Linear Approximation

A time-honored and often useful assumption about a function of interest is that it is at least approximately linear in its arguments. This is such a strong assumption that it effectively boils the entire question of the behavior of m down to a single slope parameter for each input. There can be little question as to the definition of importance of any input in this case. The linearity assumption implies that $\partial y/\partial x_i$ takes the same value everywhere in Δ, which in turn fully defines any sense of how variable y is with respect to x_i. Complexity is not an issue here unless it is also defined so as to increase with the derivative.

Local sensitivity analysis often amounts, in practice, to definition of Δ to be small enough so that an assumption of approximate linearity is plausible. Downing et al. [3] are among the may authors who have described how first-difference approximations to partial derivatives can be derived from simple one-factor-at-a-time computer experiments. More recent practitioners of this approach sometimes use orthogonal 2-level fractional factorial designs of Resolution III or IV as the basis of such studies. Minimal designs supporting this kind of analysis generally contain from approximately k to $2k$ model runs, where k is the number of inputs.

Approaches requiring even fewer model evaluations may be developed if even stronger assumptions can be made. If it is reasonable to assume that most inputs have little or no effect on the output ("effect sparsity" in some literature) and/or the signs of each derivative can be assumed to be known, then group screening plans offer sequential strategies to identify important inputs using substantially fewer than k model evaluations. See the forthcoming book edited by Dean and Lewis [2] for a description of many of these strategies.

While this general approach is often useful and usually simple, one disadvantage is that there is little basis upon which to base an objective analysis of uncertainty. Since there is no formal basis for the statistical interpretation of residuals, quantities such as the t-statistics associated with each slope have only very limited heuristic value.

2.2. Input-Output Correlations

If approximate linearity is not a justifiable assumption, it still may be acceptable in some cases to assume that the slope of y with respect to x_i's, averaged over Δ, is an acceptable measure of input importance. This is probably most reasonable when an argument can be made that y is monotonic in the arguments of interest, and that the degree of nonlinearity in its behavior is limited. In these cases, an index such as

$$\int (y(\mathbf{x}) - \bar{y})(x_i - E(x_i)) f(\mathbf{x}) d\mathbf{x}, \quad \bar{y} = \int y(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

may be reasonable, where $y(\mathbf{x}) = m(\mathbf{x})$ and f is a probability density function.

Such integrals are easily estimated using a relatively modest Monte Carlo sampling of inputs, although most guidelines would suggest the need for more function evaluations than can be used when the strict linear approximation is used. The virtues of using Latin Hypercube sampling rather than unconstrained random sampling of inputs have been argued by McKay et al. [5] and Stine [13]. Iman and Conover [4] take this approach to evaluating the importance of inputs after transforming the output data to ranks.

The connection between k and the number of runs needed for effective Monte Carlo estimation of the integrals is not so clear as it is when a Linear Approximation is used. If more than a few inputs are important, accidental correlations between selected input values can be substantial unless the number of runs is not small compared to k. These problems may be moderated by using quasi-random sequences, e.g. [8], or algorithms such as Owen's [9] that control the degree of correlation between inputs.

2.3. Stochastic Continuity

Over the last 15 years or so, a number of papers have appeared in the statistics literature suggesting that the design and analysis of computer experiments might be based on regarding (1.) m as a realization of a spatial (i.e. Δ) stochastic process (frequentist), or (2.) the generalized uncertainty about m being expressed by such a process (Bayesian). See, e.g. Sacks et al. [10] and Currin et al. [1] for overviews of this approach. The most important practical issue in such approaches is the statement of a spatial covariance function, governing the "complexity" that may be expected in the behavior of the output as each input is varied. One popular functional form is:

$$Cov[y(\mathbf{x}), y(\mathbf{x}')] = \sigma^2 e^{-\sum_i \theta_i (x_i - x_i')^2}.$$

Given data from a computer experiment, likelihood or Bayes procedures may be used to estimate parameters such at the θ_i, and these used as importance indices. The sense of importance in this example function is, again, one of scale; the value of θ_i essentially defines distance in the x_i direction over which a given degree of activity would be expected in y.

Welch et al. [14] described an algorithm, for which the overall structure is much like that of stepwise regression, for identifying the inputs for which estimates of θ_i are largest, i.e. that are most important in this sense. In demonstrating the method, they evaluated two example functions each in $k = 20$ inputs using a Latin Hypercube sample of 50

runs. The methodology worked well in these exercises, but relatively few of the 20 inputs were actually important in each case; it might be reasonably expected that more runs would be needed if more of the inputs were active. The authors suggest that, following the identification of large correlation parameters, a sensible follow-up analysis would be examination of the fitted surface (mean of the conditional or posterior stochastic process) to examine the shape of m as a function of each apparently important input. However, reliable estimation of the response surface is likely to require more runs than reliable estimation of the covariance parameters.

One somewhat philosophical sticking point with (this version of) the Stochastic Continuity approach is that the indices of importance are parameters that do not *directly* describe properties of the function of interest! In the frequentist formulation of the problem, θ_i is a property of the (physically non-existant) process of which m is supposed to be a single realization. In the Bayesian model, θ_i is part of the characterization of a generalized uncertainty (or lack of understanding) of what the model might do under specified circumstances. With sufficient data (and I am not aware of a careful analysis of what this may mean in this application), this distinction may be less important practically than it is philosophically.

2.4. Conditional Variance

The approaches described to this point are predicated on assumptions of linearity, monotonicity, and continuity, respectively, in the model function. Even an assumption of continuity, however, is not always be warranted, and even when it is strictly warranted, the degree of complexity of y as a function of some x_i may make any attempt to explicitly model m difficult or impossible for practical purposes. In such cases it may be more natural or meaningful to define importance in purely statistical terms, e.g. the degree to which y may be expected to vary as x_i varies according to some (possibly arbitrary) probability distribution, completely disregarding any attempt to match a specific change in y to a specific change in x_i.

Sobol' [12], Saltelli et al. [11], and McKay [6] are among those who have proposed input sampling plans that support estimates of conditional moments of the distribution of y, where that distribution is propagated to the output from a specified distribution on the input vector. In particular, where each component of $\mathbf{x}$ is statistically independent of the others, these authors address estimation of

$V_i[E_{(i)}[y(\mathbf{x})]]$ or "first-order variance"
$E_{(i)}[V_i[y(\mathbf{x})]]$ or "total variance"

Here the subscript i means expectation or variance with respect to the marginal distribution of x_i, and subscript (i) implies the joint distribution of all inputs except x_i. No *functional* assumptions about m are involved here, but the nonparametric nature of this approach carries a practical requirement for a large number of model evaluations. Morris et al. [7] have recently identified other sampling plans based on Balanced Incomplete Block Designs that have some advantages for this type of analysis.

While this analysis has substantial appeal for the especially assumption-averse modeler, it also carries a philosophical difficulty (although not as fundamental as the one I

described above for Stochastic Continuity methods). Here the objection is one of analysis efficiency. The indices of importance are estimated based entirely on the computed values of y, along with information about which runs share common randomly drawn values for each input. But the specific values of x_i are not used at all in the analysis; while they intuitively must carry some information of value in most practical situations, avoiding *all* assumptions about the y-to-$\mathbf{x}$ connection makes it difficult to apply this information.

3. COMPARISONS, CONCLUSIONS

The four general approaches outlined in Section 2 differ in (1.) the strength of assumptions that must be made about the model, (2.) the number of model evaluations that are required for practical purposes, and (3.) the sense in which importance is assessed for each input. Relatively strong assumptions leave relatively few degrees of freedom in defining importance, but require relatively few model evaluations for assessment. Relatively weak assumptions allow more subtle definitions of importance (or negatively, do not support the simplest interpretations), but require relatively many model evaluations.

Approach	Assumptions	Required Runs	Importance
Linear Approximation	most	least	derivative
Input-Output Correlations	↑	↓	averaged slope
Stochastic Continuity	↑	↓	complexity
Conditional Variance	least	most	variability

Variations on each of the approaches described here, and other fundamentally different approaches, have been proposed in the literature on computational science, applied mathematics, and statistics – the methods mentioned here are only an example of what has been found to be useful in many applications contexts. Future research might benefit from a broader inspection of how these methods differ, and how they might beneficially be combined to create new "points" along the assumption-data-interpretation spectrum.

REFERENCES

1. C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions, with application to the design and analysis of computer experiments. *Journal of the American Statistical Association* 86:953-963, 1991.

2. A.M. Dean and S.M. Lewis, eds. *Screening.* Springer-Verlag, New York, 2004.

3. D.J. Downing, R.H. Gardner, and F.O. Hoffman. An examination of response-surface methodologies for uncertainty analysis in assessment models. *Technometrics* 27:151-163, 1985.

4. R.L. Iman and W.J. Conover. Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. *Communications in Statistics - Theory and Methods* A9: 1749-1842, 1980.

5. M.D. McKay, W.J. Conover, and R.J. Beckman. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21:239-245, 1979.

6. M.D. McKay. Evaluation prediction uncertainty. Los Alamos National Laboratory Report NUREG/CR-6311, LA-12915-MS, 1995.

7. M.D. Morris, L.M. Moore, and M.D. McKay. Incomplete factorial sampling plans for evaluating the importance of computer model inputs. (Working paper, 2004)

8. H. Niederreiter. Quasi-Monte Carlo methods and pseudo-random numbers. *Bulletin of the American Mathematical Society* 84:957-1041, 1978.

9. A. Owen. Controlling correlations in Latin hypercube samples. *Journal of the American Statistical Association* 89:1517-1522, 1994.

10. J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science* 4:409-435, 1989.

11. A. Saltelli, T.H. Andrew, and T. Homma. Some new techniques in sensitivity analysis of computer model output. *Computational Statistics and Data Analysis* 15:211-238, 1993.

12. I.M. Sobol'. Sensitivity estimate for nonlinear mathematical models (in Russian). *Mat. Model.* 2, 1990.

13. M. Stein. Large sample properties of simulations using Latin hypercube sampling. *Technometrics* 29:143-151, 1987.

14. W.J. Welch, R.J. Buck, J. Sacks, H.P. Wynn, T.J. Mitchell, and M.D. Morris. Screening, predicting, and computer experiments. *Technometrics* 34:15-25, 1992.

The Evaluation of Combustion Mechanisms using Local and Global Sensitivity and Uncertainty Methods.

Alison. S. Tomlin

Energy and Resources Research Institute, University of Leeds, Leeds, LS2 9JT, UK.

Email: A.S.Tomlin@leeds.ac.uk

Abstract: Complex chemical mechanisms are increasingly used within models describing a range of important chemical processes. Within chemical models, kinetic parameters describing the rates of chemical steps and thermodynamics may be highly uncertain, influencing the uncertainty in final model predictions. Traditionally, local sensitivity analysis is employed within commercial modelling packages but may not be appropriate for highly uncertain data where models are nonlinear. This work compares linear sensitivity methods with global techniques such as Morris and Monte Carlo sampling for a kinetic model describing the influence of fuel sulphur on the oxidation of nitrogen within flames. The kinetics forms an important component of larger models describing pollution formation in combustion devices. The analysis reveals the most important rate and thermo-kinetic parameters contributing to the uncertainty in NO predictions for both rich and lean flames. The level of agreement between local and global techniques is highlighted. The use of reduced model representations using fitting methods is also discussed as a way of improving the efficiency of Monte Carlo based methods. Speed ups of a factor of 15 are seen without significant impact on the predicted mean output and standard deviation. For certain conditions, the mechanism is not capable of observing previous experimental data, highlighting the need for structural developments of the model such as including additional reaction steps for which data is not currently available.

Keywords: kinetic mechanism, sensitivity analysis, reduced model, Monte Carlo, MOAT.

1. INTRODUCTION

The use of computational modelling as a design tool is increasing within engineering applications. One area of importance is that of combustion reactor design. Environmental legislation means that engineers must develop combustion applications with low emissions of pollutants such as nitrogen and sulphur oxides. Understanding the impact of fuel trace elements such as nitrogen and sulphur containing compounds on pollutant emissions is important and requires the description of complex chemical mechanisms within the combustion chamber. In many cases mechanism data, such as rate constants and thermo-chemical parameters, are poorly categorised. If confidence is to be placed in the design process then the uncertainty in output predictions resulting from the use of such complex mechanisms should be investigated. Local/linear sensitivity analysis techniques are commomly used to evaluate such mechanisms. They have been developed in a generic way in the process engineering field using packages such as CHEMKIN [1], which is used for a range of applications including chemical mechanism validation in simplified flow environments such as flow reactors, premixed and diffusion flames. Linear methods are employed because they are computationally efficient, but are problematic where uncertainties in inputs are large and models are highly non-linear. This work therefore describes the development of methods for global uncertainty analysis for application within modelling packages such as CHEMKIN.

The paper will present a comparison of local and global uncertainty methods for a test case describing a 1 dimensional model of the influence of sulphur species on the emission of NO (nitrogen oxide) from methane air flames. Experiments show that the influence of sulphur within the flame can lead to both enhancement and reduction of NO emissions depending on whether the flame is fuel rich or fuel lean [2-4]. The chemical mechanism employed has been developed using evaluated rate data where possible. Predictions from the mechanism are evaluated for experimental data sets for a variety of flames and fuel nitrogen contents. Without sulphur present in the fuel the mechanism has previously been shown to exhibit good agreement with experimental profiles of key species. However, the mechanism does not capture the quantitative influence of sulphur on NOx emissions that is demonstrated in the 1D flame experiments[3]. Local sensitivity analysis has already revealed [3] several important reactions that require improved categorisation of the rate data to lower output variance. The present work presents a comparison between local and nonlocal linear methods, the global screening Morris One at a Time (MOAT) method and random sampling Monte Carlo methods using full and reduced models, coupled with scatter plot and correlation analysis.

2. MECHANISM STRUCTURE AND CONDITIONS FOR EVALUATION

The mechanism under investigation consists of an updated version of the Leeds methane/NOx mechanism [5-7]. The SOx extension (156 reversible reactions and 24 species) is based on the mechanism of Glarborg et al. [8] and Alzueta et al. [9] and has been augmented by the inclusion of additional reactions of sulphur containing species appropriate for a methane oxidation environment, and reactions describing sulphur-nitrogen interactions that have previously appeared in the literature or widely available databases. The uncertainty study here focuses on reactions of sulphur containing species and heats of formation of sulphur compounds. It is undertaken for selected experimental conditions from [3] where laser induced fluorescence studies were performed for a variety of low pressure methane flames doped with various levels of HCN and SO_2. Comparisons were made of relative NO concentrations for several dopant levels and flame stoichiometries from $\varphi = 0.7$ to $\varphi = 1.6$ (Table 2 of [3]). Whilst showing similar qualitative trends, previous mechanisms have tended to over predict the relative increase in NO on the addition of SO_2 for rich flames when compared to the experiment, and to under predict the reduction in NO for lean flames.

3. SENSITIVITY/UNCERTAINTY METHODS EMPLOYED

3.1 Uncertainties in Input Data

Well categorised kinetic rate parameters k, such as those from evaluations [10], are often quoted with an accuracy expressed as $\Delta\log k$. Here $\Delta\log k = D$ and D is defined by $\log_{10}k = C \pm D$. This is equivalent to the rate parameter k being uncertain by a factor f where $D = \log_{10}f$. For temperature dependant reactions this respresents the uncertainty at 298K ($f(298)$). A temperature dependant form for second order reactions is given by:

$$f(T) = f(298)\exp\left|\frac{\Delta E}{R}\left(\frac{1}{T}-\frac{1}{298}\right)\right|, \qquad (1)$$

where ΔE is the quoted error in the activation energy. The uncertainty of $\log_{10}k$ is usually assumed to be normally symmetric unless the parameter is stated as an upper or lower limit. One can then define a probability density function (pdf) for the rate parameter according to a distribution type. For less well categorised reactions a pdf cannot be determined and a minimum and maximum possible value are chosen, with an equal probability of the value of the

rate constant existing across the range. Local sensitivity coefficients are problematic in these cases since a most likely value cannot be reliably determined. However, if ranges of possible inputs are chosen, a full Monte Carlo analysis allows the user to determine if the model can "observe" target output values by comparison with experiments, allowing the evaluation of possible structural uncertainties in the model such as missing kinetic processes. Where target outputs are observable then the use of correlation factors or global ranking methods, allows the determination of the most important input factors leading to output uncertainties.

3.2 Linear Sensitivities.

In this study non-local linear sensitivities calculated using the *brute force method* are compared against previous local linear sensitivity studies from [3]. The use of the *brute force method* involves performing a base model run with output y_i, using the nominal values of the input parameters, and *n* extra model runs where each uncertain input parameter j is changed by a small factor Δj. The final output (y_i^*) from each run is determined and the sensitivity coefficient given by: $S_{i,j} = \frac{y_i - y_i^*}{\Delta j}$. The sensitivity at any temporal or spatial point is determined with a computational effort of order *n* (the number of uncertain inputs).

3.3 The Morris One at a Time (MOAT) Method.

Although they allow the study of non-linear interactions between parameters, global methods can be computationally expensive since thousands of model runs may be required. One example of a potentially more efficient screening method is the One at a Time analysis developed by Morris [11]. The method determines an importance ranking for parameters in terms of their mean effect on output variance as well as determining those parameters with linear additive effects and non-linear interactions [12]. In the MOAT method the inputs x_j are assumed to have values in the set {0, 1/(*p*-1), 2/(*p*-1),…,1} where in practice these values are re-scaled to values from within their uncertainty ranges. A perturbation factor Δ is defined as a multiple of 1/(p-1). A control simulation is then performed based on the random selection of parameters from the set {0,1/(*p*-1),…,1-Δ}. A single parameter is then randomly selected and modified by a factor Δ, and a second simulation performed. This is repeated until each factor has been chosen once, corresponding to *n+1* runs. This procedure is repeated *r* times until stable output statistics are obtained. The average output is computed over *r* runs and the cost of the method scales with *r*(*n*+1).

The elementary effect of the j'th component of **x** on the output y_i where x_j has been changed by a factor Δ is given by:

$$d_{ij}(\mathbf{x}) = \frac{y_i(x_1,\ldots,x_{j-1},x_j \pm \Delta, x_{j+1,\ldots,}x_m) - y_i(\mathbf{x})}{\Delta}. \qquad (2)$$

The mean effect across *r* runs is given by: $\bar{d}_{ij} = \frac{\sum_{l=1}^{r} \left| d_{ij}^l \right|}{r}$, and the variance:

$$\sigma^2(d_{ij}) = \frac{r\sum_{l=1}^{r}(d_{ij}^l)^2 - \left(\sum_{l=1}^{r} d_{ij}^l\right)^2}{r(r-1)}. \qquad (3)$$

In this study, ten runs were sufficient to produce stable outputs and a value of p=4 was chosen. Parameters with the highest mean effect have a significant impact on overall output variance and require improved categorisation to lower output uncertainty. If the variance between runs is low then the effect is said to be linear or additive. Parameters with a high variance exhibit non-linear or interactive effects, which is important since it indicates parameters with sensitivities that may change if other parameter ranges are changed through improved categorisation.

3.4 Sampling Based Methods and Monte Carlo Analysis.

The standard method of Monte Carlo analysis is to perform a large number of model runs where in each run a sample of input parameters: $\mathbf{x_k} = [x_{k1}, x_{k2}, \ldots, x_{knx}]$, $k = 1,2,\ldots,n_s$,

of size n_s is randomly selected from the possible values for **x** from the chosen distributions [12]. For a uniform distribution each sample region is assigned equal probability and therefore the sample points are chosen randomly from any region. The corresponding outputs **y** for each sample are determined by rerunning the model n_s times. Examination of the mapping from inputs to outputs can then be performed via a variety of methods including scatter plots, correlation analysis, regression analysis etc. The computational cost is of order n_s where n_s must be large enough for the mean output and the output variance to converge.

4. OVERALL METHODOLOGY

Simulation of the low pressure flames described in [3] has been performed using PREMIX [1] at a pressure of 40 Torr for fuel to air ratios Φ=1.6 and Φ=0.7. 0.3% and 0.5% of SO_2 has been added to the lean and rich flames respectively in line with the experiments [3]. The output y(**x**) is the NO mole fraction in the burnt gas region. Reactions have been treated as reversible with reverse rates calculated from the appropriate equilibrium constants. The sensitivity to the heats of formation therefore forms an important part of the study. Each PREMIX run is computationally fairly expensive since a larger number of coupled non-linear equations must be solved in order to determine the concentration profiles of over 75 species in the flame. The use of a fitted model within the Monte Carlo analysis will therefore also be discussed as a method of reducing the computational expense resulting from large numbers of PREMIX simulations.

The following methods will be presented for comparison:

For uncertainties in rate parameters of sulphur containing reactions:

1. linear sensitivities using the brute force method and a relative change of 10%,
2. MOAT analysis,
3. Monte Carlo analysis using up to 2000 model runs based on full and fitted models.

For uncertainties in heats of formation for the sulphur containing compounds:

4. Monte Carlo analysis using up to 2000 model runs.

Uncertainty ranges for kinetic rate parameters were assigned using f factors with 95% confidence limits where data evaluations existed. For parameters derived from a single experimental or modelling study a factor of 2 was used. Where data derived from a single RRKM calculation, or was estimated, a factor of 10 was assumed. Where the temperature dependence was estimated, an uncertainty factor in ΔE of 2 was assumed. Of the 155 parameters 18% derived from evaluated rate data, 18% from a low number of measurements, 7% from measurements with no evaluation, 8% from a single RRKM study and 49% were estimated. For this reason, only input ranges were determined and not pdfs. The analysis will not therefore allow pdfs of the outputs to be determined, but rather allows the evaluation of the

current mechanism in order to inform its future development. Thermodynamic data and their uncertainties were obtained from the databases of Burcat [13] and NIST [14]. Coefficients for a large number of the sulphur compounds originated from modelling techniques [15]. Where a single model value was used an uncertainty of ±100 ($kJmol^{-1}$) was assumed. Because almost half the thermodynamic data was of this type uniform distributions were used.

5. RESULTS AND DISCUSSION

5.1 Linear Sensitivity Coefficients.

Φ=1.6: The importance ranking from the non-local linear sensitivity study for the rich flame is presented in Table 1. The highest ranked reaction is SO+NH=NO+SH in agreement with the local sensitivity study in [3]. The second highest ranked reaction from the local study was SO_2+H=SO+OH, which is represented here by its reverse rate [16] and ranked 5th. The second highest ranked reaction from this study is SH+NH=NS+H_2, which was also highlighted by the local sensitivity study and has an estimated rate giving scope for improvement. The reaction H_2S+M=H+SH+M, ranked third here, was not identified using local sensitivity coefficients. SO+O_2=SO_2+O was highlighted by the previous study and again has a negative sensitivity when using non-local methods. There is therefore some broad agreement between the local and non-local linear sensitivity methods as well as notable differences in ranking.

Table 1 – Comparison of importance ranking of sulphur containing reactions in the rich flame (Φ = 1.6) from the linear brute force (lbf) and MOAT analyses.

React. No.	Reaction	lbf Rank	MOAT Rank	Pearson Coeff.	Source of Data
1	SO+NH=NO+SH	1	1	0.79	Single meas.
2	SH+NH =NS+H_2	2	5	0.12	Estimated
3	H_2S+M=H+SH+M	3	7	-0.08	Unevaluated measurements.
4	SO+O_2=SO_2+O	4	14	-0.03	Evaluated
5	SO+OH=SO_2+H	5	4	0.2	Single meas.
6	S+OH=SH+O	6	15	-0.07	Estimated
7	HSO+H=SH+OH	7	33	-0.03	Estimated
8	S+H_2=H+SH	8	10	-0.06	Unevaluated measurements.
9	SO+N=NO+S	9	2	0.56	Estimated
10	H_2S+M=H_2+S+M	10	55	0.01	Unevaluated measurements.
11	HSOH=SH+OH	69	3	0.04	Estimated
12	SH+H=H2+S	11	6	0.06	Unevaluated measurements.
13	SH+NO=SN+OH	13	8	0.01	Estimated
14	SN+O=SO+N	39	9	0.05	Estimated

Φ=0.7: In the lean case a 10% increase in the selected rate parameters was not sufficient to cause any detectable change in the NO mole fraction. An factor of 10 increase was required to produce a detectable difference, making it impossible to determine an importance ranking using a linear method. The analysis indicates that for lean conditions, the NO concentration is highly insensitive to the forward rate parameters of the sulphur reactions around their nominal values.

5.2 MOAT Analysis

Φ=1.6: Figure 1a shows the variance of the factor effects plotted against the mean effects from the MOAT analysis for the rich flame. Parameters with a low significance in terms of output

variance appear at the bottom left of the plot. Those in the bottom right segment have a high linear effect on model output and those in the upper portion show strong non-linear or interactive effects. The reaction SO+NH=NO+SH appears at the bottom right of the figure showing a strong linear effect on the output in agreement with the linear methods. SO+N=NO+S also shows a strong linear effect and ranks second in the MOAT analysis as shown in Table 1, although lower in the linear analyses. HSOH = OH+SH shows a strong mean effect and a high variance, indicating that the sensitivity to this reaction strongly depends on the values of the other parameters. Not surprisingly this reaction was not identified as important by the linear methods. Its high ranking by the MOAT analysis is strong evidence of the importance of using global uncertainty techniques.

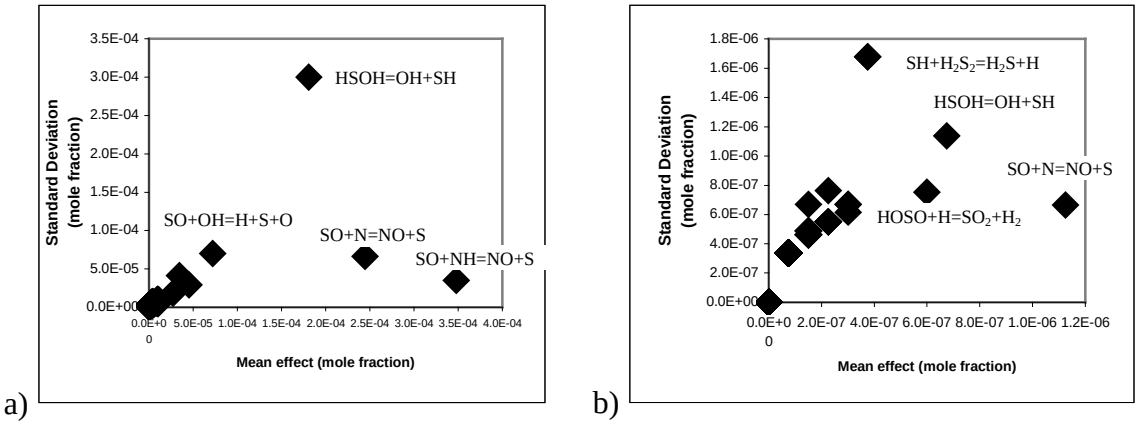


Figure 1. Morris One at a Time Analysis for (a) (Φ = 1.6) and (b) (Φ = 0.7)

Table 2 –Importance ranking of sulphur reactions in lean flame (Φ = 0.7) from MOAT analysis.

React. No.	Reaction	MOAT Rank	Source of Data
9	SO+N=NO+S	1	Estimated
11	HSOH=SH+OH	2	Estimated
15	HOSO+H=SO_2+H_2	3	Estimated
16	SH+H_2S_2=H_2S+HS_2	4	Single low temp. meas.
17	S_2+H+M=HS_2+M	5	Estimated
1	SO+NH=NO+SH	6	Single meas.
18	SO2+OH=SO3+H	7	Estimated
19	SO+M=S+O+M	8	Unevaluated measurements.

Φ=0.7: Figure 1b and Table 2 represent the output from the MOAT analysis for the lean flame. The overall mean effect of the rate parameters on the NO mole fraction in the burnt gas region is much lower than for the rich flame. SO+N=NO+S shows the highest overall mean effect and since this has an estimated rate there is some scope for improvement in predictions by its better categorisation. There are some differences between the reactions ranked highly by the MOAT analysis and by the linear studies in this and previous work [3]. For example reactions 11, 16, and 17 had no impact in the linear study despite being modified by up to a factor of 10. This indicates interaction effects between rate parameters in the scheme. The dominant uncertainty in the conversion from SO_2 to SO_3 identified by the MOAT analysis is via OH rather than the pressure dependant reaction in the linear study. The low mean effect of all reactions for this flame again highlights the low sensitivity of NO to the sulphur chemistry for lean conditions.

5.3 Monte Carlo analysis.

The output values resulting from the randomly sampled input distributions for the rich flame are presented in Fig. 2 as scatter plots for high ranking reactions from the MOAT analysis. Pearson correlation functions [12] can also be used to determine the strength of the linear response of the output to the input parameters (see Table 1). A high correlation implies a strong linear response of NO concentrations to an increase in the rate parameter. These correlation coefficients do not take into account interactive effects.

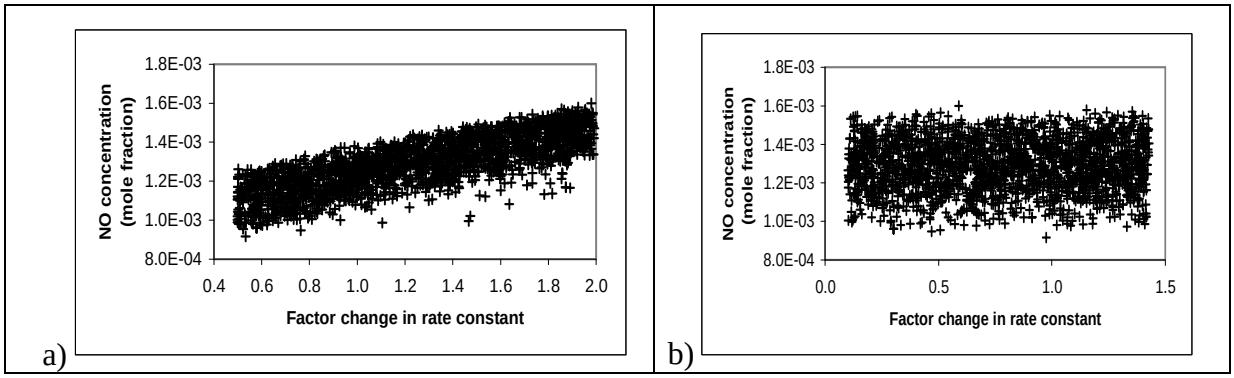


Figure 2. Scatter plot showing Monte Carlo simulation of changes in final NO concentration resulting from changes in reaction rate constant (a) SO+NH=NO+SH, (b) HSOH = OH+SH.

The Pearson coefficient of SO+NH=NO+SH is r = 0.79 (Fig. 2a) showing a strong linear response in line with both the linear sensitivity and MOAT methods. The scatter about the mean effect in Fig. 2a is due to the influence of other parameters on the output. This reaction is ranked highest by all the methods of analysis and is clearly important for the rich flame. SO+N=NO+S also shows strong linear behaviour as demonstrated by r = 0.56, in agreement with the MOAT analysis. As expected, the Monte Carlo analysis coupled with linear regression techniques agrees well with the MOAT analysis for reactions with strong linear effects. Interestingly this reaction is fairly low down the importance ranking using both linear methods. Because it is estimated however, this parameter has a highly uncertain input range and therefore its overall contribution to the output uncertainty is high, even though its sensitivity may not be. HSOH=SH+OH is ranked third by the MOAT analysis. It has an extremely low ranking using the linear method and as Fig. 2b shows there is a large amount of scatter in the Monte Carlo results. Its correlation coefficient is very low (r = 0.04). This stems from the fact that the sensitivity of NO mole fraction to this reaction rate parameter changes sign in different regions of the input parameter space. In order to highlight the importance of such non-linear responses either the MOAT method or higher order correlation techniques combined with Monte Carlo based methods must be used. The ranking of reactions SO+OH=SO_2+H and SH+NH =NS+H_2 as derived from the correlation coefficients and the MOAT analysis are similar due to their high mean effects on the output but low variances. Scatter plots and Pearson coefficients are not presented for the lean flame since insufficient changes in NO mole fraction occur. Each method therefore highlights the lack of sensitivity to the sulphur chemistry in the lean flame leaving limited scope for model improvement via better characterisation of the current rates.

5.4 Agreement with experiment.

The experimental study in [3] showed a 16% increase in the NO mole fraction in the burnt gas region of the rich flame on the addition of 0.5% SO_2, with an experimental uncertainty of

±12%. In order to overlap with the experimental predictions the NO mole fraction in the burnt gas region must be < 9.545 x 10^{-4}. The scatter plots show that in order to achieve this, several of the parameters must be chosen from the extremes of their input ranges. A view could be taken that the likelihood of these parameters being at their extreme values is low. This leads to the possibility of structural problems within the model, such as missing reaction sets or rate parameters that are significantly more uncertain than has been assumed. For the lean flame, the model does not observe the concentration determined in the experiments with any combination of the forward rate parameters for the sulphur reactions used here.

5.5 Influence of Heats of Formation.

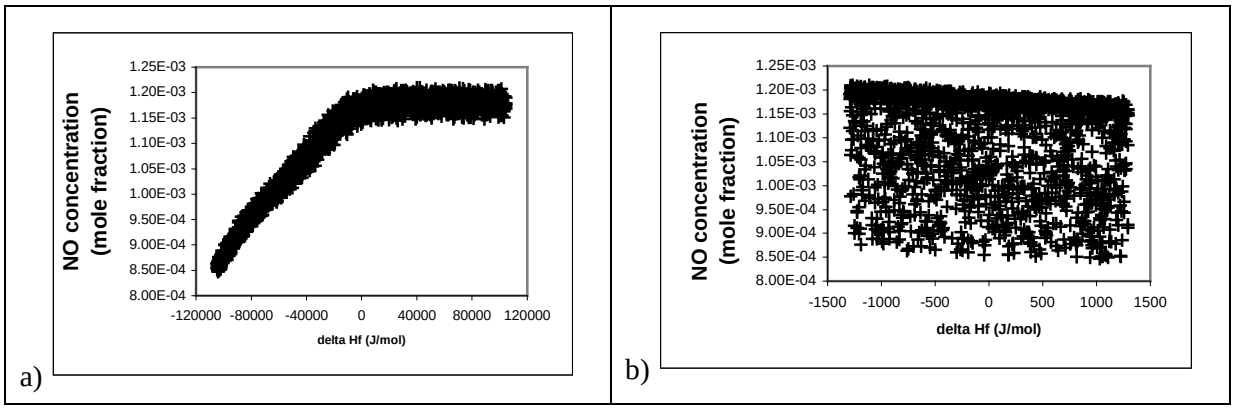


Figure 3. Scatter plot showing Monte Carlo simulation of changes in final NO concentration resulting from changes to heats of formation of a) NS, b) SO.

Uncertainties in the heats of formation of sulphur containing species may affect the reaction kinetics as the rates of the reverse reactions are determined through the equilibrium constant. Figure 3 shows scatter plots from a Monte Carlo analysis of uncertainties in heats of formation for the two most significant parameters for the rich flame. The influence of the heat of formation for NS ($\Delta_f H_{298}$(NS)) is extremely non-linear with a strong response at values lower than the quoted value, which flattens off at higher values. The output NO mole fraction will therefore be dominated by ($\Delta_f H_{298}$(NS)) only if the quoted value is too high. At higher values other compounds start to have an effect as shown by the scatter plots for SO. The large amount of scatter stems from the dominant influence of NS in its lower range. However, as the effect of NS saturates, a negative linear response to $\Delta_f H_{298}$ for SO can be seen. The influence of ($\Delta_f H_{298}$(SO)) is therefore highly dependant on the value chosen for ($\Delta_f H_{298}$(NS)).

5.6 Computational Requirements and Stabilisation of Output Statistics.

Theoretical estimations of the number of Monte Carlo runs required for the analysis of 155 uncertain parameters would be extremely high. In reality, 2000 runs are sufficient for output statistics to settle. Because of the large number of uncertain input parameters attempts could be made to reduce the computational costs by focussing the Monte Carlo analysis on a smaller number of parameters, for example those highest ranked from the MOAT analysis. For comparison purposes therefore, a second Monte Carlo run has been performed for the rich flame, where only the top 15 reactions from the MOAT analysis vary randomly within their input uncertainty ranges. All other parameters are fixed at their nominal values. The final means for the full and 15 parameter runs compare well at 1.290x10^{-3} and 1.283x10^{-3} respectively, as do the final standard deviations of 1.321x10^{-4} and 1.320x10^{-4}. This shows that the top 15

parameters identified by the MOAT analysis account for almost all the variance in the final output. However, the 15 parameter run does not settle statistically any faster than the full run, meaning that large computational savings have not been made, since a similar order of full model simulations have been required. This indicates that the number of samples used in the Monte Carlo analysis depends not on the total number of uncertain parameters, but on the number of important parameters that significantly affect the output statistics. In many cases, the number of samples required may not rise dramatically with the number of uncertain input parameters, since only a few parameters may dominate the output uncertainty.

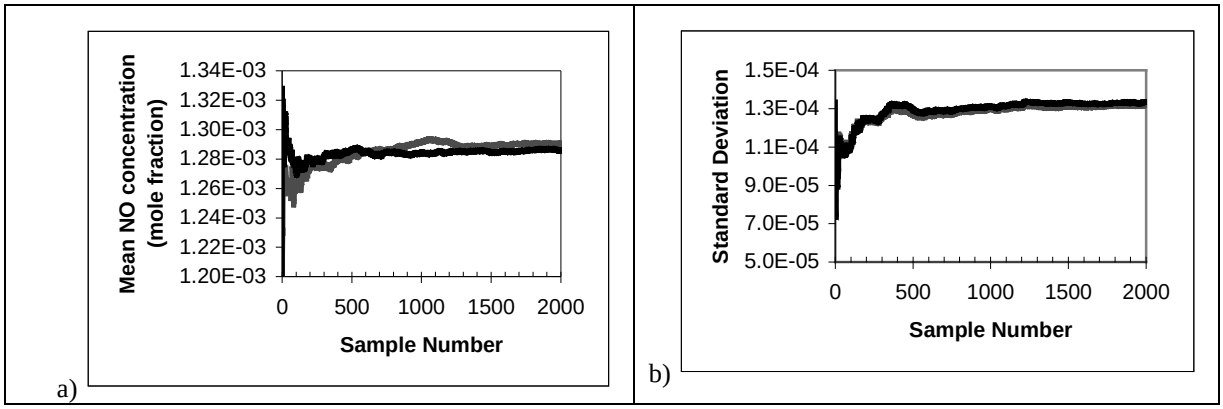


Figure 4 Comparison of output from Monte Carlo simulations using full and reduced models. Grey line - full PREMIX runs, black line – runs using fitted polynomial.

Further attempts may be made in order to reduce the computational cost of the Monte Carlo runs. In this work a second order polynomial equation has been fitted to represent the input output relationships from a small number of randomly sampled PREMIX runs using a Gram-Schmidt orthonormalisation procedure [17]. Again 15 input parameters are utilised as determined by the MOAT analysis and 120 samples are used for the fit. The polynomial coefficients are determined by minimizing the root mean square error (rms) of the fitted function with respect to the target output data derived from the full model. Terms not reducing the rms error are discarded. The final polynomial is factorised using Horner equations to minimize the number of arithmetic expressions required in the final simulation. Figure 4 shows that both the mean output and the output variance behave in a very similar way with increasing sample size for the full PREMIX simulations and those using the polynomial model. After 2000 simulations both give extremely similar results, despite the reduced model being formulated using only 120 full simulations. The final output means from the full and reduced models are 1.283×10^{-3} and 1.286×10^{-3} respectively, and the final standard deviations 1.320×10^{-4} and 1.337×10^{-4}, showing that the polynomial model gives similar results for >15 times lower computational costs. In contrast, if only 120 full model runs had been used then the final mean and standard deviations would not have settled down and would have been 1.272×10^{-3} and 1.217×10^{-4} and therefore do not represent the final values as well as using 2000 polynomial model runs. A further point is that the polynomial model directly reveals second order interactions between parameters.

6. FINAL DISCUSSION AND CONCLUSIONS

The analysis shows that useful information can be obtained from linear sensitivities, although both the linear sensitivities and the regression analysis fail to identify important reactions with strong non-linear interactions. Using combined global methodologies however, highlights a

range of reaction rates and thermo-kinetic parameters leading to output uncertainties in NO predictions providing useful information for further kinetic studies. Each method agrees that the sensitivity of NO mole fraction to sulphur containing reactions is much higher in the rich flame than the lean flame, which is extremely insensitive to the parameters tested here. Possible mechanistic problems relevant to the lean flame must therefore be identified. Since the mechanism utilised in the study contained all reactions present in the literature with measured, modelled or estimated rates, significant further improvements may involve new elementary reactions being postulated with a further requirement for the categorisation of their rate data.

The MOAT analysis performed in this study successfully identified those reactions making the major contribution to the overall output uncertainty as determined by Monte Carlo techniques. Once this group of 15 parameters had been identified it was demonstrated that the computional expense of using Monte Carlo analysis could be significantly reduced by fitting a polynomial model describing the relationship between the 15 important parameters and the chosen model output. 120 random sampling runs proved sufficient for the fit, that was then capable of predicting the mean output and standard deviation across many runs with a high degree of accuracy when compared to analysis using full model runs. Since the computation of polynomials is so fast, the use of the reduced model gave speed ups of greater than a factor of 15. This approach of combining a global screening method with random sampling analysis using a fitted model could therefore have potential benefits for the future application of global uncertainty methods where individual model runs are computationally time consuming.

ACKNOWLEDGEMENTS

The author would like to thank EPSRC and Argonne National Labs for financial support and K. Hughes, M. Pilling, J. Hessler, T. Turányi, and J. Zador for helpful discussions.

REFERENCES

1. R. J. Kee, J.F. Grcar, M.D. Smooke and J.A. Miller, Report No. SAND85-8240, 1985.
2. E. Hampartsoumian, W. Nimmo, A.S. Tomlin, K.J. Hughes, T. Mahmud, *Journal of the Institute of Energy*. 71:137-145, 1998.
3. K.J. Hughes, A.S. Tomlin, V.A. Dupont, M. Pourkashanian, *Faraday Discuss*. 119:337-352, 2001.
4. W. Nimmo, E. Hampartsoumian, K.J. Hughes and A.S. Tomlin, *Proc. Combust. Inst.* 27:1419-1426, 1998.
5. K.J. Hughes, A.S. Tomlin, E. Hampartsoumian, W. Nimmo, I.G. Zsély, M. Ujvári, T. Turányi, A.R. Clague, M.J. Pilling, *Combust. Flame* 124:573-589, 2001.
6. K.J. Hughes, T. Turányi, M.J. Pilling, A.S. Tomlin, http://www.chem.leeds.ac.uk/Combustion
7. K.J. Hughes, T. Turányi, A.R. Clague, M.J. Pilling, *Int.J.Chem.Kinet*.33:513-538, 2001.
8. P. Glarborg, D. Kubel, K. Dam-Johansen, H.M. Chiang and J.W. Bozzelli, *Int. J. Chem. Kinet.*, 28:773, 1996.
9. M. U. Alzueta, R. Bilbao, and P. Glarborg, *Combust. Flame*, 127:2234-2251, 2001.
10. R. Atkinson, D. L. Baulch, R.A. Cox, R.F. Hampson, J.A. Kerr, M. J.Rossi, and J. Troe, *J. Phys. Chem. Ref. Data*, 26:1329, 1997.
11. M.D. Morris, *Technometrics*, 33:61-174, 1991.
12. A. Saltelli, K. Chan, E.M. Scott, *Sensitivity Analysis*, Wiley, Chichester, UK, 2001.
13. A. Burcat, Burcat http://www.technion.ac.il/~aer0201/
14. NIST, http://webbook.nist.gov/chemistry/
15. C. Melius, http://z.ca.sandia.gov/~melius/
16. M.A. Blitz, K.W. McKee, and M.J. Pilling, *Proc. Combust. Inst.* 28:2491-2497, 2000.
17. R. Lowe, A. Tomlin. *Atmos. Envir.* 24:2425-2436, 2000.

Application of MCMC – Global Sensitivity Analysis Method for Model Calibration to Urban Runoff Quality Modeling

A. Kanso[a,*], *G. Chebbo*[a,b] *and B. Tassin*[a]

[a] Centre d'Enseignement et de Recherche Eau, Ville et Environnement, Ecole Nationale des Ponts et Chaussées, 6 - 8 avenue Blaise Pascal, 77455 Marne-la-Vallée – France;
[b] Faculty of engineering, Lebanese University. Airport Street, Beirut – Lebanon
E-mail: kanso@cereve.enpc.fr, chebbo@cereve.enpc.fr, tassin@cereve.enpc.fr

Abstract: In stormwater quality modeling, estimating the confidence level in conceptual model parameters is necessary but difficult. The applicability and the effectiveness of a method for model calibration and model uncertainty analysis in the case of a four parameters lumped urban runoff quality model are illustrated in this paper. This method consists of a combination of the Metropolis algorithm for parameters' uncertainties and correlation assessment and a Variance-based method for global sensitivity analysis. The use of the Metropolis algorithm to estimate the posterior distribution of parameters through a likelihood measure allows the replicated Latin Hypercube Sampling method to compute the parameters' importance measures. Calibration results illustrate the usefulness of the Metropolis algorithm in the assessment of parameters' uncertainties and their interaction structure. The sensitivity analysis demonstrates the insignificance of some parameters in terms of driving the model to have a good conformity with the data. This method provides a realistic evaluation of the conceptual description of the processes used in models and a progress in our capability to assess parameters' uncertainties.

Keywords: Uncertainty analysis, Global sensitivity analysis, Bayesian inference, Model calibration, Urban runoff, Quality modeling

1. INTRODUCTION

Since the seventies, an important number of research programs (National Urban Runoff Program, in the USA (1978-1983), French Campaign (1980-1982), Experimental Urban Catchment "le Marais" (1994-2000), …) have shown that the urban stormwater is a significant source of pollution for the receiving systems. This pollution results mostly from the erosion caused by the runoff of particulate pollutants accumulated on the urban surfaces and in sewers during the dry weather period (Figure 1). Moreover, in old urban centers combined[†] sewer systems are found, whereby, during wet weather periods, mixed rain and wastewaters may reach the receiving system through combined sewer overflows.

Within the European Union, control of this pollution was concretized in government policy and Community legislation. Concerning the urban drainage, the European Directive n°91/271 of May 1991 on wastewater treatment forces the communities to take into account the pollution discharged into receiving waters during storm events.

---

[*] *Corresponding author. Tel.:+33(0)1 64 15 36 30; Fax:+33(0)1 64 15 37 64*

[†] Combined sewer system is used in old cities to drain both the urban stormwater and the wastewater

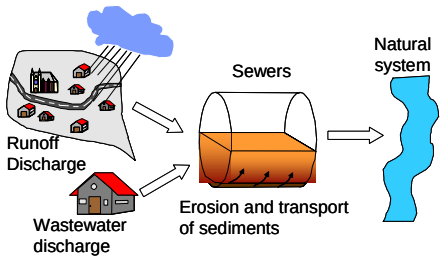


Figure 1 Sources of urban water pollution

Mathematical and computational modeling seems to be a necessary decision-making tool for the management of urban stormwater pollution. Currently, existing models are based on a combination of complex models including conceptual but empirical formulations that describe the processes of generation and transport of pollutants during rainfall. The parameters governing these functions do not have a physical interpretation and therefore, cannot be measured directly in the field. Instead, these parameters must be indirectly estimated using a calibration procedure whereby the model's parameters are adjusted until the system's and the model's outputs show an acceptable level of conformity.

However, the difficulty, expensiveness and uncertainty level of the in situ measurement of urban stormwater pollution generate data that rarely allow a satisfactory calibration and validation of these models [1]. Furthermore, classical optimization methods that are still used up to date for calibration don't allow neither an estimation of the significance of the obtained optimal parameter set, nor a realistic quantification of models' uncertainty. Thus, the existing urban stormwater quality models are rarely used for practical application.

In this paper, we present the results of testing the applicability and the effectiveness of a method for model calibration/validation/sensitivity analysis in urban runoff quality modeling. This method based on the Monte Carlo Markov Chain sampling techniques "MCMC" consists of a combination of a Metropolis algorithm for statistical inference and a Variance-based method for the Global Sensitivity Analysis. This test will be done using data resulting from a survey conducted on the «Marais» catchment in the center of Paris – France [2].

This paper is organized as follows: In section 2, we discuss the difficulties encountered in urban runoff quality modeling. In section 3, we present a general overview of the uncertainty and sensitivity analysis methods. In section 4, we describe the MCMC-GSA method by introducing the Metropolis algorithm, the replicated Latin Hypercube sampling method and their use in the model's calibration and sensitivity analysis. In section 5, we examine the applicability of this method in the case of urban runoff quality modeling. Finally, in section 6, we summarize the methodology and discuss the results.

2. URBAN RUNOFF QUALITY MODELING

It is obvious that modeling represents a necessary tool for understanding the behavior of the urban drainage system and a predictive tool in decision making. For this purpose, models have been developed to simulate the urban water cycle for both quantitative and water quality aspects. Concerning quantitative stormwater management, researchers developed runoff and water flow models that are widely used by managers. However, concerning storm water quality management, researchers built complex models whose structure corresponds to the course of pollution. These models simulate the pollutants' accumulation on the urban catchments, their erosion by runoff, the erosion of sediments in the sewers, and finally the

transport of pollutants through sewers to the outlet. However, despite that many models have been proposed since 1971 (first version of SWMM by US-EPA), several difficulties are facing attempts of stormwater quality modeling.

First of all, the physical, chemical and biological phenomena occurring simultaneously at each stage of the processes of generation and transport of pollution in the system make the system very complex. Moreover, space scales vary greatly considering the heterogeneity of the system's characteristics (topography, watersheds, pipes, sediments size), and time scales vary from several days corresponding to the dry weather period, to few minutes during the wet weather period. Therefore, the only possible modeling approach is the conceptual one.

Second, despite the efforts that have been done to understand the sources and the mechanisms governing the processes involved, the dynamics of accumulation, erosion and transport of pollutants are not well known especially in what concerns the sources and processes of pollution generation in sewers. Currently, modelers tend to divide the urban catchment to a number of sub-catchments of few tens of hectares connected by a sewer network. Runoff models, which are initially developed for surfaces, are used to conceptually describe the accumulation and erosion processes on sub-catchments for which little knowledge is currently available. Erosion and transport models of in-sewers solids' are derived from alluvial hydrodynamics, which poorly describe the real behavior of a sewer system during a rain event. So, great discrepancies exist between the current state of knowledge concerning phenomena and the models used.

Third, field surveys for collecting data necessary for the development of models are difficult and expensive. In consequence, input data (topography, sediment sewer deposits, rain intensity, etc…) and quality measurement data (pollutants concentrations) are rare and characterized by great uncertainties (in the range of 30%) [1]. They rarely allow a satisfactory calibration of the model's parameters.

Finally, while considerable attention has been given to develop global calibration procedures that estimate a best set of parameter values, noting that this is not an easy task especially that most of the models are non-linear [3, 4], much less attention has been given to both the assessment of the significance of the obtained optimal set of parameters, and the realistic quantification of models' uncertainty. Thus the estimated parameters from these models are generally error-prone leading to considerable uncertainty in the calibrated model.

Improving these models and their usefulness requires modelers to use a more robust methodology for calibration and validation of models. Such methodology should be able to provide both an assessment of the uncertainties in the model's parameter values and an evaluation of the confidence level of the model's predictions. Uncertainty and sensitivity analysis are therefore indispensable for any modeling improvement attempt in this field.

3. UNCERTAINTY AND SENSITIVITY

In the last decade, great attention has been given to the Bayesian inference for model calibration and uncertainty assessment particularly in the case of complex hydrological models [5, 6]. Nevertheless, its application in environmental modeling is very rare.

Bayesian approach, expresses uncertainties in the model's parameters θ in terms of probability. Parameter uncertainty is quantified first by introducing a prior probability distribution $P(\theta)$,which represents the knowledge about θ before collecting any new data, and

second, by updating this prior probability on θ to account for the new data collected (D). This updating is performed using Bayes' theorem, which can be expressed as:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{\int P(D|\theta) \cdot P(\theta) \cdot d\theta} \quad (1)$$

Where $P(\theta|D)$ is the posterior distribution of θ; $\int P(D|\theta) \cdot P(\theta) \cdot d\theta$ is a normalizing constant required so that $\int P(\theta|D) \cdot d\theta = 1$, and $P(D|\theta)$ is the conditional probability for the measured data given the parameters. $P(D|\theta)$ is often referred to as the likelihood function.

Unlike traditional statistical theories based on first order approximations and multi-normal distributions that may fail especially when dealing with nonlinear complex models [5], Monte Carlo Markov Chain "MCMC" technique have become increasingly popular as a general method that provides a solution to the difficult problem of sampling from a high dimensional posterior distribution [7]. The idea behind MCMC for Bayesian inference is to generate enough samples from a random walk which adapts to the true posterior distribution $P(\theta|D)$. A variety of appropriate Markov chains can be constructed, but all of them are special cases of the Metropolis algorithm [8]. A study conducted by Kuczera and Parent (1998) demonstrated the capability of the Metropolis algorithm to produce reliable inferences for the parameter's uncertainty assessment in the case of hydrological models.

This posterior distribution represents the uncertainty in the model's parameters and can be propagated through a Monte Carlo method to assess the uncertainty in the model's output attributable to the parameters' uncertainties. However, as the obvious objective of calibration is to reduce the uncertainty in the model's output, it seems necessary to conduct global sensitivity analysis to determine on one hand, which parameters contribute the most to the output variation and require reducing their variances to minimize the variance in the model's output; and on the other hand, which parameters are insignificant and can be discarded from the model. Thus, using this method we can determine the type of research that is required to reduce the output's uncertainty by reducing the variance in some of the model's parameters.

There are many different ways to perform a sensitivity analysis, the method that will be used in this paper is called a "Variance based" method where the uncertainty in the model's output Y is measured by its variance $V(Y)$ and thus can be partitioned to the sum of a top marginal variance and a bottom marginal variance as follows:

$$V(Y) = V[E(Y|U)] + E[V(Y|U)] \quad (2)$$

Where U is a subset of one or more elements θi. $V[E(Y|U)]$ is the variance of the conditional expectation of Y given U and it will be equal to zero if Y is completely independent of U, $E[V(Y|U)]$ is the expectation of the conditional variance of Y given U and it will be equal to zero if Y depends only on U [9]. In this context, the main effect, or first order sensitivity index S_U, representing the sensitivity of Y to the parameter U is defined as $S_U = V[E(Y|U)]/V(Y)$. The total effect, or total sensitivity index S_{TU} is defined as $S_{Ti} = E[V(Y|\theta_{\sim U})]/V(Y)$ where $\theta_{\sim U}$ indicates all the factors but U.

Many estimation procedures of S_U and S_{TU} are available in case of independent parameters. However, when the parameters are correlated, a replicated Latin Hypercube sampling method [9] for the estimation of the importance measure of parameters can be used.

4. MODEL ASSESSMENT METHOD

In this paper, a combination of two complementary and model – independent techniques is used to quantitatively assess the uncertainties associated with the model's parameters as well as the output of the model itself.

4.1. Metropolis algorithm

Although the Metropolis algorithm is not the most efficient Markov Chain sampler, it is chosen in this study because of the simplicity of its implementation, and its generality. It only requires knowledge about the likelihood function to update simultaneously the parameters set for each iteration. Supposing that residuals between model and observation are $N(0, \sigma^2)$, the likelihood function can be written in the multiplicative form:

$$P(D \mid \theta) = \prod_{t=1}^{n} \frac{1}{\left(2 \cdot \pi \cdot \sigma^2\right)^{1/2}} \cdot e^{-\frac{(Y_t - f(X_t,\theta))^2}{2 \cdot \sigma^2}} \qquad (3)$$

Where $(Y1,\ldots,Yn)$ is the vector of the measured response Y, $(X_1,\ldots, X_n)$ is a vector of input data, $\theta = (\theta_1,\ldots, \theta_p)$ is the vector of p unknown parameters, and $f(\,)$ is the model's output. σ is considered, as well as θ, as a set of parameters to be estimated during calibration.

At each iteration, candidate values of parameters are drawn from a multi-normal transition probability distribution for which the variance could be tuned up in a way to increase the speed of convergence. However, updating periodically (automatically) the variance during the simulation, as proposed by Kuczera [5] is subject to difficulties: how can one be sure that the samples used to update the variance contain information of a good quality that can help to ensure the convergence of the chain to the limit distribution? We suggest fixing a prior value of the variance according to the information about the parameters during all the simulation.

An interesting feature of the Metropolis algorithm is that the interaction among the model's parameters is reflected in the likelihood function, so there will be no need to incorporate correlation in the prior distributions of parameters. In order to avoid favoring any initial value, the use of a uniform prior distribution over the range of parameters may seem reasonable [6].

4.2. Replicated Latin Hypercube sampling

The Replicated Latin Hypercube Sampling method r-LHS has been employed in this study to assess the importance measure of the parameters. This method use r replicate Latin hypercube samples of size k to produce $m = r \times k$ parameter vectors θ in total. The same k values of each component U of θ will appear in each replicate but the matching within each one will be done independently. For this application the k values of each parameter U are sampled from its posterior distribution inferred with the Metropolis algorithm. The Iman & Conover rank correlation method [10] has been considered for the r-LHS in order to induce parameters' correlation in the sample. After making the computer runs using the m replicated samples, the importance of U is assessed by computing the ratio S_U:

$$S_U = \frac{SSB}{SST},\ SSB = r\sum_{i=1}^{k}(\overline{y_i} - \overline{y})^2,\ SST = \sum_{i=1}^{k}\sum_{j=1}^{r}(\overline{y_{ij}} - \overline{y})^2,\ \overline{y_i} = \frac{1}{r}\sum_{j=1}^{r} y_{ij},\ \overline{y} = \frac{1}{k}\sum_{i=1}^{k}\overline{y_i} \qquad (4)$$

y_{ij} represents the output value that corresponds to the *i*th value U_i, in the *j*th replicate. In this paper, we are interested in the sensitivity analysis for the likelihood measure in order to identify the parameters that are mainly driving the model to have a good conformity with the data. Ratto [11] showed that sensitivity analysis for the likelihood gives useful information for model calibration especially when great interaction exists between parameters.

5. CASE STUDY

In this paper, we apply the method on the case of urban runoff modeling firstly on the scale of a sub-catchment as used in practice and secondly on the scale of a street surface.

5.1. Site description

Two different watershed scales have been used in this study: the first one WS1 is a 42 ha urban catchment (91% imperviousness) drained by a combined sewer system and the second one WS2 is a 160 m^2 street surface. The used rain event database covers a continuous period of 16 months (1996-1997) with 151 rain events. Suspended solid SS pollutographs* were measured for 40 rain events at the outlet of the combined sewer, and for 11 rain events at a street gully collecting discharge from the street. These data were acquired on the experimental catchment "le Marais" in the centre of Paris [2].

5.2. Model description

The model used in this study to simulate the Suspended Solids pollutograph is a very classical one. It describes the particulate pollutants' erosion during the storm event and their accumulation on the watershed during the preceding dry weather period. This model was at first proposed to be used on street surface scales. However, it is currently used in all available urban stormwater pollution software at the scale of urban subcatchment where both sewers and urban surfaces are described as one entity.

Equation 5 and Equation 6 represent the two accumulation models tested in this paper. Equation 5 calculates the accumulation of pollutants assumed to follow an asymptotic behavior that depends on two parameters: an accumulation rate *Daccu* (kg/ha/day) and a dry erosion rate *Dero* (day^{-1}) [12].

$$\frac{dMa(t)}{dt} = Daccu \cdot Simp - Dero \cdot Ma(t) \quad (5)$$

$$\frac{dMa(t)}{dt} = Kaccu \cdot (M_{\lim} \cdot Simp - Ma(t)) \quad (6)$$

Where *Ma(t)* (kg) is the available pollutants' mass at time t and S_{imp} (ha) is the impervious area. Equation 6 represents a mathematical reformulation of the previous model and was chosen in regard to the obtained results. This model depends on two parameters: an accumulation coefficient *Kaccu* and a maximum accumulated mass *Mlim*. It supposes that the accumulation is proportional to the mass still to be accumulated before reaching the maximum *Mlim*, which is equivalent to the *Daccu*/*Dero*.

Equation 7 represents the evolution of the available pollutant mass during storm weather period. It is supposed that the eroded mass is proportional to the available mass and to the

* Suspended Solid pollutograph represents the profile of SS *C(t)* concentration during time *t*

rainfall intensity. The erosion model depends on two parameters: the erosion coefficient *Wero* and a coefficient *w* [13].

$$C(t) = \frac{1}{q(t)} \cdot \frac{dMa(t)}{dt} \text{ and } \frac{dMa(t)}{dt} = -Wero \cdot I(t)^{w} \cdot Ma(t) \quad (7)$$

Where *C(t)* (mg/l) is the SS concentration produced by erosion, *q(t)* is the discharge (m3/s) at the outlet of the watershed at time t, and *I(t)* is the rainfall intensity (mm/hr).

5.3. Results

12,000 iterations were performed with the Metropolis algorithm, and the first 2,000 samples generated were removed allowing the Chain to "forget" the initial parameter set. Results showed that the Chain converged successfully to the same posterior probability distribution of the parameters regardless of the initial parameter set used. However, the speed of convergence has been found to be sensitive to the variance of the transition distribution. In the present case we chose a value of the standard deviation equal to 1/15 of the prior value of parameter to ensure the convergence.

5.3.1. Marais catchment scale

Figure 2 represents the confidence intervals of the model's output obtained by applying Monte Carlo to the model with the estimated posterior distribution of parameters. In the present case, the range of the possible responses is very large. The value of the estimated variance of errors (σ = 130mg/l), which is quite large compared to the variance of the data (σ_{data} = 150 mg/l), indicates that the variation in the measured pollutographs are considered as randomness in regard to the predictive capacity of this calibrated model. Obviously, the proposed model seems to be unable to reproduce accurately the measured pollutographs, and the Metropolis results indicate clearly that it is not due to calibration problems.

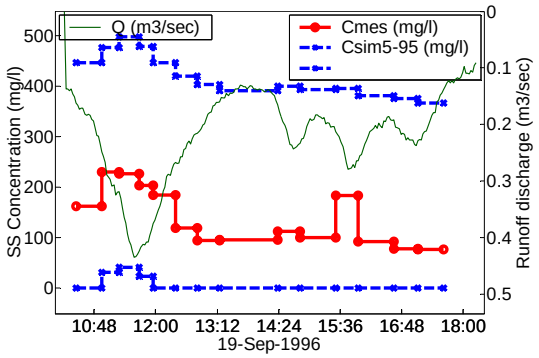


Figure 2 5-95% prediction intervals of the SS concentration at the Marais catchment scale

This is not surprising regarding the experimental results showed by Gromaire [2] where the deposits in combined sewer systems contribute to 60% of pollution. The complexity of sediments' deposition, erosion and transport processes in sewers make the sub-catchment scale by far outside the domain of validity of the conceptual model used. Thus, it seems important to apply the MCMC method for the calibration of this model on a space scale having an acceptable range of conformity to the model's domain of validity.

5.3.2. Street Surface scale

Figure 3 presents the posterior probability distribution obtained for the parameters *Daccu*, *Dero*, *Wero*, *w* and for the standard deviation of errors σ with the Metropolis algorithm.

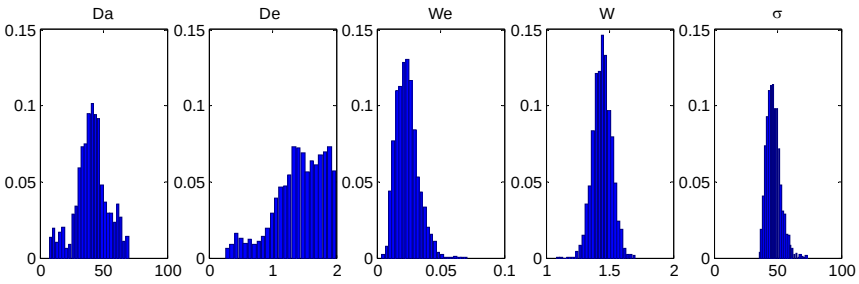


Figure 3 Posterior distribution of the 4 parameters estimated at the street catchment using Eq. 5

The analysis of the posterior distributions of the parameters shows large uncertainties related to the dry weather model parameters *Daccu* and *Dero* (Figure 4). We also found a linear correlation between these two parameters (correlation = 0.7). This correlation is due to the mathematical formulation of the accumulation model (Eq. 5). As a consequence, the accumulation model could be better calibrated if mathematically reformulated.

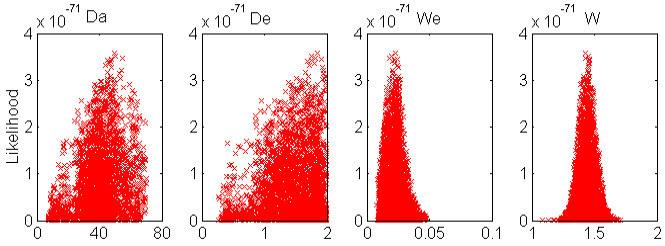


Figure 4 Scatter plot of the Likelihood measure vs. the parameters at the street catchment using Eq. 5

However, despite that the results obtained for the reformulated model (using Eq. 6) show a better identification of the maximum mass accumulated *Mlim* as shown in Figure 5, calibration results indicate a large uncertainty related to the parameter *Kaccu* representing (like the parameter *Dero*) the speed of the accumulation process during dry weather.

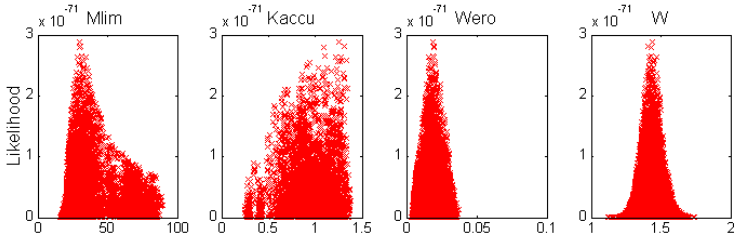


Figure 5 Scatter plot of the Likelihood measure vs. the parameters at the street catchment using Eq. 6

50 replicates of the 200 LH samples are used to estimate the importance measures of the parameters for the likelihood of the model's output for the two used models (Figure 6). Results show that the maximum accumulated mass *Mlim* represents an important parameter that has a significant impact on the likelihood measure of the model. However, the *Kaccu* parameter has an insignificant effect on the model's output. This conclusion is also provided using the scatter plot of the likelihood measure vs. the parameters as shown in Figure 4 & 5.

One can conclude that the estimation of the initial accumulated stock available before the rain event is very essential for the good performance of the model. However, the sensitivity analysis results indicate clearly that using the length of the dry weather period as an explicative parameter for the accumulation process, described by an asymptotic behavior, is not sufficient to explain the variability of the available mass just before the rain event.

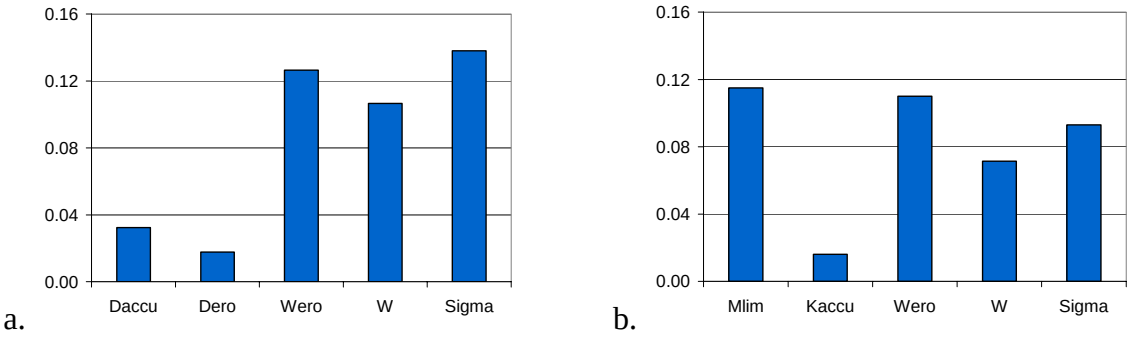


Figure 6 Importance measures for the likelihood measure of the model output using a. Eq 5 & b. Eq 6

Nevertheless, calibration results indicate a clear correlation between the maximum mass *Mlim* and the erosion parameter *Wero* (Figure 7.a.). Such correlation is not surprising regarding the mathematical structure of the erosion model (Eq. 7), which represents a multiplicative form of *Ma(t)* and *Wero*.

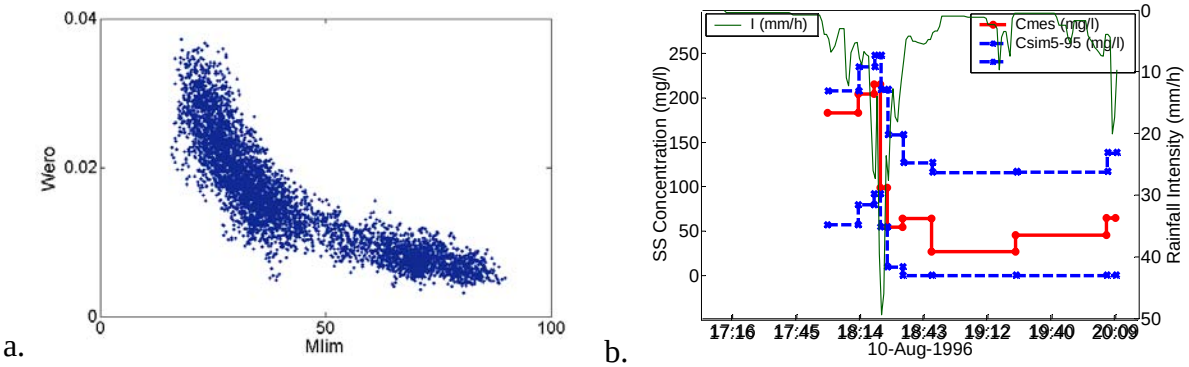


Figure 7 a. Correlations between *Mlim* and *Wero*. b. 5-95% prediction intervals of the pollutants concentration simulated by model

Figure 7.b. presents the confidence interval of the model's output *C(t)*. It shows large uncertainties in the model's predictions. This is not surprising regarding the fact that an important part of this uncertainty is attributable to the value of the variance of errors (σ = 47mg/l) which is quite large compared to the variance of the data (σ_{data} = 62mg/l). In other words, the predictive power of the calibrated model is low.

6. CONCLUSION

In this paper, we tested the applicability and effectiveness of a method used for model calibration/validation/sensitivity analysis in urban runoff quality modeling. This method, based on the MCMC sampling technique, consists of a combination of the Metropolis algorithm and a Variance based method. Metropolis algorithm provides an estimation of the posterior distributions describing parameters' uncertainties, as well as, their interaction structure. On the basis of the parameters' distributions, the Monte Carlo method determines the conceptual model's confidence intervals reflecting its prediction capacity. Using the posterior distribution, the performance of the replicated LHS method in regard to the likelihood measure leads to the quantitative identification of the main parameters that drive the model to have best fit to data.

Calibration results demonstrate that the tested conceptual model seems unable to represent the complexity of the system at the scale of urban sub-catchments. However, the application of the method to calibrate the model on a street surface scale shows that the mathematical concept of the accumulation model, using two parameters *Daccu* and *Dero*, contains linear interaction between its parameters, and implies much more uncertainty in their calibration. Furthermore, despite that a reformulation of this model using two parameters (*Mlim* and *Kaccu*) allows a better identification of the parameter *Mlim*, sensitivity analysis results show that the parameter *Kaccu* provides negligible contribution to the likelihood variation, or in other words, have no significant effect on the behavior of the model. This hypothesis casts doubts on the utility of using an asymptotic behavior, which depends only on the length of the dry weather period to describe the accumulation process. Such a conclusion needs to be validated on other sites to test its generality.

However, this method delivers much information, which would have been unreachable with classical calibration methods, and which are very useful for modeling attempts.

ACKNOWLEDGMENTS

Authors gratefully acknowledge the financial support of the “Réseau Génie Civil et Urbain” and the “Syndicat Interdépartemental pour l’Assainissement de l’Agglomération Parisienne”. We also would like to thank the Joint Research Center for providing the SimLab software used for the computation of importance measures.

REFERENCES

1. M. Ahyerre, G. Chebbo, B. Tassin,& E. Gaume (1998). Storm water quality modeling, an ambitious objective?. *Wat. Sci. Tech.*, vol. 37, No. 1, pp. 205 - 213.
2. M.C. Gromaire. *La pollution des eaux pluviales urbaines en réseau d’assainissement unitaire, caractéristiques et origines*. PhD thesis, ENPC, France, 1998
3. Q. Duan, S. Sorooshian, & V. Gupta (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. *Wat. Res. Res.*, vol. 28, No. 4, pp. 1015 - 1031.
4. G. Kuczera (1997). Efficient subspace probabilistic parameter optimization for catchment models. *Wat. Res. Res.*, vol. 33, No. 1, pp. 177 - 185.
5. G. Kuczera & E. Parent (1998). Monte Carlo assessment of parameter inference in catchments models: The Metropolis algorithm. *Journal of Hydrology*, 211, pp. 69 – 85
6. K.J. Beven & A.M. Binley (1992). The future of distributed models: model calibration and predictive uncertainty. *Hydrol. Processes*, 6, pp. 279 – 298
7. C.P. Robert & G. Casella (1999). *Monte Carlo statistical methods*. Springer Press. 1999
8. M. Tanner. *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. Springer Press, 1996
9. M.D. Mckay (1996) *Variance-Based Methods for Assessing Uncertainty Importance in NEUREG-1150 Analyses*, LA-UR-96-2695, pp. 1-27
10. R.L. Iman and W.J. Conover (1982). A distribution free approach to inducing rank correlation among input variables. *Comm. Statist.* B11(3), pp. 311-334
11. M. Ratto, S. Tarantola and A. Saltelli (2001). Sensitivity analysis in model calibration: GSA-GLUE approach. *Comp. Phys. Comm.* 136, pp. 212-224
12. Alley W. & Smith P. (1981). Estimation of accumulation parameters for urban runoff quality modeling. *Wat. Res. Res.*, vol. 17, No. 6, pp. 1657 – 1664.
13. Huber W., Heaney J., Nix S., Dickinson R. & Polmann D. *SWMM3 user’s Manual*. Technical report, EPA, 1981

Global Sensitivity Analysis: An Introduction

Andrea Saltelli

European Commission, Joint Research Centre of Ispra, Italy
andrea.saltelli@jrc.it

Abstract: This presentation aims to introduce global sensitivity analysis (SA), targeting an audience unfamiliar with the topic, and to give practical hints about the associated advantages and the effort needed. To this effect, we shall review some techniques for sensitivity analysis, including those that are not global, by applying them to a simple example. This will give the audience a chance to contrast each method's result against the audience's own expectation of what the sensitivity pattern for the simple model should be. We shall also try to relate the discourse on the relative importance of model input factors to specific questions, such as "Which of the uncertain input factor(s) is so non-influential that we can safely fix it/them?" or "If we could eliminate the uncertainty in one of the input factors, which factor should we choose to reduce the most the variance of the output?" In this way, the selection of the method for sensitivity analysis will be put in relation to the framing of the analysis and to the interpretation and presentation of the results. The choice of the output of interest will be discussed in relation to the purpose of the model based analysis. The main methods that we present in this lecture are all related with one another, and are the method of Morris for factors' screening and the variance-based measures. All are model-free, in the sense that their application does not rely on special assumptions on the behaviour of the model (such as linearity, monotonicity and additivity of the relationship between input factor and model output). Monte Carlo filtering will be also be discussed to demonstrate the usefulness of global sensitivity analysis in relation to estimation.

Keywords: global sensitivity analysis, factor prioritisation, main effects, second-order interaction effects, nonlinear models

INTRODUCTION

The material in this presentation is taken from a primer on global sensitivity analysis entitled "*Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*" by Andrea Saltelli, Stefano Tarantola, Francesca Campolongo and Marco Ratto. This will appear with John Wiley & Sons by early 2004, and we shall refer to it as to Saltelli et al., 2004 in the following. The primer aims at guiding a non-expert user in the choice of the method to adopt for the user own problem. The methods recommended include the variance based measures, the method of Morris, and Monte Carlo filtering, e.g. some effective methods for global sensitivity analysis.

Global sensitivity analysis is the study of how the uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the

model input". Global could be an unnecessary specification here, were it not for the fact that most analysis met in the literature are local or one-factor-at-a-time.

All models have use for sensitivity analysis. Applications worked by the Joint Research Centre group for Applied Statistics include: Atmospheric chemistry (Campolongo et al., 1999a), transport emission modelling, fish population dynamics (Campolongo et al. 1999b), composite indicators (Tarantola et al. 2002), portfolios, oil basins models (Saltelli, 2002), capital adequacy modelling (for Basle II), macroeconomic modelling, radioactive waste management (Saltelli and Tarantola, 2002). Applications from several practitioners can be found in Saltelli et al. Eds. 2000, a multi-author book.

Prescriptions have been issued for sensitivity analysis of models when these used for policy analysis.
In Europe, the European Commission recommends sensitivity analysis in the context of the extended impact assessment guidelines and handbook (2002). Similar recommendation in the United States EPA's White Paper on model use acceptability (1999)

The EC handbook for extended impact assessment, a working document by the European Commission, 2002, states: "A good sensitivity analysis should conduct analyses over the full range of plausible values of key parameters and their interactions, to assess how impacts change in response to changes in key parameters". The EPA paper (1999) is less prescriptive, but insists on the need for uncertainty and sensitivity analysis.

Even leaving prescriptions aside, one cannot ignore that models have not escaped the post-modern critique of the role of science in society. Specific critiques of simulation modelling and model validation have been frequent in recent years. One example: <<*…most simulation models will be complex, with many parameters, state-variables and non linear relations. Under the best circumstances, such models have many degrees of freedom and, with judicious fiddling, can be made to produce virtually any desired behaviour, often with both plausible structure and parameter values.*>>, Hornberger and Spear 1981.

Also, from within the modelling community reminders of the problem were frequent: Konikov and Bredehoeft, 1992, proclaims: "*Groundwater models cannot be validated*". This cry of alarm was taken up by Oreskes et al. 1994, in an article on Science entitled "*Verification, Validation and Confirmation of numerical models in the earth sciences*", both works focusing on the impossibility of model validation. Two established laboratory, IIASA and RIVM, had considerable trouble with the perceived quality of their models, see Mac Lane 1989, and van der Sluijs 2002 respectively. The post-modern French thinker Jean Baudrillard (1990) presents 'simulation models' as unverifiable artefact which, used in the context of mass communication, produce a fictitious hyper realities that annihilate truth. Science for the post modern age is discussed in Funtowicz and Ravetz 1990, 1993, 1999, mostly in relation to Science for policy use, a settings which Gibbons (1994) calls "mode 2" scientific production.

Faced with these critiques, the modelling community may consider that a quality check as that which is provided by a careful sensitivity analysis is worth its effort.

Before we discuss the methods for sensitivity analysis, we would like to say a few words about the output Y of interest. In our experience, the target of interest should not be the model output per se, but the question that the model has been called to answer. To make an example, if a model predicts contaminant distribution over space and time, it is the total area where a given threshold is exceeded at a given time which would play as output of interest, or the total health effects per time unit.

One should seek from the analyses conclusions of relevance to the question put to the model, as opposed to relevant to the model, e.g.

- Uncertainty in emission inventories [in transport] are driven by variability in driving habits more than from uncertainty in engine emission data.
- In transport with chemical reaction problems, uncertainty in the chemistry dominates over uncertainty in the inventories.
- Engineered barrier count less than geological barriers in radioactive waste migration.

This remark on the output of interest clearly applies to model use, not to model building, where the analyst might have interest in studying a variety of intermediate outputs.

FIRST EXAMPLE: THE OBVIOUS TEST CASE

We move now to a self-evident problem, to understand the methods as applied to it. This is a simple linear form:

$Y = \sum_{i=1}^{r} \Omega_i Z_i$

Y is the output of interest (a scalar), Ω_i are fixed coefficients, Z_i are uncertain input factors distributed as

$Z_i \sim N(\bar{z}_i, \sigma_{Z_i}), \quad \bar{z}_i = 0, \quad i = 1,2,...r$.

Y will also be normally distributed with parameters:

$\sigma_Y = \sqrt{\sum_{i=1}^{r} \Omega_i^2 \sigma_{Z_i}^2}$

$\bar{y} = \sum_{i=1}^{r} \Omega_i \bar{z}_i$

To make our point we stipulate as additional assumptions:

$\sigma_{Z_1} < \sigma_{Z_2} < .... < \sigma_{Z_r}$

$\Omega_1 > \Omega_2 > .... > \Omega_r$

According to most of the existing literature, SA should be done by taking derivatives, such as: $S_{Z_i}^d = \frac{\partial Y}{\partial Z_i}$, which would give for our model of Y: $S_i^d = \frac{\partial Y}{\partial Z_i} = \Omega_i$.

Hence the factors' ordering by importance would be $Z_1 > Z_2 > ... > Z_r$, based on our previous assumption that $\Omega_1 > \Omega_2 > .... > \Omega_r$, and this in spite of the fact that $\sigma_{Z_1} < \sigma_{Z_2} < .... < \sigma_{Z_r}$. This would seem to suggest that if our purpose is to rank input factors in terms to their contribution to the variability of the output, then simple derivatives such as $S^d_{Z_i} = \frac{\partial Y}{\partial Z_i}$ are not the best instrument to use.

A better measure could a normalised derivative of the type: $S^{\sigma}_{Z_i} = \frac{\sigma_{Z_i}}{\sigma_Y}\frac{\partial Y}{\partial Z_i}$, which, applied to our model, gives $S^{\sigma}_{Z_i} = \Omega_i \frac{\sigma_{Z_i}}{\sigma_Y}$

Comparing this with our previous expression $\sigma_Y = \sqrt{\sum_{i=1}^{r} \Omega_i^2 \sigma_{Z_i}^2}$, we obtain $\sum_{j=1}^{r} \left(S^{\sigma}_{Z_i}\right)^2 = 1$.

This is a nice result: the terms add to 1, and each of them gives the fractional contribution of the factor to the variance of the output. Unfortunately this only works for linear models.

If we want to tackle nonlinear models as well, we have to abandon derivatives and move into "exploration" of the input factors space, e.g. via Monte Carlo.

We generate a sample

$$\mathbf{M} = \begin{matrix} z_1^{(1)} & z_2^{(1)} & ... & z_r^{(1)} \\ z_1^{(2)} & z_2^{(2)} & ... & z_r^{(2)} \\ ... & ... & ... & ... \\ z_1^{(N)} & z_2^{(N)} & ... & z_r^{(N)} \end{matrix}$$

and run our computer program estimating the corresponding model output

$$\mathbf{y} = \begin{matrix} y^{(1)} \\ y^{(2)} \\ ... \\ y^{(N)} \end{matrix}$$

A natural thing to do at this point is to regress the y' s on the z_i's to obtain a regression model

$y^{(i)} = b_0 + \sum_{i=1}^{r} b_{Z_i} z_i^{(i)}$, where asymptotically $\hat{b}_0 \cong 0, \hat{b}_{Z_i} \cong \Omega_i$, $i = 1,2,...r$. Most regression packages will already provide the regression in terms of standardised regression coefficients $\hat{\beta}_{Z_i} = \hat{b}_{Z_i}\sigma_{Z_i}/\sigma_Y \cong \Omega_i \sigma_{Z_i}/\sigma_Y$. Comparing $\hat{\beta}_{Z_i} \cong \Omega_i \sigma_{Z_i}/\sigma_Y$ with

$S^{\sigma}_{Z_i} = \Omega_i \frac{\sigma_{Z_i}}{\sigma_Y}$,

it is easy to conclude that for linear models $\beta_{Z_i} = S^{\sigma}_{Z_i}$.

In summary, $\sum_{j=1}^{r} \left(S^{\sigma}_{Z_i}\right)^2 = \sum_{j=1}^{r} \left(\beta_{Z_i}\right)^2 = 1$, but only for linear models. Yet the regression coefficients are better than the derivatives in several respects.

Although for nonlinear models $\sum_{j=1}^{r} \left(\beta_{Z_i}\right)^2 \leq 1$, at least we now know how much linear the model is. This is given by the model coefficient of determination $R^2_y = \frac{\sum_{i=1}^{N} \left(\hat{y}^{(i)} - \bar{y}\right)^2}{\sum_{i=1}^{N} \left(y^{(i)} - \bar{y}\right)^2}$.

We now know that we can decompose a fraction R^2_y of the model variance using the β_{Z_i}. Furthermore the coefficients β_{Z_i} offer a measure of sensitivity that is multi-dimensionally averaged, unlike the $S^{\sigma}_{Z_i}$. For linear model this does not matter but it does, and a lot, for nonlinear ones. The drawback is when $R^2_y << 1$; typically R^2_y can be zero or near it for non-monotonic models.

In summary, we like the idea of decomposing the variance of the model output according to source (the input factors), but would like to do this for all models, independently from their degree of linearity or monotonicity. We would like a model-free approach.

In order to get there, we take a somehow twisted path and start asking ourselves the question: If I could determine the value of an uncertain factor, e.g. one of our Z_i and thus fix it, how much would the
variance of the output decrease? E.g. imagine the true value is z_i^* and hence we fix Z_i to it obtaining a "reduced" conditional variance: $V\left(Y \middle| Z_i = z_i^*\right)$. There are two problems with this quantity being a good measure of sensitivity. First I do not know where to fix the factor, and secondly for nonlinear model one could have $V\left(Y \middle| Z_i = z_i^*\right) \geq V(Y)$.

This difficulty can be overcome by averaging this measure over the distribution of the uncertain factors obtaining $E\left(V\left(Y \middle| Z_i\right)\right)$, or $E_{Z_i}\left(V_{\mathbf{Z}_{-i}}\left(Y \middle| Z_i\right)\right)$ where we have made explicit the variables over which mean and variance operators are applied. This measure has the property that $E\left(V\left(Y \middle| Z_i\right)\right) \leq V(Y)$ always, and in particular
$E\left(V\left(Y \middle| Z_i\right)\right) + V\left(E\left(Y \middle| Z_i\right)\right) = V(Y)$, where the term $E\left(V\left(Y \middle| Z_i\right)\right)$ is called a residual, and the term $V\left(E\left(Y \middle| Z_i\right)\right)$ is known as the first order effect of Z_i on Y. A nice property of the main

effect is that it is large when a factor is influential. Furthermore it is easy to verify that for linear models $S_{Z_i} = \frac{V\left(E\left(Y|Z_i\right)\right)}{V(Y)} = \beta_{Z_i}^2$.

We have made a real progress, as while $\sum_{j=1}^{r}\left(\beta_{Z_i}\right)^2 = 1$ only holds for linear models, $\sum_{j=1}^{r}\left(S_{Z_i}\right) = 1$ holds for a much larger class of models: that of the additive models. For non-additive models, $\sum_{j=1}^{r}\left(S_{Z_i}\right) \leq 1$, which is also a way to define non-additive models. Yet the measure S_{Z_i} is very useful for <u>all</u> models, as it provides a rigorous answer to a precise sensitivity analysis setting: <u>setting FP</u>, for factors prioritisation. Let us then make a digression here, and describe this setting.

FACTORS' PRIORITISATION (FP) SETTING

Imagine that I must bet on a factor that, once "discovered" in its true value and fixed, would reduce the most V(Y). Of course I do not know where the true values are for the factors, hence I cannot compare the $V\left(Y\middle|Z_i = z_i^*\right)$ for the various factors. Hence the best choice I can make is, by definition, to choose the factor with the highest $V\left(E\left(Y\middle|Z_i\right)\right)$ or , which is the same, the highest $S_{Z_i} = \frac{V\left(E\left(Y|Z_i\right)\right)}{V(Y)}$, whether the model is additive or not (Saltelli and Tarantola, 2002).

To complete all this, we must say something about non-additive model treatment, so let us complicate our model $Y = \sum_{i=1}^{r}\Omega_i Z_i$ by allowing both the Ω_i and Z_i to be uncertain, i.e. $Z_i \sim N\left(\bar{z}_i, \sigma_{Z_i}\right)$, $\bar{z}_i = 0$, $i = 1,2,...r$ as before and $\Omega_i \sim N\left(\bar{\omega}_i, \sigma_{\omega_i}\right)$, $\bar{\omega}_i = ci, i = 1,2,...r$, where *c* is a constant greater than zero (note: if the mean of the Ω_i were also null as that of the Z_i , then the model would be fully non-additive, as we shall see in a moment).

Our set of uncertain input factors is now $\mathbf{X} \equiv (\Omega_1, \Omega_2, ...\Omega_r, Z_1, Z_2, ...Z_r)$. We start crunching number estimating the sensitivity measures and we obtain the following results:

All S_{Ω_i} are zero.

All S_{Z_i} are > zero.

S_{Ω_i} is zero because the distribution of Z_i is centred in zero, and hence for any fixed value ω_i^* of Ω_i

$E\left(Y\middle|\Omega_i = \omega_i^*\right) = 0$, and *a fortiori* $V\left(E\left(Y\middle|\Omega_i\right)\right) = 0$.

Given that $\sum_{j=1}^{r}\left(S_{Z_i}\right)\le 1$ where is the remaining variance? To find it out we must compute sensitivity indices on more than one factor. If we do that, we find that $\frac{V\left(E\left(Y\middle|Z_i,Z_j\right)\right)}{V_Y}=S_{Z_i}+S_{Z_j}$, while, instead: $\frac{V\left(E\left(Y\middle|\Omega_i,Z_i\right)\right)}{V_Y}>S_{\Omega_i}+S_{Z_i}$. The difference $S_{\Omega_i Z_i}=\frac{V\left(E\left(Y\middle|\Omega_i,Z_i\right)\right)}{V_Y}-S_{\Omega_i}-S_{Z_i}$ is the second order (or two-way) effect of the two factors. We have discovered that our model is additive with respect to S_{Z_i}, S_{Z_j}, and non-additive with respect to S_{Ω_i}, S_{Z_i}.

Adding all the non-zero first order terms and all the non-zero second order terms gives back 1, i.e. 100% of the variance of Y is accounted for.
I.e. $\sum_{i=1}^{r}S_{Z_i}+S_{\Omega_i Z_i}=1$

For our model, all other terms of whatever order (1,2,3…2r) is zero. In general, if k is the total number of independent factors, then $\sum_i S_i+\sum_i\sum_{j>i}S_{ij}+\sum_i\sum_{j>i}\sum_{l>j}S_{ijl}+....S_{12...k}=1$ (Sobol', 1993).

It is quite rare that in practical applications one computes all terms in the development above. The number of terms grows exponentially with k.

We are customarily happy with computing all the S_i plus a full set of synthetic terms called S_{Ti} which give for each factor X_i, the effect of all terms including that factor.

What are the total effect terms S_{Ti} and why do we need them? Let us compute one of them, by starting with the measure $\frac{V\left(E\left(Y\middle|\mathbf{X}_{-\Omega_i}\right)\right)}{V_Y}=\frac{V\left(E\left(Y\middle|\Omega_1,\Omega_2,...\Omega_{i-1}\Omega_{i+1},...\Omega_r Z_1,Z_2,...Z_r\right)\right)}{V_Y}$. We have taken factor Ω_i as an example. Analogy with previous formulae should suggest that, by definition, this is the [first order] effect of all-but- Ω_i. Hence $S_{T\Omega_i}\equiv 1-\frac{V\left(E\left(Y\middle|\mathbf{X}_{-\Omega_i}\right)\right)}{V_Y}$ will be the effect of all terms [any order] that include Ω_i; for our model this is simply $S_{T\Omega_i}=S_{\Omega_i}+S_{\Omega_i Z_i}$, provided we remember that the S_{Ω_i} are zero as well, so that $S_{T\Omega_i}=S_{\Omega_i Z_i}$. Note that because of an algebraic relation already mentioned

$1-\frac{V\left(E\left(Y|\mathbf{X}_{-\Omega_i}\right)\right)}{V_Y}=\frac{E\left(V\left(Y|\mathbf{X}_{-\Omega_i}\right)\right)}{V_Y}$., so that the right hand expression is often used for the S_{Ti}.

There is a considerable symmetry between the S_i and S_{Ti}. Both indices can be computed in a single
shot at the cost of about N(k+2) simulations, where N is between 100 and 1000, to give an idea. In Saltelli, 2002, we use an extension of the method of Sobol', 1993. Both indices can also be computed using the Fourier based FAST method, as extended in Saltelli et al., 1999.

Furthermore S_i is ideal for factor prioritisation setting, already described, while S_{Ti} is ideal for the "factors fixing" setting (of which more in a moment).

A nice property of S_{Ti} is that if one is desperate for less expensive simulations, a rough estimate of these can be obtained via the method of Morris, at less than 1/10 of the cost, see Morris 1991. (We prefer to compute a "modulus" version of the test statistics, as described in Chapter 4, Campolongo et al., in Saltelli et al. Eds., 2000).

Finally one last useful property of variance based methods is their application "by groups", e.g.
$S_{\mathbf{\Omega}}+S_{\mathbf{Z}}+S_{\mathbf{\Omega},\mathbf{Z}}=1$, where $\mathbf{\Omega}=\Omega_1,\Omega_2,...\Omega_r$. The computational cost of this is just 3N. Or I can regroup as $\sum_{i=1}^{r}S_{\mathbf{A}_i}=1$, where $A_i=(\Omega_i,Z_i)$. The computational cost of this is kN. Note that in this latter expression all higher order terms are zero because there are interactions only within $A_i=(\Omega_i,Z_i)$.

Although in the first regrouping we save a lot in terms of model execution, and in the second we don't, there might be reasons other than economy to regroup factors. I might want to groups factors in different submodels. In this way, if I can fix all factors in the submodels may be I can skip the submodel altogether. I might want to separate controllable factors from uncontrollable ones, and so on.

A SECOND EXAMPLE: WHAT CAN SENSITIVITY OFFER FOR PARAMETER ESTIMATION

Let us now move to an estimation/calibration problem for a computational model with six parameters. We do not know how the model is done – imagine it is a computer code. The output of interest *Y* is a measure of likelihood is obtained after comparing the model prediction Y' with data, e.g.

Y=exp(-[sum of squared residuals of the predicted Y' versus the data]).

How can we characterise the good parameter set for calibration? A scatter plots of log-likelihood (e.g. of the sum of scores) vs. parameters is not very informative (Figure 1). Even "filtering", e.g. taking the best outcomes, those with the highest log-likelihood, leaves us in the dark (Figure 2). Plotting the factors value for the input (Figure 3) as well as for the input corresponding to the best values (Figure 4) is likewise noninformative. Note that if we computed on the filtered input factors (Figure 4) the pairwise correlation coefficients we would obtain zeros. Also Principal Component Analysis would not be informative as applied to the filtered input sample, as there are no correlations among the filtered factors. Computing the first order sensitivity indices for the log-likelihood and the second order ones (Figure 5), a story starts to emerge; there are non-zero second order effects, but only within the closed groups involving factors (1,2,3) and (4,5,6). Computing the third order effect (Figure 6) again only those pertaining to (1,2,3) and (4,5,6) are non-zero. Regrouping and adding the terms up gives an interesting result:

$$S^c_{123} = S_1 + S_2 + S_3 + S_{12} + S_{13} + S_{23} + S_{123} = 0.5$$
$$S^c_{456} = S_4 + S_5 + S_6 + S_{45} + S_{46} + S_{56} + S_{456} = 0.5$$

where we have used the supescript c symbol to denote the effects closed within the indices. The variance of the problem is characterised by two groups of three factors. Higher term orders are zero.

This leads the investigator to conclude that what could be reasonably estimated are two unknown functions of two parameter sub-sets. We can now reveal that the unknown function, our computer program, was the sum of two speres.

$$f(X_1,...,X_6) =$$
$$= -\left(\sqrt{X_1^2 + X_2^2 + X_3^2} - R_1\right)^2 / A_1 - \left(\sqrt{X_4^2 + X_5^2 + X_6^2} - R_2\right)^2 / A_2$$

Were the investigator to identify this structure, by trial and error, he/she would conclude that all that estimation can provide are the two radiuses.

This concludes our illustration of sensitivity analysis as applied to a diagnostic setting, and we would now like to come back to our discussion of the settings for sensitivity analysis.

MORE ON THE SETTINGS FOR SENSITIVITY ANALYSIS

We have already mentioned that the sensitivity measure of the first order, $S_i \equiv \frac{V\left(E\left(Y|X_i\right)\right)}{V_Y}$ is the ideal measure for factor prioritisation. It is also easy to see that the total effect measure $S_{Ti} \equiv \frac{E\left(V\left(Y|\mathbf{X}_{-i}\right)\right)}{V_Y}$ is appropriate for a setting that we could call

"Factors Fixing": Can I fix a factor [or a subset of input factors] at any given value over their range of uncertainty without reducing significantly the output variance? If factor X_i is totally non-influential, then all the variance is due to $\mathbf{X}_{-i}$, and fixing this vector results in $V\left(Y|\mathbf{X}_{-i}\right)=0$. It is easy to see that the reverse is also true so that necessary and sufficient condition for X_i to be totally non-influential is $S_{Ti}\equiv 0$.

Other settings that we have found useful are the following.

Factors mapping: Which factor is mostly responsible for producing realisations of Y in the region of interest? This can be treated with Monte Carlo Filtering and related tools (described elsewhere at this workshop).

Variance cutting: Reducing the variance of the output of a prescribed amount fixing the smallest number of factors. This setting can be dealt with using a combination of the S_i and S_{Ti} measures (Saltelli and Tarantola, 2002).

Why do we need settings? One way in which a sensitivity analysis can go wrong is because its purpose is left unspecified or vague (e.g. "find the most important factors"). One throws different statistical tests and measures to the problem and obtains different factors rankings. What can then be concluded? Models can be audited and settings for sensitivity analysis can be audited as well. For this reason we believe that importance must be defined beforehand.

A FEW MORE COMMENTS ON PRACTICES

What else can go wrong in a sensitivity analysis? Two instances come to mind:

There are too many outputs of interest, as we discussed at the beginning. What is the question asked from the model? Is the model relevant to the question? The optimality of a model must be weighted with respect to the task, according to a current mode of thinking. According to Beck et al. 1997, a model is "relevant" when its input factors actually cause variation in the model response that is the object of the analysis. Model "non-relevance" could flag a bad model, or a model used out of context (e.g. a gun to kill a fly). Excess complexity could also be used to silence or to fend off criticism from stakeholders, e.g. in environmental assessment studies.

Patchy or piecewise sensitivity (performed by sub-model, or one possible model at a time, or one factor at a time): Not only conflicts with the requirement of focus just mentioned, but leads to a dangerously incomplete exploration of the uncertainties; interactions are overlooked. All uncertainties should be explored simultaneously. Also the procedure of fixing non-influential factors should be conducted in this way, as fixing factors based on their first order effect can be dangerous as discussed above. The Ω_i of our initial example all have first order equal zero.

A posteriori sensitivity: Once an analysis has been produced, its revision via sensitivity analysis by a third party is not something most modellers will willingly submit to. Sensitivity analysis should be used in the process of model development, prior and within model use in analysis.

One should never forget that an unpleasant (or pleasant, depending from the viewpoint) feature of sensitivity analysis is that it might falsify the analysis altogether, e.g. by showing that the model cannot answer the question given the uncertainties, or that the model is irrelevant, or that the variation in the output of interest (e.g. a contamination level in an estuary) is insensitive to the available policy options given the uncertainties. A nice example that shows how SA can falsify a model as applied to a policy issue is described in Chapter 20, Tarantola et al., of Saltelli et al., Eds. 2000.

CONCLUSIONS

We can itemise our main conclusions as follows. There is an increased need, scope and prescription for quantitative uncertainty and sensitivity analyses. Methods are mature for use, e.g. in terms of literature, software, computational cost, tested practice, ease of communication.

In spite of this one observes a “slow start” of quantitative methods in practical analyses

Variance based measure are concise, easy to understand and to communicate, reduce to the elementary test (the standardised regression coefficients $\beta_{i_i}^2$) for linear model, relate to the popular method of Morris.

We also like and use methods in the MC filtering family.

Whatever the method one uses, we think it important that the framing of the analysis be defensible and meaningful to its users.

Figure 1. Log-likelihood for the six input factors.

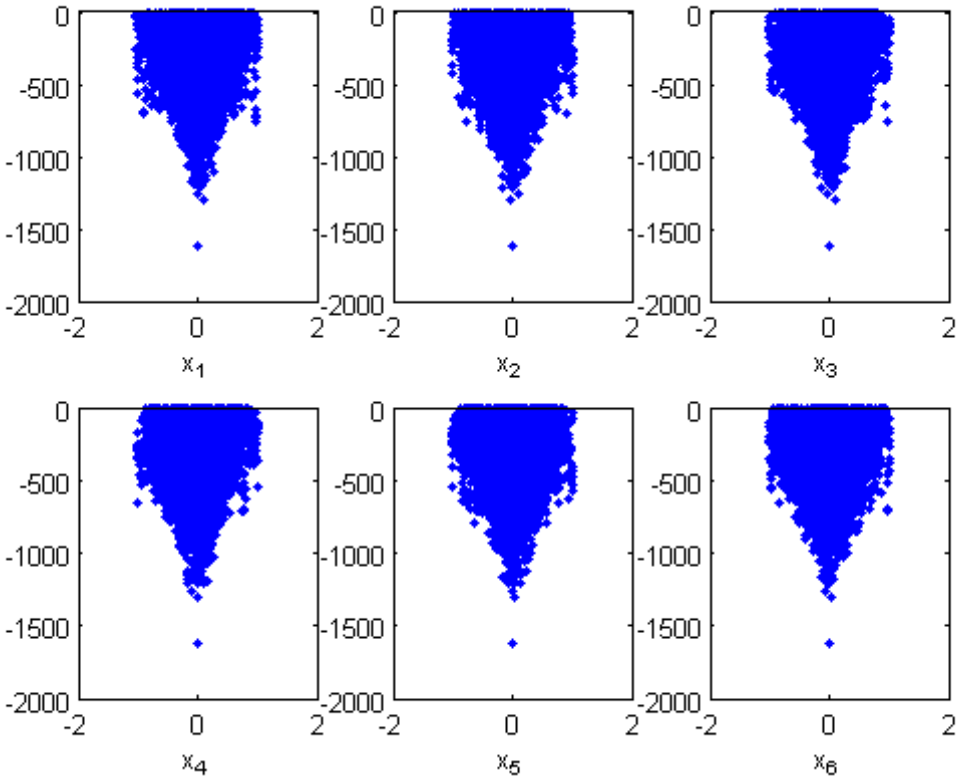


Figure 2. Same as Figure 1, for values of log-likelihood > -200.

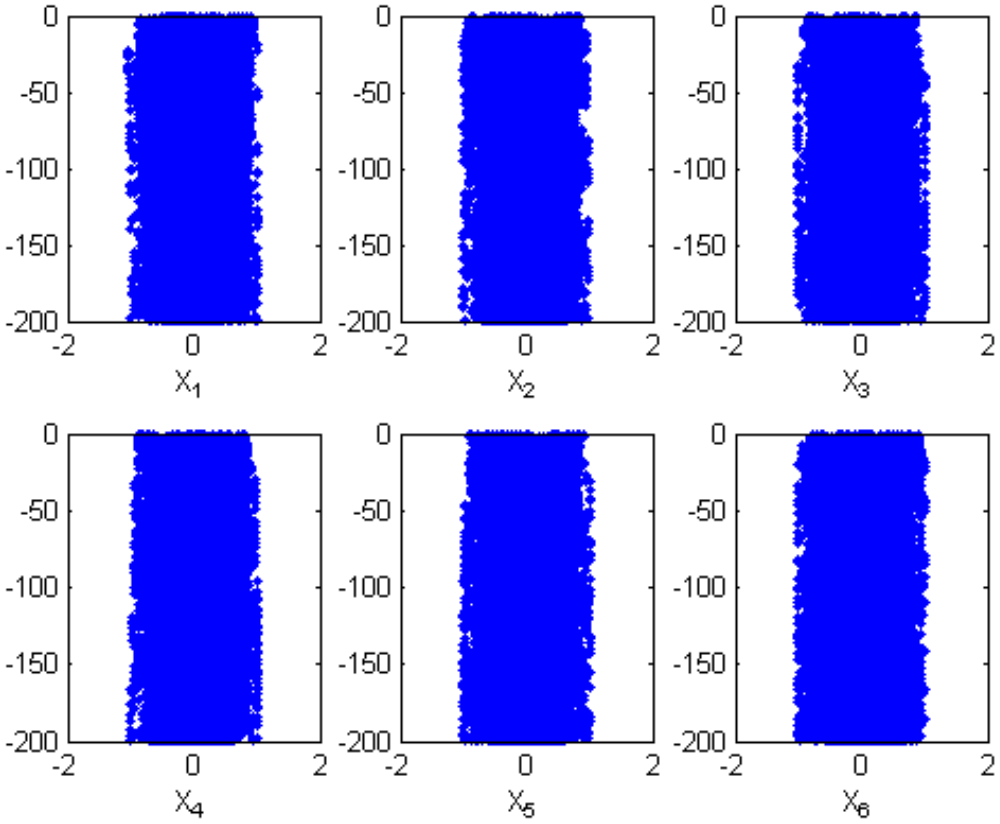


Figure 3. Pair-wise scatter plots of input factors.

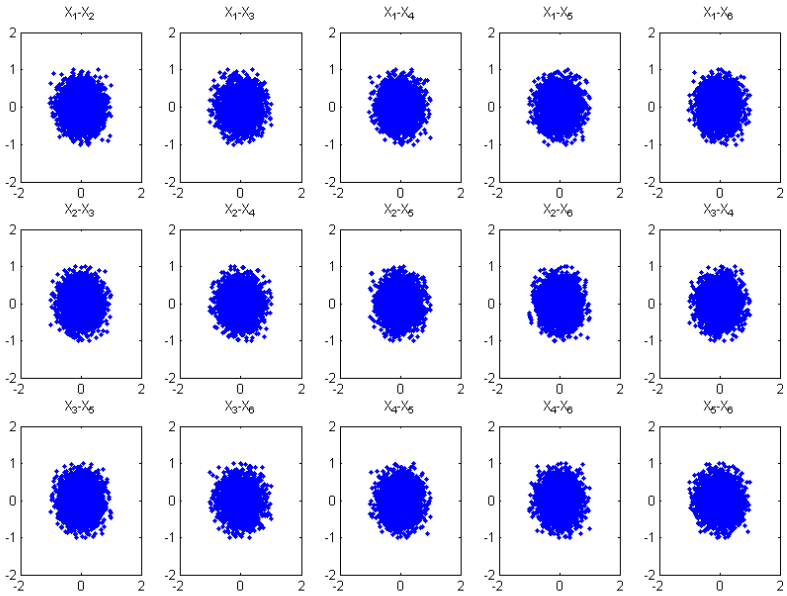


Figure 4. Same as the previous figure, for values of log-likelihood > -200.

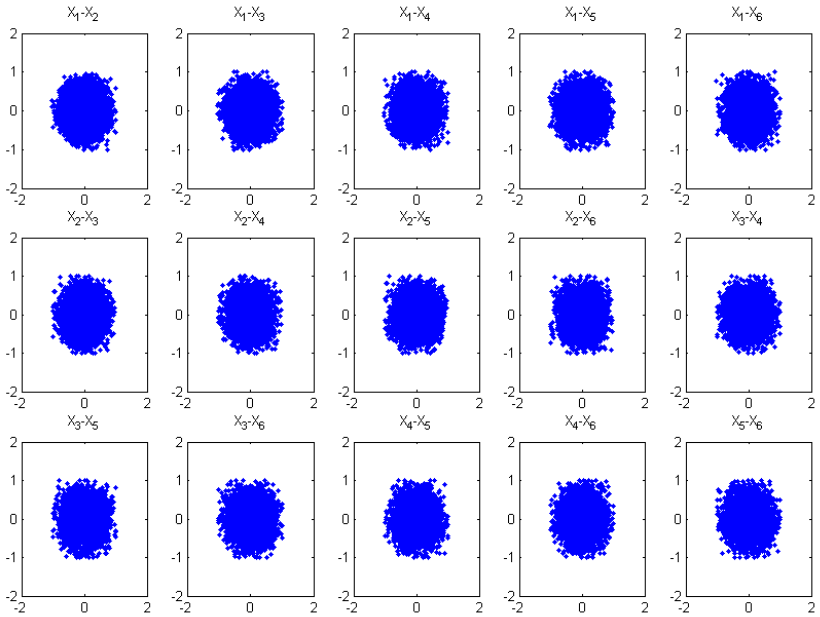


Figure 5. First- and second-order sensitivity indices for the log-likelihood.

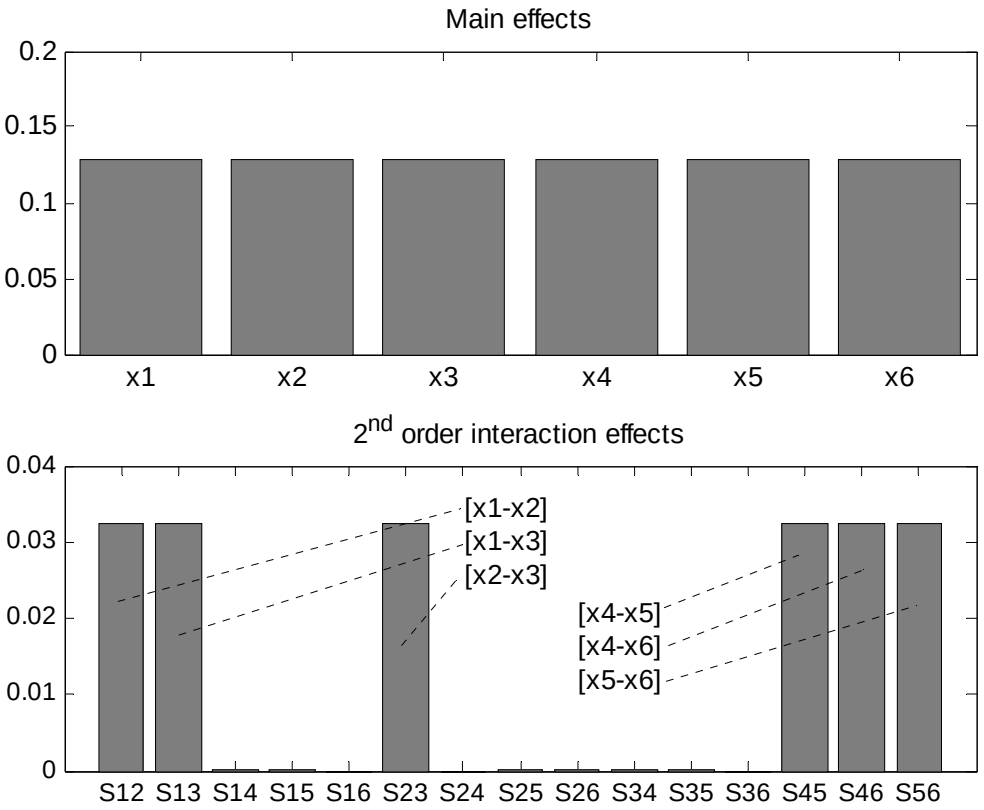
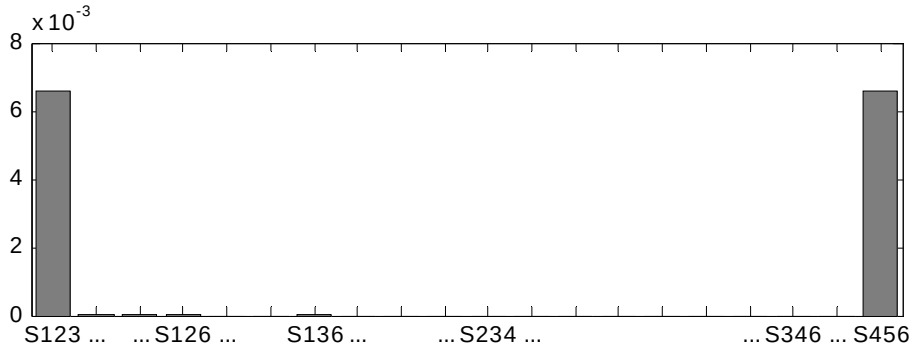


Figure 6. Third-order sensitivity indices for the log-likelihood.



NOTES

The Joint Research Centre distributes freely the software SIMLAB for uncertainty and sensitivity analysis. More information from stefano.tarantola@jrc.it. Marco Ratto (marco.ratto @jrc.it) has developed a set of scripts in Matlab to run global sensitivity analysis in diagnostic settings (e.g. with filtering plus variance based methods, see our two-sphere example). This is also available.

A forum to discuss sensitivity analysis issues is available at http://sensitivity-analysis.jrc.cec.eu.int/.
It includes a FAQ section, introduction to the main methods and a bibliography.

REFERENCES

Baudrillard, J., 1990, The Revenge of the Crystal, London, Pluto Press.

Beck, M. B., J. R. Ravetz, L. A. Mulkey and T. O. Barnwell, 1997, On the problem of model validation for predictive exposure assessments, *Stochastic Hydrology and Hydraulics*, **11**, 229-254.

Campolongo, F., Saltelli A,. Jensen, N.R., Wilson, J., and Hjorth, J., 1999a, The role of multiphase chemistry in the oxidation of dimethylsulphide (DMS). A latitude dependent analysis, *Journal of Atmospheric Chemistry*, **32**, 327-356.

Campolongo, F., Tarantola, S., and Saltelli, A., 1999b, Tackling quantitatively large dimensionality problems, *Computer Physics Communications*, **117**, 75-85.

Crosetto, M. and Tarantola, S., 2001 Uncertainty and Sensitivity Analysis: Tools for GIS-based Model Implementation, *International Journal of Geographic Information Science*, **154**, 415-437.

EC, 2000, European Commission's Communication on Extended Impact Assessment Brussels, 05/06/2002 COM(2002) 276 final. Guidelines for implementing the directive are available at the Governance page of the EC, http://europa.eu.int/comm/governance/docs/index_en.htm.

EPA, 1999, The US Environmental Protection Agency Science Policy Council, White Paper on the Nature and Scope of Issues on Adoption of Model Use Acceptability Guidance, http://www.epa.gov/osp/crem/library/whitepaper_1999.pdf.

Funtowicz, S. O., and J. R. Ravetz. (1990). Uncertainty and Quality in Science for Policy. Kluwer Academic, Dordrecht, The Netherlands.

Funtowicz S, and Ravetz J.R., 1993, Science for the Post-Normal Age, *Futures*, **25**, 735-755.

Funtowicz S, and Ravetz J.R., 1999, Post-Normal Science - an insight now maturing, *Futures*, **31**, 641-646.

Gibbons, M., 1994, The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. London: Sage Publications.

Hornberger G. M., and R. C. Spear, 1981, “An approach to the preliminary analysis of environmental systems”, *Journal of Environmental management*, **12**, 7-18.

Konikov L.F. and Bredehoeft J. D., 1992, Groundwater models cannot be validated, *Advances in Water Resources*, **15**(1), 75-83.

Leamer E. E., 1990, “Let's take the con out of econometrics” and “Sensitivity Analysis would help”, in Modelling Economic Series, C.W.J. Granger, Ed., Oxford, Clarendon Press.

Lemons, J., K. Shrader-Frechette, and C. Cranor, “The precautionary principle: scientific uncertainty and type I and II errors”, *Foundations of Science*, **2**, 207-236, 1997.

Mac Lane, S., 1988, *Science*, Letters, **241**, 1144 and **242**, 1623,1624.

Oreskes, N., K. Shrader-Frechette, K. Belitz, 1994, Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences, *Science*, **263**, 641-646.

Ratto, M., Tarantola, S. and A. Saltelli 2001, Sensitivity analysis in model calibration: GSA-GLUE approach, *Computer Physics Communications*, **136**, 212-224.

Saltelli, A., S. Tarantola, K. Chan, 1999, A quantitative, model independent method for global sensitivity analysis of model output, *Technometrics*, **41** (1), 39-56.

Saltelli, A., K. Chan and M. Scott, Eds., 2000, Handbook of Sensitivity Analysis, John Wiley & Sons publishers, Probability and Statistics series.

Saltelli, A., 1999, Sensitivity analysis. Could better methods be used?, *Journal of Geophysical Research*, **104**(D3), 3789-3793. N.B. see also Errata Corrige for a table’s error on the same journal, 1999, **104**(D19), 24,013.

Saltelli, A. (2002). Making best use of model valuations to compute sensitivity indices. *Computer Physics Communications*, **145**, 280-297.

Saltelli, A., Tarantola, S. Campolongo, F., 2000, Sensitivity analysis as an ingredient of modelling, *Statistical Science*, **15**(4), 377-395.

Saltelli A. Tarantola S., 2002, On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal, *Journal of American Statistical Association*, **97** (459), 702-709.

Saltelli A. Tarantola S., 2004, Campolongo, F. and Ratto, M., Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models, John Wiley & Sons publishers, Probability and Statistics series, 2003 (to appear).

SOBOL' I.M., 1993, Sensitivity Estimates for Nonlinear Mathematical Models, *Mathematical Modelling and Computational Experiments*, **1**(4) 407-414

Stirling, A., 1999, On science and precaution in the management of technological risk, Vol I: A synthesis report of case studies, JRC-IPTS/ESTO, May 1999, EUR 19056 EN.

Stirling, A., Multi Criteria Mapping, in Foster J. Ed., Valuing Nature, Routledge 2002.

Tarantola, S., J. Jesinghaus, M. Puolamaa, 2000, Global sensitivity analysis: a quality assurance tool in environmental policy modelling, in Saltelli, et al., eds, 2000, 385-397.

Tarantola S., Saisana M., Saltelli A., Schmiedel F. and N. Leapman (2002) Statistical techniques and participatory approaches for the composition of the European Internal Market Index 1992-2001, European Commission, EUR 20547 EN.

van der Sluijs, J. P., 2002, A way out of the credibility crisis of models used in integrated environmental assessment, *Futures* **34**, 133–146.

Sensitivity Analysis of the e-Business Readiness Composite Indicator

S. Tarantola, M. Nardo, A. Saltelli, I. Kioutsioukis and R. Liska

Joint Research Centre, IPSC, Applied Statistics Sector, TP 361, Ispra(VA), 21020, Italy

E-mail: stefano.tarantola@jrc.it

Abstract: The initiative "e-Readiness of European enterprises" is part of the European Commission "e-Europe 2005" action plan. As part of this initiative, the European Council of ministers has requested the compilation of a composite indicator to assess the preparedness of the internet business environment of European Countries. Underlying data for the component indicators have been obtained through enterprise surveys. The Joint Research Centre, as part of the European Commission, has been asked to carry out a pilot study on this composite indicator for the year 2002. The study includes the testing for robustness and sensitivity, as recommended by the European Commission guidelines for impact assessment. We illustrate here the uncertainty and sensitivity tests that have been carried out for this pilot study.

Keywords: e-business, robustness, uncertainty, weights, imputation.

1. INTRODUCTION

Composite indicators are weighted combinations of selected sub-indicators into single aggregated measures, via underlying models of the policy domains of interest. Discussion on the legitimacy of such indicators is incessant. Composites are increasingly used by media and policy makers to communicate information on the situation of countries or regions in various policy fields such as environment, economy or technological development (reviews in [1,2]). Opponents lament that composites are mixes of dubious interpretation yet expensive to obtain. Organisms such as the UN, the OECD and the European Commission make use of such measures. In particular the OECD and the JRC have recently undertaken the joint preparation of a handbook of good practices of composite indicators building [3].

In this paper we study the construction of a composite indicator of e-business readiness (see Section 2). This composite indicator is aimed at measuring the progress of Member States towards a more extensive take up and use of digital technologies. We report part of the results of a pilot study commissioned to JRC by the Directorate General Enterprise of the European Commission. In particular, we focus our analysis on the weighting scheme used to aggregate sub-indicators, and on the sensitivity of the composite indicator to different weighting schemes and to incomplete data.

As far as weighting is concerned, JRC suggested and deployed a participatory technique, called "budget allocation", which allows any expert of a panel to express, from a policy perspective, their opinion upon the relative importance of sub-indicators (see Section 4).

The issue of sensitivity is crucial for the characterisation of composites. The Communication from the European Commission on Structural Indicators [4] recognises the importance to assess the sensitivity of the message provided by composites with respect to the weights employed. Here we consider an additional source of uncertainty in the evaluation of the composite indicator, the uncertainty due to missing data.

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

As we shall see in Section 3 we use a *Multiple Imputation* technique (based on Markov Chain Monte Carlo algorithms) for the treatment of missing data. This is appealing in that it provides confidence bounds for the imputed data [5,6]. Imputed data are, indeed, estimated values. Different imputed data may result in different values for the composite indicator. Thus their effect on the resulting composite indicator must be acknowledged using both uncertainty and sensitivity analysis.

2. THE e-READINESS COMPOSITE INDICATOR

The eEurope 2005 Action Plan [7] calls for a benchmarking of the target that 'by 2005, Europe should have (…) a dynamic e-business environment', specifying that 'e-business comprises both e-commerce (buying and selling on-line) and restructuring of business processes to make best use of digital technologies'. Besides proposing guidelines for the benchmarking, the resolution sets out a number of policy indicators to monitor progress in the implementation of the Action Plan.

One of these benchmarking indicators is the composite indicator on e-business readiness. According to the Council's recommendation, this is made of two core groups (see Table 1): (a) *Adoption of ICT* by business, and (b) *Use of ICT* by business; each group is composed by six sub-indicators.

The composite indicator, Y_c, for a given country c, is a weighted sum of k sub-indicators X_{ic} (5 available for *Adoption* and 6 for *Use of ICT*) and k weights w_i : $Y_c = \sum_{i=1}^{k} X_{ic} \cdot w_i$. The analysis is conducted using an incomplete dataset (data availability is 81%) for the year 2002. Therefore, the first step in our analysis is that of "filling up" empty spaces.

Table 1. List of sub-indicators for the composite indicator on e-business readiness

Adoption of ICT by business	
a1	% of enterprises that use Internet
a2	% of enterprises that have a web site/home page
a3	% of enterprises that use at least two security facilities at the time of the survey
a4	% of total number of persons employed using computers in their normal work routine (at least once a week)
a5	% of enterprises having a broadband connection to the Internet
a6	% of enterprises with a LAN and using an Intranet or Extranet
Use of ICT by business	
b1	% of enterprises that have purchased products / services' via the internet, Electronic Data Interchange or any other computer mediated network where these are >1% of total purchases
b2	% of enterprises that have received orders via the internet, Electronic Data Interchange or any other computer mediated network where these are >1% of total turnover
b3	% of enterprises whose IT systems for managing orders or purchases are linked automatically with other internal IT systems
b4	% enterprises whose IT systems are linked automatically to IT systems of suppliers or customers outside their enterprise group
b5	% of enterprises with Internet access using the internet for banking and financial services
b6	% of enterprises that have sold products to other enterprises via a presence on specialised internet market places

3. MULTIPLE IMPUTATION OF MISSING DATA

3.1 Methodology

We do not attempt any imputation for countries and sub-indicators that are totally missing. Therefore, Belgium, France, The Netherlands, Portugal and the sub-indicator *a3* were not included in the analysis (see Table 2).

The explanation of the MCMC-based technique for multiple imputation is given in Refs. 5 and 6. The technique yields, simultaneously, one estimate for each of the 22 missing data. This process is repeated M=50 times. Against the prevailing practice of using, for each of the 22 cells, the mean over the M individual estimates, we use their full distribution in our study. The gray values in Table 2 are the medians of the (normal) distributions. The sample means and standard deviations, calculated over the M values, are given in Table 3. The dataset in Table 2 is the starting point for the calculation of the composite indicator.

Table 2. Data set for the e-business readiness composite indicator. The50-th percentiles of the distribution of the imputed values are marked in grey. NA stands for 'not available'.

2002	a1	a2	a3	a4	a5	a6	b1	b2	b3	b4	b5	b6
B	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
DK	95%	76%	NA	59%	56%	53%	24%	12%	5%	9%	68%	2%
D	84%	66%	NA	51%	28%	45%	39%	16%	11%	11%	55%	1%
EL	64%	34%	NA	43%	3%	25%	7%	6%	7%	8%	39%	1%
E	82%	38%	NA	34%	45%	31%	3%	1%	3%	9%	64%	0%
F	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
IRL	83%	53%	NA	42%	9%	41%	23%	11%	6%	10%	57%	3%
I	74%	46%	NA	42%	15%	39%	3%	3%	1%	1%	38%	0%
L	78%	51%	NA	44%	18%	69%	22%	11%	5%	10%	42%	1%
NL	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
A	85%	64%	NA	51%	29%	50%	27%	17%	6%	12%	58%	2%
P	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
FIN	96%	69%	NA	57%	44%	51%	30%	13%	6%	14%	81%	2%
S	95%	80%	NA	72%	33%	62%	31%	16%	8%	16%	71%	2%
UK	74%	67%	NA	57%	20%	39%	18%	12%	14%	18%	45%	1%

4. SELECTION OF WEIGHTS

A rather common way to assign weights is to involve experts opinion. In the *budget allocation* method [8], each expert is given a "budget" of 100 points, and is asked to distribute the budget over the sub-indicators by allotting more points to those indicators which are felt as more important. For each sub-indicator, the average weight across the experts (last row in Tables 4 and 5) is used in the aggregation procedure.

Table 3. Mean and standard deviation of the 22 imputed data over M values

Indicator	Country	Mean	Standard deviation
a4	FIN	53%	11%
a5	D	32%	31%
b1	S	38%	16%
b1	UK	20%	11%
b2	UK	10%	5%
b3	DK	16%	5%
b3	D	13%	9%
b3	E	10%	5%
b3	L	12%	8%
b3	A	12%	8%
b3	FIN	16%	9%
b3	S	13%	8%
b4	DK	20%	8%
b4	D	16%	7%
b4	E	14%	8%
b4	L	8%	4%
b4	A	11%	4%
b4	FIN	19%	8%
b4	S	16%	8%
b5	UK	56%	9%
b6	DK	1%	8%
b6	UK	2%	11%

The budget allocation method was employed at the steering group meeting of the e-business support network (e-BSN), held in Paris on October, 28th, 2003. Fourteen experts coming from the European Union and the Accession Countries were involved in the exercise. The sets of weights obtained, for each core group, are given in Tables 4 and 5.

Contrarily to the common use of average weights, where the information from the single expert vanishes, we believe it is important to retain the identity of the experts and acknowledge, in our model of composite indicator, the uncertainty to due expert selection.

Table 4. Results of the budget allocation exercise for 'Adoption of ICT'. Data for a3 not available. Weights originally attributed to indicator a3 have been equally distributed over the remaining 5 indicators, and re-scaled so that the sum is 100.

Expert's Nationality	a1	a2	a4	a5	a6	SUM
UK	-	15	15	35	35	100
NL	-	20	20	50	10	100
LV	35	35	10	20	-	100
L	30	25	30	15	-	100
DK	25	25	25	25	-	100
SL	-	30	20	30	20	100
F	-	25	25	25	25	100
LT	-	10	20	40	30	100
IRL	-	-	31.2	50	18.8	100
N	-	-	42.9	35.7	21.4	100
S	11.2	-	-	44.4	44.4	100
HU	16.7	16.7	16.7	22.2	27.7	100
EL	-	15	25	30	30	100
E	40	40	10	10	-	100
Average	**11.7**	**19.2**	**20.3**	**30.4**	**18.4**	

Table 5. Results of the budget allocation exercise for 'Use of ICT'.

Expert's Nationality	b1	b2	b3	b4	b5	b6	SUM
UK	15	15	35	35	-	-	100
NL	10	30	-	30	-	30	100
LV	35	30	20	15	-	-	100
L	25	25	-	-	25	25	100
DK	25	25	25	25	-	-	100
SL	40	20	-	-	30	10	100
F	25	25	25	25	-	-	100
LT	-	10	-	30	20	40	100
IRL	15	15	30	40	-	-	100
N	15	35	35	15	-	-	100
S	10	30	-	40	-	20	100
HU	5	10	20	20	20	25	100
EL	20	-	30	30	20	-	100
E	40	-	40	-	20	-	100
Average	**20**	**19.3**	**18.6**	**21.8**	**9.6**	**10.7**	

5. UNCERTAINTY ANALYSIS

Given the variability of expert selection, and the uncertainty coming from the imputation of the missing data, the composite indicator for the different countries is also affected by uncertainty. We have carried out the following tests:

- uncertainty analysis to assess how the variability in the weights and the uncertainty in the imputed data influence the composite indicator of e-readiness;
- sensitivity analysis of the composite indicator to assess how much uncertainty is due to choice of weights and how much to imputation errors. This is helpful to know whether collecting more data permits drawing more accurate inferences.

The variability in expert selection has been accounted for by considering a trigger factor ω, i.e. a discrete random variable uniformly distributed between 1 and 14 (the number of experts). For example, for ω=7 the expert from France is chosen (see Tables 4 and 5).

The uncertainty coming from imputation of missing data depends on how many imputations have been done for a given country. For example, for Denmark three imputations have been made. Therefore, we define one uncertain factor for each imputed data. The factors are normal distributions with means and standard deviations given in Table 3. For Denmark we have four uncertain factors; for Italy only one uncertain factor (ω), hence no sensitivity analysis can be carried out. Let Y be the composite indicator for a given country:

$$Y = w_1 a_1 + w_2 a_2 + w_3 a_3 + w_4 a_4 + w_5 a_5 + w_6 a_6 + w_7 b_1 + w_8 b_2 + w_9 b_3 + w_{10} b_4 + w_{11} b_5 + w_{12} b_6$$

where $\omega \equiv (w_1, w_2, ..., w_{12})$ is the set of weights proposed by a given expert (a given row in Tables 4 and 5). For UK, for example, the composite has five sources of uncertainty: ω, b1, b2, b5 and b6; for Greece the composite has only one source of uncertainty: ω.

Different countries have different (number of) uncertainty sources; this implies that the uncertainty analysis is carried out independently for each country. For each country, a LHS sample of size *N =1500* is generated for the uncertain factors based on their distributions, and the composite indicator is evaluated N times. Figure 1 displays the empirical distributions of composite indicators for the eight countries that had both uncertainty on weights and on data. The other countries, Greece, Ireland, and Italy have a complete dataset, thus the uncertainty analysis is a histogram with 14 bins, one for each expert in the budget allocation exercise.

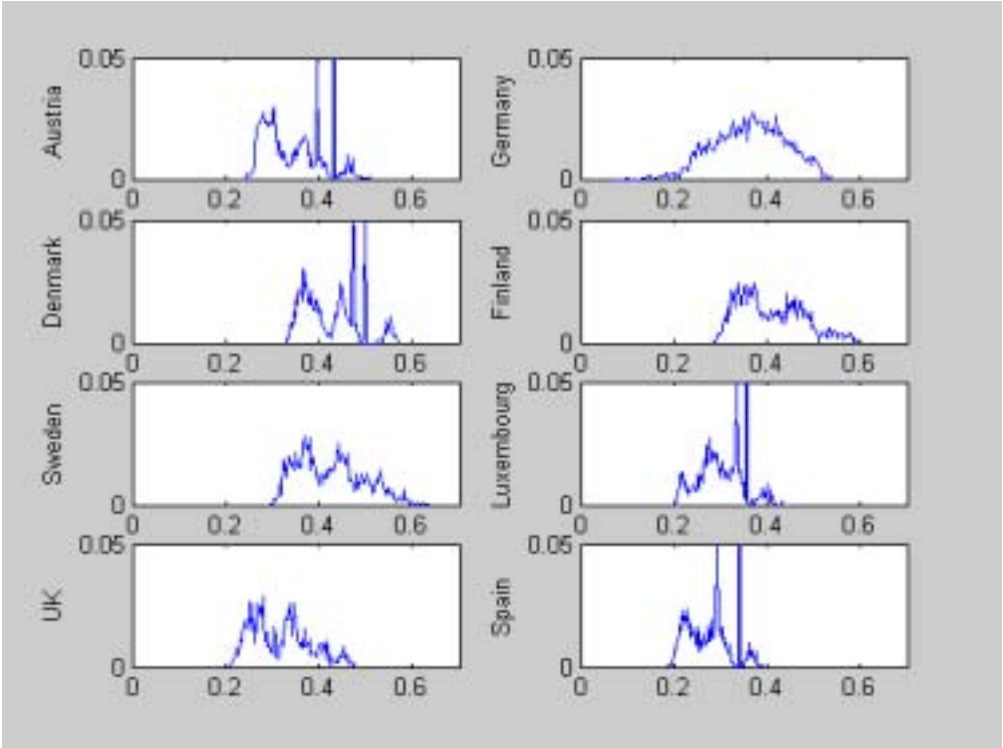


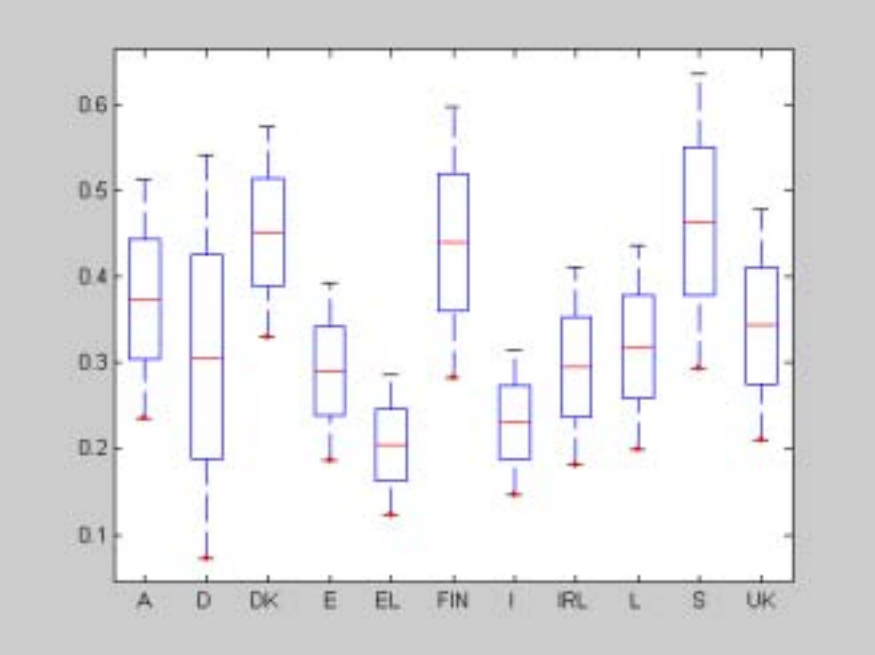
Figure 1. Uncertainty analysis of the composite indicator of e-business readiness for eight countries, based on a LHS sample of size 1,500. On the horizontal axis the values of the composite indicator; on the vertical axis their frequency of occurrence.

For Austria, b3 and b4 are imputed, and experts 4 (from Luxembourg) and 6 (from Slovenia) give zero weight to both b3 and b4. This causes the presence of two peaks for Austria (the left one due to expert 4, the right one due to expert 6). Similar peaks occur for Luxembourg, Denmark, and Spain. Figure 1 displays multi-modal distribution functions for most countries. Each modal function is the result of the convolution of particular combinations of weights with uncertainty in imputation. Discrete distributions are obtained for Greece, Ireland and Italy. While for Ireland uncertainty on weights does not favour any particular output value, for Greece and Italy medium and low values respectively of e-business are more likely.

Figure 2 displays the composite indicator of e- readiness with its confidence bounds for all countries in terms of box-plots. Sweden and Italy have non overlapping bounds: the policy inference is robust, no matter uncertainty in weights or in data. When the box plots of two countries overlap, the degree of uncertainty determines the relative score of the countries

considered. Spain, Greece, and Italy unambiguously have lower e-readiness than Denmark, Finland and Sweden. Germany overlaps with almost all other countries pointing to the crucial effects played by uncertainty in weights and data (mainly a5, see Table 3) for this country.

Figure 2. Box plots of the composite indicator on e-business readiness. Uncertainty is due to different weighting schemes as well as to imputation of missing data.



Also Sweden and Finland have large boxes. For Greece, Spain and Italy the boxes are narrower, indicating a less important role for uncertainty in data and weights. For UK, with 4 indicators imputed (b1, b2, b5, b6), the box is not so large (very low weight associated to b5 and b6). How much of the uncertainty in the composite indicators is due to different weighting schemes rather than to imputation of missing data values? This will be the object of the sensitivity analysis.

6. SENSITIVITY ANALYSIS

6.1 Methodology

The question answered by the sensitivity analysis is how much of the uncertainty on the composite indicator Y for a given country depends on the uncertainty on its input factors **X** (imputed data and weight selection). Using the variance of Y as yardstick of importance, the issue becomes, how much does the prediction variance, V(Y), decrease, on average, when some components of **X** are held fixed. The starting point of the variance-based methods is the variance decomposition $V(Y)=V(E(Y|X_i))+E(V(Y|X_i))$, where X_i is any uncertain factor. The first order sensitivity indices can be calculated as $S_i=V(E(Y|X_i))/V(Y)$, for each uncertain factor. The higher S_i, the higher the importance of X_i, as the larger the average drop in variance $V(Y)$ obtained when fixing X_i within its range.

The method used here to evaluate the sensitivity indices is a generalisation of that proposed in [9] (a review is also offered in [10]) at no extra cost for the analysis. We illustrate the generalisation briefly here. The first order indices are calculated by:

$$\hat{S}_j = \frac{\hat{U}_j - \hat{E}^2(Y)}{\hat{V}(Y)} \qquad (1)$$

where Monte Carlo estimates for U_j, $E(Y)$ and $V(Y)$ are computed as:

$$\hat{U}_j = \frac{1}{n-1}\sum_{r=1}^{n} f(x_{r1}, x_{r2},..., x_{rk}) f(x'_{r1}, x'_{r2},..., x'_{r(j-1)}, x_{rj}, x'_{r(j+1)},..., x'_{rk}) \tag{2}$$

$$\hat{E}^2(y) = \left\{\frac{1}{n}\sum_{r=1}^{n} f(x_{r1}, x_{r2},..., x_{rk})\right\}^2 \tag{3}$$

$$\hat{V}(y) = \frac{1}{n}\sum_{r=1}^{n} f(x_{r1}, x_{r2},..., x_{rk})^2 - \frac{1}{n^2}\left(\sum_{r=1}^{n} f(x_{r1}, x_{r2},..., x_{rk})\right)^2 \tag{4}$$

Let us simplify the notation by setting:

$$y_A = f(x_1, x_2,..., x_k) = f(A) \qquad y_C = f(x_{r1}', x_{r2}',...x_{r(j-1)}', x_{rj}, x_{r(j+1)}',...x_{rk}') = f(C) \tag{5}$$
$$y_B = f(x_1', x_2',..., x_k') = f(B) \qquad y_D = f(x_{r1}, x_{r2},...x_{r(j-1)}, x_{rj}', x_{r(j+1)},...x_{rk}) = f(D)$$

A and *B* are independent sample matrices; *C* and *D* are independent re-sample matrices as well. Ref. 9 suggests that, when calculating U_j as sum of products $y_A y_C$, more accurate estimates for S_j are obtained when $E^2(y)$ is based on products of independent matrices:

$$\hat{E}^2(y) = \frac{1}{n}\sum_{r=1}^{n} f(x_{r1}, x_{r2},..., x_{rk}) f(x'_{r1}, x'_{r2},..., x'_{rk}) = \frac{1}{n}\sum_{r=1}^{n} y_A y_B \tag{6}$$

Therefore, it is also legitimate to estimate $E^2(y)$ using products of $y_C y_D$, which are also independent. When (6) is employed, the denominator of (1) can then be calculated from either $y_A y_A$ or $y_B y_B$. Similarly, when $E^2(y)$ is estimated using products of $y_C y_D$, the denominator in (1) can be estimated from either $y_C y_C$ or $y_D y_D$. We end up with four sensitivity indices:

$S_j^I = \left(\sum y_A y_C - \sum y_A y_B\right) / \sum y_A y_A$	$S_j^{III} = \left(\sum y_A y_C - \sum y_C y_D\right) / \sum y_C y_C$
$S_j^{II} = \left(\sum y_A y_C - \sum y_A y_B\right) / \sum y_B y_B$	$S_j^{IV} = \left(\sum y_A y_C - \sum y_C y_D\right) / \sum y_D y_D$

Exploiting the symmetry property of the design (Ref. [9]), we obtain additional indices:

$S_j^V = \left(\sum y_B y_D - \sum y_A y_B\right) / \sum y_A y_A$	$S_j^{VII} = \left(\sum y_B y_D - \sum y_C y_D\right) / \sum y_C y_C$
$S_j^{VI} = \left(\sum y_B y_D - \sum y_A y_B\right) / \sum y_B y_B$	$S_j^{VIII} = \left(\sum y_B y_D - \sum y_C y_D\right) / \sum y_D y_D$

The indices $S_j^I, S_j^{II}, S_j^{III}, S_j^{IV}$ are positively correlated. So are the indices $S_j^V, S_j^{VI}, S_j^{VII}, S_j^{VIII}$. The two groups of indices are negatively correlated. Comparison tests between the indices used in [9] and the average of the eight estimates confirm that the convergence of this latter is generally more rapid. Same symmetry properties allow the estimation of four total indices.

7. RESULTS

We test the sensitivity of the composite indicator for each country to both weights and imputation results using the method illustrated so far. A total cost of about 1,000 model runs has been required to estimate both the (eight) first order indices and the (four) total effects for all the factors with an accuracy of 1% on the indices. The time required to execute one model run is approximately nil, as the model output is a weighted average of the input. Is uncertainty coming from imputation (other than for Greece, Italy and Ireland) more relevant than the uncertainty due to choice of weights?

Table 5. First order and total effects of uncertain factors (weights trigger and imputed indicators) on e-business readiness for eight countries. Calculations performed with the enhanced version of the method of Saltelli.

	A	D	DK	E	FIN	L	S	UK
Weights	0.97	0.56	0.95	0.91	0.91	0.95	0.91	0.94
a_4					0.03			
a_5		0.38						
b_1							0.04	0.03
b_2								0.01
b_3	0.01	0.01	0.00	0.01	0.01	0.02	0.01	
b_4	0.00	0.01	0.03	0.05	0.02	0.01	0.01	
b_5								0.00
b_6			0.00					0.00
	A	D	DK	E	FIN	L	S	UK
Weights	0.98	0.61	0.97	0.93	0.93	0.97	0.94	0.96
a_4					0.04			
a_5		0.42						
b_1							0.06	0.04
b_2								0.01
b_3	0.02	0.01	0.06	0.02	0.02	0.04	0.01	
b_4	0.00	0.01	0.08	0.07	0.03	0.02	0.02	
b_5								0.01
b_6			0.05					0.00

Table 5 shows that for all countries a large fraction of the composite indicator variability is due to the set of weights used. The uncertainty brought by weights is an implicit part of the participatory approach used to build the composite indicator. In other terms a "true value" for weights cannot exist because of different objectives, viewpoints and interests at stake; uncertainty in the composite indicator due to weights cannot be eliminated and has a visible impact on the results.

The uncertainty due to the imputation of missing data does not account for more than 5% of the e-business readiness variance for all countries but Germany. For Germany indicator a_4 accounts for 38% of the composite indicator variance. This means that, being able to find the real value of a_4, would reduce (on average) the variance of the composite indicator of 38%.

The total indices look very similar to those of the first order. This highlights the additive structure of the model. Some indicators (e.g., b_3 and b_4 for Germany; b_2, b_5 and b_6 for UK) have total effect index lower than, or equal to, 0.01. This means that it is worthless to spend resources collecting data for those indicators and those countries, because this would not help improving the accuracy of the composite indicator.

8. CONCLUSIONS

Media and policy-makers look with increasing interest at composite indicators as appealing tools to attract the attention of the community and to help focusing policy debates. But methodological gaps in their design and construction may invite politicians to draw simplistic conclusions or the press to communicate misleading information. That is why national and international organisations believe that it is important to focus on methodological issues in the design of composite indicators [3].

This study focuses on the design stage of composite indicators, where rarely robustness and sensitivity analysis are applied. Yet, quite recently, the European Commission has recognised the role of such investigation and requires the use of sensitivity analysis in the development of any new composite indicator. The Joint Research Centre supports various Directorates General of the European Commission in a number of projects that involve the development and use of composite indicators. The case of e-business readiness presented in this paper is the latest exercise carried out so far.

REFERENCES

1. M. Saisana and S. Tarantola. State-of-the-art report on current methodologies and practices for composite indicator development, *EUR 20408 EN Report*, European Commission-JRC: Italy, 2002.
2. M: Saisana, A. Saltelli and S. Tarantola
3. CINDI (2004) Web-forum on Composite INDIcators, Eurostat CIRCA web-site
4. European Commission. *Communication from the Commission on Structural Indicators* COM (2002) 551 final, 2002.
5. D. Rubin. Multiple Imputation after 18+ years (with discussion), *Journal of the American Statistical Association*, 91, 473-489, 1996.
6. J. Schaffer. *Analysis of incomplete multivariate data*, Chapman & Hall, London, 1997.
7. European Commission. *Communication from the Commission on e-Europe 2005 Action Plan* COM (2002) 263 final, 2002.
8. B. Moldan and S. Billharz. Sustainability indicators. Scope 58, Whiley Pub., 1997.
9. Saltelli, A. Making best use of model valuations to compute sensitivity indices, *Computer Physics Communications*, 145, 280-297, 2002.
10. A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*, Wiley, to appear in March 2004.

Sensitivity indices for imprecise probability distributions

Jim W. Hall[a]

[a]Department of Civil Engineering, University of Bristol, Queen's Building, University Walk, Bristol, BS8 1TR, UK
E-mail: jim.hall@bristol.ac.uk

Abstract: Conventional variance-based sensitivity indices are extended to deal with the case when information is available as closed convex sets of probability measures, a situation that exists when probability distributions are specified with interval-valued parameters. The generalization to closed convex sets of probability measures yields lower and upper sensitivity indices. An example demonstrates a numerical method for estimating these sensitivity indices.

Keywords: Variance-based sensitivity indices, coherent lower and upper probabilities

1. INTRODUCTION

The information input into computer models may be imprecise for several reasons. Imprecision is often a consequence of measurement processes, for example using digital sensors. Prior information is sometimes recorded in the literatures as intervals without any information about probability distributions [1]. Given only finite time, it is argued that it may be impossible to elicit precise probability distributions from experts [2]. Indeed experts may deliberately use imprecision to express their uncertainty.

The extension of probabilistic analysis to include imprecise information is now well established in the theory of imprecise probabilities [3], robust Bayesian analysis [4, 5] and fuzzy statistics [6]. In this paper we explore the notion of sensitivity within this framework. We confine ourselves to the theory of coherent lower and upper probabilities, which, whilst not the most general theory of imprecise probabilities, is sufficient to deal with the situation in which probability distributions are specified by interval-valued parameters.

2. COHERENT LOWER AND UPPER PROBABILITIES

Consider a probability density function $f(x, \mathbf{a})$, where $x \in \mathbb{R}$ and $\mathbf{a} = (a_1, a_2, \ldots, a_m)$, a vector of parameters of the probability density function. By definition

$$\Pr(A) = \int_A f(x, \mathbf{a}) dx, \forall A \subseteq \mathbb{R}. \tag{1}$$

If each parameter a_i in $\mathbf{a}$ is specified by a closed interval $[l_i, u_i]$ then $\mathbf{a}$ is constrained by an m-dimensional box Q, defining a closed set of probability measures that imply lower and upper probabilities, $P(\underline{A})$ and $P(\overline{A})$:

$$\Pr(\underline{A}) = \inf_{\mathbf{a} \in Q} \int_A f(x, \mathbf{a}) dx \tag{2}$$

Further author information: (Send correspondence to Jim W. Hall)
Jim W. Hall: Telephone: +44 117 928 9763

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

$$\Pr(\overline{A}) = \sup_{\mathbf{a} \in Q} \int_A f(x, \mathbf{a}) dx. \tag{3}$$

$P(\underline{A})$ and $1 - P(\overline{A})$ will be located at the same point $\mathbf{a}$, so $P(\underline{A}) = 1 - P(\overline{A})$, meaning that $P(\underline{A})$ and $P(\overline{A})$ are coherent lower and upper probabilities [7].

The lower and upper expectations, $E(\underline{X})$ and $E(\overline{X})$, are given by

$$\underline{E}(X) = \inf_{\mathbf{a} \in Q} \int_{-\infty}^{\infty} x f(x, \mathbf{a}) dx \tag{4}$$

$$\overline{E}(X) = \sup_{\mathbf{a} \in Q} \int_{-\infty}^{\infty} x f(x, \mathbf{a}) dx. \tag{5}$$

The definitions in Equations 2 to 5 can be extended to the case when $f(\mathbf{x}, \mathbf{a})$ is a joint probability distribution on $\mathbb{R}^n$ and $\mathbf{x} = (x_1, \ldots, x_n)$.

2.1. Lower and upper variance

The standard definition of the variance $V(X)$ of a random variable X is

$$V(X) = E([X - E(X)]^2). \tag{6}$$

If $\mathcal{M}$ is a closed convex set of probability measures $P : X \to [0, 1]$, then the lower and upper variances $\underline{V}(X)$ and $\overline{V}(X)$ are given by:

$$\underline{V}(X) = \min_{P \in \mathcal{M}} V(X) \tag{7}$$

$$\overline{V}(X) = \max_{P \in \mathcal{M}} V(X). \tag{8}$$

2.2. Natural extension of imprecise probabilities

Let g be a function such that $y = g(\mathbf{x}) : \mathbf{x} = (x_1, \ldots, x_n)$, and let $B_y \subseteq \mathbb{R}^n$ containing all of the points $(x_1, \ldots, x_n)$ such that $g(\mathbf{x}) \in C : C \in \mathbb{R}$, then the lower and upper probabilities $\underline{P}(C)$ and $\overline{P}(C)$ are:

$$\underline{P}(C) = \inf_{\mathbf{a} \in Q} \int_{B_y} \cdots \int f(x_1, \ldots, x_n, \mathbf{a}) dx_1 \ldots dx_n \tag{9}$$

and

$$\overline{P}(C) = \sup_{\mathbf{a} \in Q} \int_{B_y} \cdots \int f(x_1, \ldots, x_n, \mathbf{a}) dx_1 \ldots dx_n. \tag{10}$$

3. VARIANCE-BASED SENSITIVITY ANALYSIS

Consider now the conventional probabilistic case in which the uncertainties in $x_1, \ldots, x_n$ are expressed as precise probability distributions, i.e. $x_1, \ldots, x_n$ and y are replaced by random variables $X_1, \ldots, X_n$ and Y respectively. In variance-based sensitivity analysis, the first order sensitivity indices S_i represents the fractional contribution of a given variable X_i to the variance in a given output variable Y [8]. In order to calculate the sensitivity indices the total variance V in the model output Y is apportioned to all the input factors X_i as [9]

$$V = \sum_i V_i + \sum_{i<j} V_{ij} + \sum_{i<j<k} V_{ijk} + \ldots + V_{12\ldots n} \tag{11}$$

where

$$V_i = V[E(Y|X_i = x_i^*)] \tag{12}$$

$$V_{ij} = V[E(Y|X_i = x_i^*, X_j = x_j^*)] - V_i - V_j \tag{13}$$

and so on. $V[E(Y|X_i = x_i^*)]$ is the Variance of the Conditional Expectation (VCE) and is the variance over all values of x_i^* in the expectation of Y given that X_i has a fixed value x_i^*. The first order (or 'main effect') sensitivity index S_i for variable X_i is:

$$S_i = V_i/V \tag{14}$$

and the 'total effect' sensitivity index is [10]

$$S_{Ti} = 1 - \frac{V[E(Y|X_{\sim i} = x_{\sim i}^*)]}{V(Y)} \tag{15}$$

where $X_{\sim i}$ denotes all of the variables other than X_i.

4. IMPRECISE SENSITIVITY INDICES

In the case when the uncertainty in the variables $X_1 \ldots X_n$ is described by a closed convex set $\mathcal{M}$ of probability measures P, the lower and upper variances introduced in Equations 7 and 8 above can be extended to lower and upper sensitivity indices, $\underline{S}_i$ and $\overline{S}_i, i = 1, \ldots, n$:

$$\underline{S}_i = \min_{P \in \mathcal{M}} S_i \tag{16}$$

and

$$\overline{S}_i = \max_{P \in \mathcal{M}} S_i \tag{17}$$

where

$$\sum_{i=1}^{n} \overline{S}_i \leq 1. \tag{18}$$

The additional constraint in Equation 18 means that the upper sensitivity indices $\overline{S}_i$, $i = 1, \ldots, n$ may not co-exist. Indeed there is a closed convex set $\mathcal{S}$ of sensitivity indices $\mathbf{S} \in \mathcal{S} : \mathbf{S} = \{S_1, \ldots S_n\}$ constrained such that $\forall S_i, i = 1, \ldots, n : \underline{S}_i \leq S_i \leq \overline{S}_i$ and $\sum_{i=1}^{n} \overline{S}_i \leq 1$.

4.1. Numerical method

Estimating the lower and upper sensitivity indices in Equations 16 and 17 is a problem of non-linear optimization. Each iteration j of the optimization involves estimating the precise sensitivity indices for some $P_j \in \mathcal{M}$, specified by a vector of parameters $\mathbf{a}_j = (a_1, \ldots, a_m)$. For each $\mathbf{a}_j$ the corresponding precise joint probability distribution $f(\mathbf{x}, \mathbf{a}_j)$ is randomly sampled d times, yielding a precise estimate of the variance [8]:

$$\hat{V}(Y_j) = \frac{1}{d} \sum_{k=1}^{d} g^2(\mathbf{x}_k, \mathbf{a}_j) - \hat{g}_{0,j}^2 \tag{19}$$

where

$$\hat{g}_{0,j} = \frac{1}{d} \sum_{k=1}^{d} g(\mathbf{x}_k, \mathbf{a}_j). \tag{20}$$

The Monte Carlo estimate $\hat{V}_i(Y_j)$ of the ith partial variance is given by

$$\hat{V}_i(Y_j) = \frac{1}{d} \sum_{k=1}^{d} g(\mathbf{x}_{\sim i,k}^{(1)}, \mathbf{x}_{i,k}^{(1)}, \mathbf{a}_j) g(\mathbf{x}_{\sim i,k}^{(2)}, \mathbf{x}_{i,k}^{(1)}, \mathbf{a}_j) - \hat{g}_{0,j}^2 \tag{21}$$

where

$$\mathbf{x}_{\sim i,k} = (x_{1,k}, x_{2,k}, \ldots, x_{i-1,k}, x_{i+1,k} \ldots, x_{n,k}). \tag{22}$$

The superscripts (1) and (2) in Equation 21 indicate that two sampling matrices are being used for $\mathbf{x}_k$. Both matrices have dimensions $d \times n$. In computing $\hat{V}_i(Y_j)$ the values of Y_j corresponding to $\mathbf{x}_k$ from matrix (1) are multiplied by the values of Y_j computed using a different matrix (2), but for the ith column, which is kept constant [8]. This resampling yields a precise estimate of the sensitivity indices $S_{i,j}$. The lower and upper variances are then given by

$$\underline{V}(Y) = \min_j(V(Y_j)) \tag{23}$$

$$\overline{V}(Y) = \max_j(V(Y_j)) \tag{24}$$

and the lower and upper sensitivity indices are given by

$$\underline{S_i}(Y) = \min_j(S_i(Y_j)) \tag{25}$$

$$\overline{S_i}(Y) = \max_j(S_i(Y_j)), i = 1, \ldots, n \tag{26}$$

where $S_i(Y_j) = V_i(Y_j)/V(Y_j)$.

5. APPLICATION

Oberkampf et al. [11] have proposed a series of Challenge Problems to compare and evaluate alternative theories of uncertainty. One of the Challenge Problems relates to a

damped linear oscillator (a single degree of freedom mass-spring-damper system), whose steady-state magnification factor D_s is given by

$$D_s = \frac{k}{\sqrt{(k - m\omega^2)^2 + (c\omega)^2}} \tag{27}$$

where k is the spring constant, m is the mass of the oscillator, ω is the frequency of oscillation and c is the damping coefficient. In this Challenge Problem, the variables in Equation 27 were specified as follows:

m is given by a precise triangular probability distribution defined on the interval [10,12], with a median value 11.

k is given by an imprecise triangular probability distribution, specified by three imprecise parameters k_{min}, k_{mod} and k_{max}, whose values are contained in the closed intervals $k_{min} \in [90, 100]$, $k_{mod} \in [150, 160]$ and $k_{max} \in [90, 100]$.

c is given by a closed interval of possible values $c \in [5, 10]$. No probability distribution over this interval is specified or to be assumed.

ω is given by an imprecise triangular probability distribution, specified by three imprecise parameters ω_{min}, ω_{mod} and ω_{max}, whose values are contained in the closed intervals $\omega_{min} \in [2.0, 2.3]$, $\omega_{mod} \in [2.5, 2.7]$ and $\omega_{max} \in [3.0, 3.5]$.

In the Challenge Problem specification, the information concerning k and c was given by three independent sources. The problem of aggregation of evidence from multiple sources is beyond the scope of the present paper and is not addressed. The information is used from the first source only.

There are 6 interval-valued distribution parameters, k_{min}, k_{mod}, k_{max}, ω_{min}, ω_{mod}, ω_{max}, and one interval-valued variable, c, in the analysis. If the sensitivity indices S_i were a monotonic function of these imprecise quantities then it would only be necessary only to test the vertices of the 7 dimensional hypercube that contains all of the possible values of these quantities. There is, however, no reason to believe that S_i should be a monotonic function of these interval-valued quantities, so in order to find the imprecise sensitivity indices it was necessary to search the volume contained within these interval constraints. Besides testing each of the 2^7 vertices, the volume was searched by uniformly sampling the space with a total of 30000 samples. At each test point $\mathbf{a}_j = (k_{min,j}, k_{mod,j}, k_{max,j}, \omega_{min,j}, \omega_{mod,j}, \omega_{max,j}, c_j)$ (Equations 19 to 26) 50000 Monte Carlo samples were used in the sensitivity estimates.

The lower and upper upper probability distributions on D_s are shown in Figure 1. The lower and upper expectations were estimated as $\underline{E}(D_s) = 1.78$ and $\overline{E}(D_s) = 2.86$ and the lower and upper variances were estimated as $\underline{V}(D_s) = 0.09$ and $\overline{V}(D_s) = 1.57$. The imprecise sensitivity indices are listed in Table 1. Note the additional condition in Equation 18 means that the upper sensitivity indices cannot all coexist.

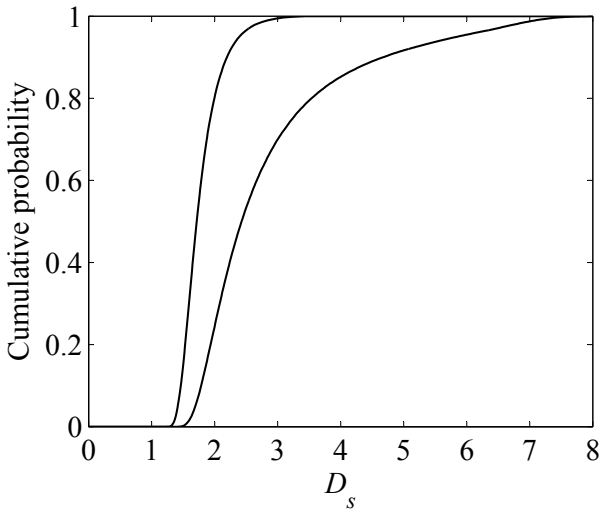


Figure 1. Lower and upper cumulative probability distributions of D_s

Table 1. Imprecise sensitivity indices

i	Variable	$\underline{S}_i$	$\overline{S}_i$
1	m	0.00	0.07
2	k	0.18	0.76
3	ω	0.19	0.70

6. CONCLUSIONS

Variance-based sensitivity indices provide an intuitive and practical expression of the contribution of model input variables to the variance in the model output [10, 12]. To date, variance-based sensitivity analysis have been restricted to the situation where uncertain information is presented as precise probability distributions, yielding precise sensitivity indices. In this paper this precise probabilistic case has been extended to the situation in which information appears as imprecise probability distributions or intervals, yielding interval-valued sensitivity indices for the (precise or imprecise) probabilistic variables. These imprecise indices complement the insights into the effects of imprecision and randomness provided by generalized uncertainty analysis [13]. A further challenge, which has not been addressed in this paper, is the problem of aggregation of imprecise and probabilistic information from multiple sources [14, 15]. Sensitivity analysis has further potential in this respect in highlighting the influence of different information sources.

The computational expense of calculating imprecise sensitivity indices is considerable. Furthermore, the advantage over Monte Carlo approaches of efficient methods for calculating variance-based sensitivity indices, such as FAST and Sobol' methods [8], is less clear

than in the precise case. Monte Carlo methods can make use of function evaluations from previous steps in the optimization to find the lower and upper sensitivity indices, whereas the FAST and Sobol' methods would usually require a new sample at each optimization step. Whilst for the example addressed in this paper little computational advantage was to be gained by reusing previous function evaluations, clearly this will be desirable in many practical situations, so methods of this type are the subject of ongoing research.

ACKNOWLEDGMENTS

Dr Hall's research is supported by a Royal Academy of Engineering Post-Doctoral Research Fellowship.

REFERENCES

1. J.W. Hall, E. Rubio, and M.J. Anderson. Random sets of probability measures in slope hydrology and stability analysis. *ZAMM: Journal of Applied Mathematics and Mechanics*, in press, 2004.
2. J. Berger. The robust bayesian viewpoin (with discussion). In J. Kadane, editor, *Robustness of Bayesian Analyses*. North-Holland, Amsterdam, 1984.
3. P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
4. P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.
5. F.R. Hempel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, Chichester, 1986.
6. R. Viertl. *Statistical Methods for Non-Precise Data*. CRC Press, Boca Raton, Florida, 1996.
7. P. Walley. Towards a unified theory of imprecise probabilities. *Int. J. Approximate Reasoning*, 24(2-3):125–148, 2000.
8. K. Chan, S. Tarantola, A. Saltelli, and I.M. Sobol'. Variance-based methods. In A. Saltelli, K. Chan, and E.M. Scott, editors, *Sensitivity Analysis*, chapter 8, pages 167–198. Wiley, Chichester, 2000.
9. I.M. Sobol'. Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling Computational Experiment*, 1:407–414, 1993.
10. T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of model output. *Reliability Engineering and Systems Safety*, 52(1):1–17, 1996.
11. W.L. Oberkampf, J.C. Helton, C.A. Joslyn, S.F. Wojtkiewicz, and S. Ferson. Challenge problems: uncertainty in system response given uncertain parameters. *Reliability Engineering and System Safe*, in press.
12. A. Saltelli, S. Tarantola, and F. Campolongo. Sensitivity analysis as an ingredient of modelling. *Statistical Science*, 15(4):377–395, 2000.
13. S. Ferson. Probability bounds analysis is global sensitivity analysis. In *this volume*, 2004.
14. C. Genest and J.V. Zidek. Combining probability distributions: A critique and annotated bibliography. *Statistical Science*, 1(1):114–148, 1986.
15. K. Sentz and S. Ferson. Combination of evidence in dempster-shafer theory. Technical Report SAND2002-0835, Sandia National Laboratories, Albuquerque, New Mexico, 2002.

Accelerated estimation of sensitivity indices using State Dependent Parameter models

Marco Ratto[a], Stefano Tarantola[a], Andrea Saltelli[a] and Peter C. Young[b]

[a]Institute for the Protection and Security of the Citizen
Joint Research Centre – European Commission
TP 361, Via E. Fermi 1, 21020 Ispra (VA), Italy;
[b]Centre for Research on Environmental Systems and Statistics
CRES/IENS,Lancaster University
Lancaster LA1 4YQ, U.K.
E-mail: marco.ratto@jrc.it, stefano.tarantola@jrc.it, andrea.saltelli@jrc.it, p.young@lancaster.ac.uk

Abstract: In this paper we use State Dependent Parameter (SDP) models (a non-parametric model estimation approach, based on recursive filtering and smoothing estimation) to estimate the main effect sensitivity indices of computational models. Especially when coupled with efficient sampling methods, such as the quasi-random LP-tau sequence, this method is extremely efficient, allowing for drastic reduction in the cost of the sensitivity analysis. Moreover, the method allows us also to estimate the first order terms of the High Dimensional Model Representation of the model under analysis, at no additional cost.

Keywords: State Dependent Parameter models, Variance based methods, High Dimensional Model Representation, Sensitivity analysis.

1. STATE OF THE ART

Consider the mathematical or computational model $Y = f(X_1, X_2, ..., X_k)$, where some of the input factors X_i are uncertain. For the non-correlated case, sensitivity indices are related to the Sobol' decomposition [1]

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + ... + V_{12...k} \tag{1}$$

where $V_i = V_{X_i}\left(E_{\mathbf{X}_{-i}}\left(Y|X_i\right)\right)$, $V_{ij} = V_{X_iX_j}\left(E_{\mathbf{X}_{-ij}}\left(Y|X_i, X_j\right)\right) - V_i - V_j$ and so on. This is also related to a decomposition of the function f itself into terms of increasing dimensionality (HDMR, [2-3]), i.e.

$$f(Y) = f_0 + \sum_i f_i + \sum_i \sum_{j>i} f_{ij} + ... + f_{12...k} \tag{2}$$

where each term is a function only of the factors in its index, i.e. $f_i = f_i(X_i)$, $f_{ij} = f_{ij}\left(X_i, X_j\right)$ and so on. The various terms can be expressed as:

$$
\begin{aligned}
&f_0 = E(Y) \\
&f_i(X_i) = E_{\mathbf{X}_{-i}}(Y \mid X_i) - f_0 \\
&f_{ij}(X_i, X_j) = E_{\mathbf{X}_{-ij}}(Y \mid X_i, X_j) - f_i(X_i) - f_j(X_j) - f_0 \\
&\ldots
\end{aligned}
\tag{3}
$$

Variance-based sensitivity measures are based on the terms in the decomposition (1), normalised by the total unconditional variance $V(Y)$:

$$
\begin{aligned}
&S_i = V_i / V, \\
&S_{ij} = V_{ij} / V, \ldots \ S_{ij}^c = V_{X_i X_j}\left(E_{\mathbf{X}_{-ij}}\left(Y \middle| X_i, X_j\right)\right) / V \\
&\ldots
\end{aligned}
\tag{4}
$$

The S_i are called the main effects, the S_{ij} are the second order pure interaction terms, while the S_{ij}^c are the called the second order closed effects, giving the overall effect of a group of two factors, i.e. for orthogonal input factors $S_{ij}^c = S_i + S_j + S_{ij}$, and so on until the closed term of order *k* equal to 1. All indices are nicely scaled in [0, 1] and, as discussed in [4] and in the Keynote lecture by A. Saltelli at this conference [5], are related to rigorous settings, applicable to different contexts for SA. In particular, main effects are related to the setting "Prioritising Factors", i.e. to identify the factor which, if determined (i.e., fixed to its true, albeit unknown, value), would lead to the greatest reduction in the variance of the target output , and so on for the second most important factor etc., till all factors are ranked.

Let us then concentrate on the main effects S_i. The classical strategy for global sensitivity analysis methods is to directly estimate the V_i terms, without passing through the elementary functions f_i, and then to normalise by *V*. These methods (FAST, Extended FAST, correlation ratios, Sobol', etc, see [4-6] for reference) are conceived as black-box methods and do not try to use information present in the Monte Carlo sample, e.g. analysing scatter plots and trying some smoothing of the pattern, if any, between to model output and a given input. So, even if they are robust, unbiased and applicable to whatever non-linear and complex computational model, they do not make the best use of all the information contained in the Monte Carlo sample. This makes such methods computationally expensive, with a required number of model evaluations that is proportional to the number of factor *k*, e.g. of at least some thousands for a good approximation of the solution. This limits the application of variance based methods to not too complex computational models, which allow the required number of model evaluations to be carried out in a reasonable time. A lot of effort has been expended in recent years to reduce the cost of the analysis, either by improving the efficiency of the available methods (see e.g. [7]), or by exploring more efficient routes, such as the Bayesian approach presented by Oakley and O'Hagan [8]. In the latter case, Bayesian tools are used to exploit the information about the input-output mapping more efficiently than classical variance based methods, thus reducing drastically the computational cost of the analysis.

In this paper, we first estimate the f_i's, using recursive filtering and Fixed Interval Smoothing (FIS) algorithms to fit SDP models to the input-output mapping [9], then we

compute the variance of f_i to estimate the main effects. As in [8] this method allows us to estimate *both* f_i, and V_i (i.e. S_i), adding valuable information to the sensitivity analysis at a much smaller computational cost than classical methods. The convergence rate is of the same order of the Bayesian approach by Oakley and O'Hagan [8] and, at the same time, the method presented here is simple, since it is based on 'classical' recursive algorithms, such as the Kalman filter [10-11] and recursive FIS.

2. THE METHOD

The present methodology exploits signal processing and time series analysis tools, in particular an approach to non-stationary and nonlinear signal processing based on the identification and estimation of stochastic models with time variable (TVP) or state dependent (SDP) parameters. The works of P.C. Young [12-13] illustrate TVP/SDP algorithms and provide full references on the subject.

Often non-stationary and nonlinear systems can be approximated well by TVP (or piece-wise linear) models, the parameters of which can be estimated using recursive methods of estimation, where parameters are assumed to evolve in a simple stochastic manner (e.g. [12-14]). When instead the changes in the parameters are functions of the state or input variables (i.e. they actually constitute stochastic state variables), then the system is truly nonlinear and likely to exhibit severe nonlinear behaviour. Normally, this cannot be approximated in a simple TVP manner; in which case, the alternative and more powerful SDP modelling methods must be used.

In SDP time series modelling, the natural ordering of the data along the time coordinate is replaced by an ordering based on the ascending value of the state variables (or inputs), making the SDP model estimation similar to 'pattern recognition', i.e. to analysing scatter plots between a model input X_j and the output Y. In the SA framework, the analyst has a set of Monte Carlo simulations from which sensitivity indices and HDMR terms have to be estimated. Nothing impedes to consider such a set of Monte Carlo model evaluations as a time series and therefore to try to apply SDP modelling to estimate the first order terms in the decomposition of the computational model given in (2).

2.1. SDP models and HDMR

The general SDARX (State Dependent Auto-Regressive with eXogenous variables) specification for a dynamical system is:

$$Y_t = \mathbf{Z}_t^T \mathbf{p}_t + e_t \qquad e_t \sim N(0, \sigma^2) \tag{5}$$

where

$$\mathbf{Z}_t^T = [-Y_{t-1}, -Y_{t-2}, \ldots, -Y_{t-n}, \mathbf{X}_{t-\delta}^T, \mathbf{X}_{t-\delta-1}^T, \ldots, \mathbf{X}_{t-\delta-m}^T]$$

$$\mathbf{X}_t^T = [X_1, \ldots, X_k]$$

$$\mathbf{p}_t = [a_1(\mathbf{Z}_t), a_2(\mathbf{Z}_t), \ldots, a_n(\mathbf{Z}_t), \mathbf{b}_0(\mathbf{Z}_t), \mathbf{b}_1(\mathbf{Z}_t), \ldots, \mathbf{b}_m(\mathbf{Z}_t)]$$

and $a_i(\mathbf{Z}_t), i = 1,\ldots,n$ $\mathbf{b}_i(\mathbf{Z}_t), i = 0,\ldots,m$ are the state dependent parameters, which are assumed to be functions of the state vector $\mathbf{Z}_t$. We keep here the time series notation, with the index *t* spanning the set of Monte Carlo simulations, $t = 1,\ldots,N$.

In the SDP modelling framework the identification of the model structure itself is a critical issue. Observations of the input and output series are available and the analyst has to identify the dynamical model which best fits the observations, in the most parsimonious way (Data-Based Mechanistic modelling, [15]).

This involves finding which and how many lags and/or delays characterise the input state variables as well as if and how many autoregressive terms of the output have to be included. Moreover, the analyst has to make hypotheses on (i) which state variable are parameters dependent to and (ii) whether all parameters are state dependent or some of them are only time-dependent or simply constant.

In the present context, however, considerable simplifications can be achieved considering that model is deterministic and that, from (2), we *know* that, truncating all terms of order two and higher, the model can be written as:

$$Y_t - f_0 = f_1(X_{1,t}) + f_2(X_{2,t}) + \ldots + f_k(X_{k,t}) + e_t \quad e_t \sim N(0,\sigma^2) \tag{6}$$

where we assume that all terms of high order can be approximated by a Gaussian white noise with zero mean and variance σ^2, i.e. the truncated HDMR is seen as a stochastic non-linear system. This can be justified by a version of the central limit theorem [16], since the truncated terms can be seen as the sum of a large number of independent random variables with equal zero mean and arbitrary probability distribution.

Comparing the representation (6) to the SDP model definition (5) we can see that:

1. no autoregressive terms of the output variable are present in (6), i.e. *n*=0;
2. no lags or delays in the input variables are present in (6), i.e. *m*=0, δ=0;
3. items 1 and 2 imply that the state vector reduces to the vector of input variables, i.e. $\mathbf{Z}_t = \mathbf{X}_t$ and that the vector of time dependent parameters reduces to $\mathbf{p}_t = \mathbf{b}_0(\mathbf{X}_t)$;
4. each term of the sum (6) is a function of a single input variable, so each state dependent parameter b_{0i} depends only on the corresponding input variable X_i, i.e. $p_{i,t} = b_{0j}(\mathbf{X}_t) \equiv b_{0j}(X_{j,t})$. Without loss of generality, we can then re-write each term of (6) as $f_i(X_{j,t}) \equiv b_{0j}(X_{j,t})X_{j,t} = p_{j,t}X_{j,t}$.

So, the general, dynamic, time series specification (5), including lagged variables and delays, can be specialised to the HDMR of the computational model stopped to the first order (6) as follows:

$$Y_t - f_0 = p_{1,t}X_{1,t} + p_{2,t}X_{2,t} + \ldots + p_{k,t}X_{k,t} + e_t \qquad e_t = N(0,\sigma^2) \tag{7}$$

Estimating the state dependent parameters $p_{j,t}$ is hence equivalent to estimating the first order terms of the HDMR. The estimation of $p_{j,t}$ requires the following steps (see [12-13] for more details):

1. characterise the variability of $p_{j,t}$ in some stochastic manner: this is done using generalised random walk processes (GRW), specifically our choice is limited to the random walk (RW) and the integrated random walk (IRW) processes;

2. estimate via maximum likelihood (ML) the hyper-parameters (i.e. the variance of the innovations) of the RW/IRW processes of the state dependent parameters;

3. estimate the state dependent parameters and hence the first order terms of (2) applying recursive filtering and smoothing estimation, within an iterative backfitting procedure (The ML estimation in step 2. can also be iterated within the backfitting procedure).

With this procedure we estimate all the terms simultaneously, allowing us to use a single sample to estimate all indices. Moreover, the Monte Carlo sample is a standard one (pure random sample, Latin Hypercube, LP-tau, etc) and does not require a particular design, such as the classical variance based methods. This also allows it to be applied in the case of dependent inputs. We warn however that the convergence rate depends somehow on how the sample is generated. If quasi-random LP-tau random numbers are used, the convergence rate is very high, while using Latin Hypercube or pure random samples convergence is slower. This is clearly due to the more efficient exploration of the parameter space provided by the LP-tau quasi-random sequence.

2.2. The backfitting algorithm

The 'time scale' of the SDP model used for SA is just given by the sequence of the Monte Carlo evaluations of the computational model, so no 'logical' ordering can be expected in this sequence. In practice, the $\mathbf{p}_t$ values will continuously 'jump' in an extremely noisy way from one run to the subsequent in the Monte Carlo sample. So, it cannot be assumed that the simple GRW model is appropriate to describe such a variation over 'time'. However, it is possible to solve this problem if we sort the data in an ad-hoc manner. Specifically, if the ordering is chosen so that the SDP variations associated with the sorted series are smoother, it is more likely that a simple GRW process can be utilized to describe their evolution.

In our case, it is logical to assume that the most suitable ordering for each parameter $p_{j,t}$ should be done with respect to the corresponding input factor $X_{j,t}$. In this way, we can expect that the recursive estimation will be able to identify the pattern of Y_t vs. $X_{j,t}$. This also implies that each SDP needs a different sorting strategy, each with respect to its input factor. To solve this further problem, the backfitting procedure described in [13] can be exploited. Here, each parameter is estimated in turn, based on the modified dependent variable series obtained by subtracting all the other terms on the right hand side of (7) from Y_t. At each such backfitting iteration, the sorting can then be based on the single variable associated with the current SDP being estimated.

2.3. Specific issues for SA applications

2.3.1. Rescaling the inputs

Suppose that the input factor distribution of a factor X_j contains the zero. Suppose also that the corresponding f_j term assumes a non-zero value for $X_j = 0$. This poses a singularity problem, since, if $X_{j,t} = 0$ at a given sample point *t*, $p_{j,t}$ should assume an infinite value to yield a non-zero $f_{j,t}$. So, we propose that in the standard procedure for SA estimation, while leaving the output unmodified, all the input factors are transformed via a monotonic (invertible) operator into cumulative probabilities, obtaining a uniform distribution for all input factors in the [0, 1] interval. Then, to eliminate the zero, all values are then shifted to the interval [1, 2]. This allows the elimination of the singularity problems, while preserving the scale and shape of the model output.

Then, after estimation, the f_j^* terms estimated for the transformed variable X_j^*, can be plotted by putting in the axis of the abscissas the original values X_j, fully recovering the true f_j terms. In fact, if $P_j(X_j)$ is the cumulative distribution of factor X_j, at each sample point *t* we can write: $f_{j,t}^*(X_{j,t}^*) = f_{j,t}(P_j^{-1}(X_{j,t}^* - 1)) = f_{j,t}(X_{j,t})$.

2.3.2. The choice of the GRW model

A last methodological issue on the use of SDP models for sensitivity analysis of computational models concerns the choice of the GRW model, i.e. RW of IRW. Usually, sensitivity analysis tools, such as the software SIMLAB [17], aim to be 'black-box'. Whatever the model, inputs and outputs are fed to the SA tool to get the sensitivity indices. In principle, *a priori* one does not know whether RW or IRW model is more appropriate for each f_j term of the HDMR decomposition. In order to make this choice 'automatic', we propose the following preliminary step to the backfitting algorithm:

1. for each factor X_j perform the ML estimation of both RW and IRW models of the univariate model $Y_t = p_{j,t} X_{j,t} + e_t$;
2. for each estimated model, compute the R^2 measure or fit;
3. select the model with the highest R^2 in the subsequent SDP estimation.

3. APPLICATION

We have tested the method with several models, with up to 15 input factors. We first considered models with known analytical solutions for sensitivity indices and HDMR, such as the *g*-function of Sobol' [9]. Such tests allowed the convergence rate of the numeric to analytic solution to be measured. In general, a number of model runs of about 1,000 is sufficient for quite accurate estimates, with an absolute errors of about 0.01 on a scale [0, 1]. An extremely important improvement with respect to the classical estimation methods is that, in all the tests done, the computational cost was almost independent of the number of input factors. Clearly, it has to be expected that for a number of factors larger than 15-20, the convergence rate will start decreasing, but cases with such a large number of input factors are

usually preliminary passed through a screening method (e.g. the method of Morris or others) and only at a later stage the variance based/HDMR analysis is performed for fewer inputs. Convergence also becomes slower for non-monotonic models for which the parameters tend to be all rather equally important. Note that the latter condition is not usual, as most often factors follow a Pareto law, with few factors accounting for most of the importance. In all cases, the convergence is reached with a few thousand runs, with a considerable saving in computational effort with respect to the standard estimation methods (FAST, Sobol'). Even in the nastiest cases, convergence did not require more than 8000 runs.

Here, we show some significant results for the Level E model. Level E was used both as a benchmark of Monte Carlo computation [18] and as a benchmark for sensitivity analysis methods [19]. This test case has been extensively used by several authors; see [20] for a review. The model predicts the radiological dose to humans over geological time scales due to the underground migration of radionuclides from a nuclear waste disposal site. In a companion paper to this conference [21], readers can find another application of SDP modelling for the SA of a basin model to evaluate hydrocarbon exploration risk.

The Level E model has 12 input factors and is characterised by a strong non-linearity. Among the 12 parameters, X_4 (=$v^{(1)}$, water velocity in the first geosphere layer) and X_{12} (=W, stream flow rate) have the largest main effect over the simulated period. In Figures 1 and 2 we show the sensitivity indices versus time for these two parameters and compare the asymptotic values estimated with standard SA tools (Sobol' method), taking 1,000,000 runs, with the SDP estimation having total costs of 1024 and 8192. The samples for SDP model analysis were generated using LP-tau quasi-random sequences. We can see that already with only 1024 runs, which is a very small sample size for this kind of model, the absolute errors of the SDP estimates with respect to the asymptotic values is of the order of 0.01-0.02 in the sensitivity scale range of [0, 1]. With 1024 runs there is a critical point for W, where the drop of the sensitivity index at t=200,000 yr is shifted to the next time point t=300,000 yr. Increasing the total cost to 8192, results converge to the asymptotic values.

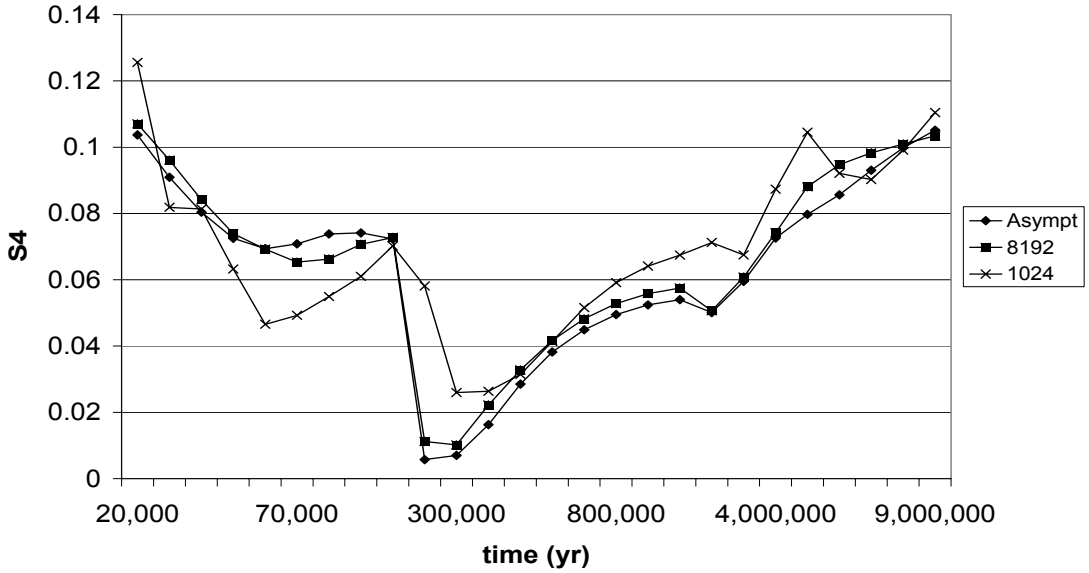


Figure 1. First order sensitivity index vs, time for parameter $v^{(1)}$ (X_4).

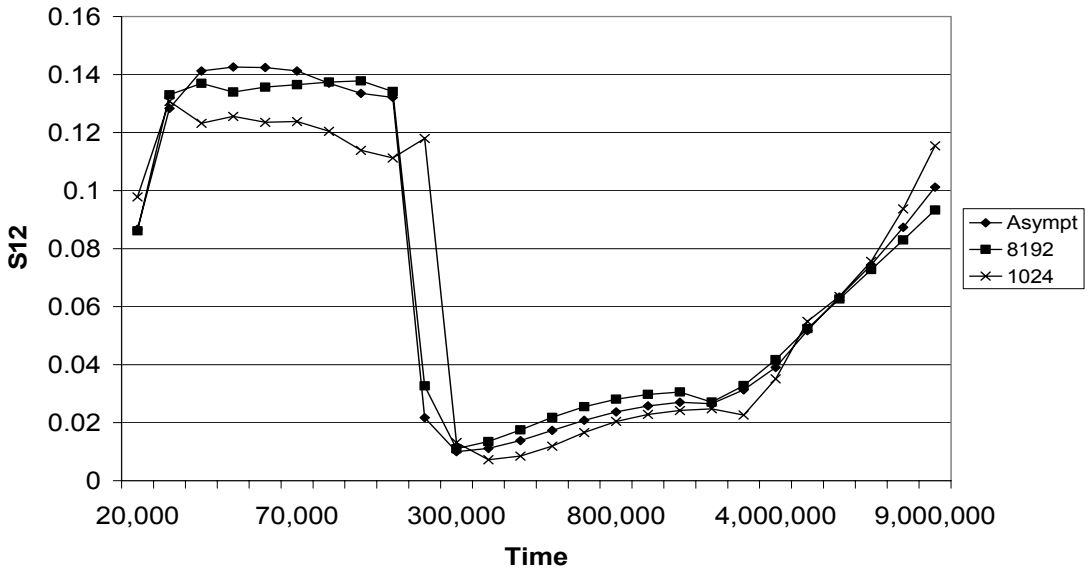


Figure 2. First order sensitivity index vs. time for parameter *W* (X12).

Comparing total costs, with the Sobol' technique we would need about 40,000 model runs to reach an accuracy comparable to the cheaper SDP model estimation of 1,024 model runs, i.e. the SDP modelling approach reduces the computational time by a factor 40 in this case. Conversely, 1024 runs for the Sobol' estimates are too few, with absolute errors that can reach 0.7-0.8, i.e. totally unreliable estimates.

In addition to sensitivity estimates, the SDP modelling approach also allows us to estimate the first terms in the high dimensional model representation decomposition. The plots of such functions for $v^{(1)}$ and *W* at the time *t*=100,000 are shown in Figure 3. The added value of the SDP modelling approach is evident by looking at the clear representation of the first order input-output mapping between $v^{(1)}$ and *W* and the output *Y* (the radiological dose). It is interesting to note in Figure 3 that the pattern estimated with 1024 runs for $f(X_4)$ slightly passes the zero axis for high values of X_4, while this is corrected increasing the number of runs to 8192. To better appreciate the 'pattern recognition' performed by the SDP estimation, in Figure 4 we also compare the scatter plots to the SDP estimates of $f(X_4)$ and $f(X_{12})$ (solid lines), for 8196 runs.

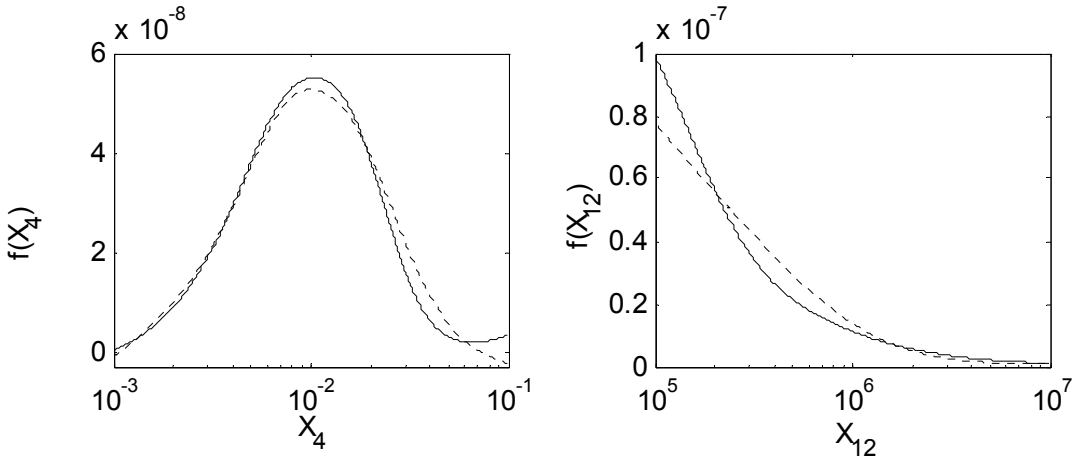


Figure 3. First order terms of the HDMR of the Level E model for $v(1)$ ($=X_4$) and W ($=X_{12}$) at t=100,000 yr. Solid lines are for the total cost of 8192 runs; dotted lines for the total cost of 1024 runs.

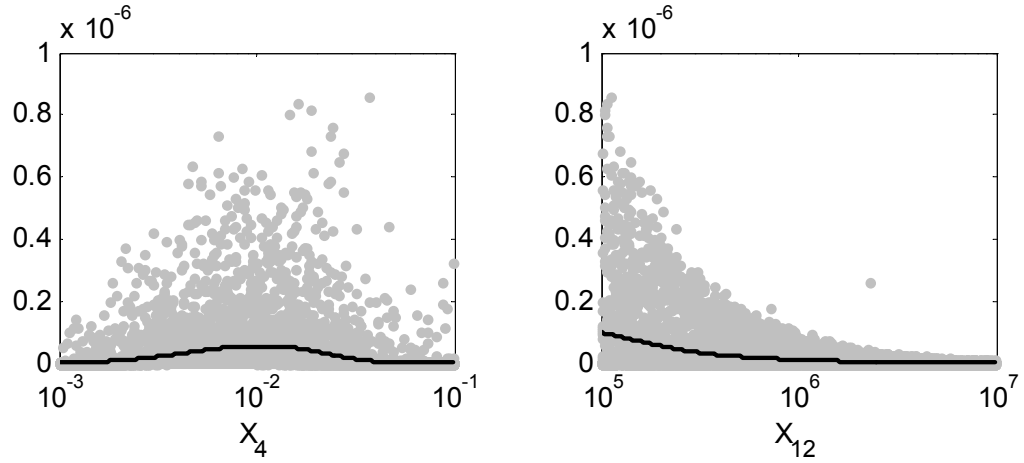


Figure 4. Scatter plots of Y and estimated first order terms of (2) for the Level E model for $v(1)$ ($=X_4$) and W ($=X_{12}$) at t=100,000 yr (8192 runs).

4. CONCLUSIONS

The use of SDP models is a powerful tool for a fast and accurate estimate of the first order terms of the HDMR and of the main effects sensitivity indices of computational models. All the estimates are performed with a unique sample, which can be any standard Monte Carlo sample. However, if efficient quasi-random number generators are used, such as the LP-tau sequence, the efficiency of the method is further enhanced, with a significantly faster convergence. We have tested the method with different models, always with extremely rapid convergence rates: 1,000 runs are in most cases sufficient for good estimates with models having up to 10-15 input factors. The dependence of the computational cost of the method to the number of input factors is very small: this is an extremely important improvement with respect to classical estimation methods. The convergence becomes slower in cases where the model is non-monotonic *and* the input factors share similar and relatively small levels of importance, i.e. they do not follow a Pareto law.

REFERENCES

1. I. M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Matematicheskoe Modelirovanie*, 2: 112-118, 1990 (translated as: I. M. Sobol'. Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling & Computational Experiment*, 1: 407-414, 1993).
2. H. Rabitz, Ö. F. Aliş, J. Shorter, and K. Shim. Efficient input-output model representations. *Computer Physics Communications*, 117: 11-20, 1999.
3. H. Rabitz, and Ö. F. Aliş. Managing the Tyranny of Parameters in Mathematical Modelling of Physical Systems. In A. Saltelli, K. Chan and M. Scott, Eds, *Sensitivity Analysis*, 199-223. Wiley & Sons, Ltd, New York, 2000.
4. A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*, John Wiley & Sons publishers, Probability and Statistics series, to appear in March 2004.
5. A. Saltelli. Global Sensitivity Analysis: An Introduction. SAMO2004, Keynote lecture.
6. A. Saltelli, K. Chan, and M. Scott, Eds. *Sensitivity Analysis*. John Wiley & Sons, Ltd, New York, 2000.
7. A. Saltelli. Making best use of model valuations to compute sensitivity indices. *Computer Physics Communications*, 145: 280-297, 2002.
8. J. Oakley, and A. O'Hagan. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Research Report No. 525/02* Department of Probability and Statistics, University of Sheffield, 2002. Submitted to *Journal of the Royal Statistical Society, Series B*.
9. Ratto M., Saltelli A., Young P., in preparation.
10. R. E. Kalman. A new approach to linear filtering and prediction problems. *ASME Trans., Journal Basic Eng.*, 82D: 35-45, 1960.
11. R. E. Kalman, and R. S. Bucy. New results in Linear filtering andpred iction theory. *ASME Trans., Journal Basic Eng.*, 83-D: 95-108, 1961.
12. P. C. Young. Nonstationary time series analysis and forecasting, *Progress in Environmental Science*, 1: 3-48, 1999.
13. P. C. Young. Stochastic, Dynamic Modelling and Signal Processing: Time Variable and State Dependent Parameter Estimation. Chapter from W. J. Fitzgerald et al. Eds, *Nonlinear and Nonstationary Signal Processing*, 74-114. Cambridge University Press, Cambridge, 2000.
14. P. C. Young. *Recursive Estimation and Time-Series Analysis*, Springer, 1984.
15. P.C. Young, S. D. Parkinson and M. Lees. Simplicity out of complexity: Occam's razor revisited. *Journal of Applied Statistics*, 23: 165-210, 1996.
16. W. Feller. The fundamental limit theorems in probability. *Bull. Amer. Math.Soc*., 51:800–832, 1945.
17. SIMLAB, Version 2.2, Simulation environment for Uncertainty and Sensitivity Analysis, http://www.jrc.cec.eu.int/uasa/prj-sa-soft.asp.
18. OECD/NEA PSAC User group. *PSACOIN Level E intercomparison.* B. W. Goodwin, J. M. Laurens, J. E. Sinclair, D.A. Galson, and E. Sartori., Eds. Nuclear Energy Agency, Organisation for Economic Cooperation and Development, Paris, 1989.
19. OECD/NEA PSAG User group. *PSACOIN Level S intercomparison*. A. Alonso, P. Robinson, E. J. Bonano and D. A. Galson, Eds. Nuclear Energy Agency, Organisation for Economic Cooperation and Development, Paris, 1993.
20. A. Saltelli, and S. Tarantola. On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. *Journal of American Statistical Association*, 97: 702-709, 2002.
21. L. P. Ruffo, L. Bazzana, A. Corradi, A. Saltelli. S. Tarantola, M. Ratto, Exploration Risk Evaluation Through UASA Techniques, SAMO2004, oral presentation SAMO-04-20.

Sensitivity Analysis of Monte Carlo Estimates from Computer Models in the Presence of Epistemic and Aleatory Uncertainties

Bernard Krzykacz-Hausmann

Gesellschaft für Anlagen- und Reaktorsicherheit (GRS)
85748 Garching
Germany
E-mail: krb@grs.de
Fax: ++49-89-32004 301

Abstract: Results from complex computer models are often subject to both *aleatory* and *epistemic* uncertainty. The natural straightforward procedure to analyze these uncertainties by Monte Carlo simulation is a double-loop nested sampling: the epistemic parameters are sampled in the outer loop and the aleatory variables are sampled in the nested inner loop. For time-demanding codes, however, the computational effort of this procedure may be prohibitive. Therefore a method of an approximate sensitivity analysis ("sensitivity" in the sense of "uncertainty importance") has been suggested which is based on a single-loop sampling procedure with epistemic parameters and aleatory variables being sampled "simultaneously" from their respective distributions. From the results of such sampling one can obtain approximate estimates of many of the commonly used sensitivity measures for the aleatory probability distributions of model outcomes of interest with respect to the underlying epistemic parameters. The reliability of these estimates depends on the relative contribution of epistemic uncertainties **U** to the overall joint epistemic & aleatory uncertainty in the outcome Y expressed by the quantity $c^2 = \mathrm{var}E[Y|\mathbf{U}]/\mathrm{var}Y$. This quantity can be estimated in several ways depending on the feasibility of additional sampling and model computations.

Keywords: sensitivity analysis, aleatory and epistemic uncertainty, uncertainty importance, conditional expectation.

1. INTRODUCTION

The effect of model input variables subject to aleatory uncertainty ("random behavior") on the results of a complex model can be analyzed by Monte Carlo simulation. To this end the aleatory variables are sampled according to their random laws and the results of the corresponding model runs are summarized in form of empirical distributions which represent the aleatory uncertainty of the model outcomes. From these empirical distributions statistical estimates of the probabilities of the process states of interest and other useful probabilistic quantities like expectations etc. may be obtained.

Often, however, the exact types of the random laws, their distributional parameters, the model formulations, the values of model parameters, the input data of the model application etc are not known precisely, i.e. they are subject to epistemic ("lack-of-knowledge") uncertainty. These uncertainties, denoted as epistemic input uncertainties, are quantified by probability distributions representing the respective subjective state of knowledge.

The aim of epistemic sensitivity analysis ("uncertainty importance analysis") in this case is to quantify the effect of the epistemic input uncertainties on the epistemic uncertainty of the

probabilistic quantities representing aleatory output uncertainty, e.g. probabilities, expectations etc.

It is widely recognized and accepted that these two types of uncertainty must very carefully be distinguished and therefore it wouldn't make sense to perform a "simultaneous" Monte Carlo simulation of both types of variables and a sensitivity analysis of a direct model outcome with respect to the variables of both types.

It is intuitively clear and has often been pointed out by many authors, e.g. [1], that the natural method to appropriately account for both types of uncertainty by Monte Carlo simulation is a "double-loop" nested sampling procedure (also called "two-stage" or "two-dimensional" sampling,). It consists of (1) an "outer loop" where the values of the epistemic parameters are sampled according to their epistemic marginal probability distributions and (2) a nested "inner loop" where the values of the aleatory variables are sampled according to their aleatory conditional probability distributions given the values of the epistemic variables chosen in the outer loop. Each "inner loop" provides an empirical conditional aleatory distribution of the process outcome of interest such that finally a sample of empirical distributions is obtained. This sample could be used for a standard epistemic sensitivity analysis for various (aleatory) probabilistic quantities.

However, for complex and computationally expensive models, as used e.g. in probabilistic safety analysis of nuclear power plants, the computational effort for the double-loop procedure will be prohibitive. In such cases the consequence would be to do without an uncertainty and sensitivity analysis.

Therefore, an approach of an approximate epistemic sensitivity analysis is suggested in the following sections. Instead of the nested double-loop sampling procedure the above-mentioned simple single-loop sampling procedure is employed with both types of variables being sampled "simultaneously" according to their joint probability distribution. From the results of this sampling appropriate sensitivity measures can be computed.

2. FUNDAMENTALS

Being subject to both epistemic and aleatory uncertainties, any scalar process variable or model outcome Y may be represented as

$$Y = h(\mathbf{U},\mathbf{V})$$

with

- **U** = set of all epistemic uncertainties (**u**ncertain parameters),
- **V** = set of all aleatory uncertainties (random **v**ariables),
- h = the computational model considered as a deterministic function of both aleatory and epistemic uncertainties **U** and **V**.

When holding the epistemic variables **U** fixed at a value **u** , i.e. **U**=**u**, the resulting outcome Y is a function of the aleatory uncertainties **V**, solely. Its probability distribution, i.e. the conditional distribution $F(y|\mathbf{U}=\mathbf{u})$ of Y given **U**=**u**, quantifies the corresponding (conditional) aleatory uncertainty in Y. Its expectation

$$E[Y|\mathbf{U}=\mathbf{u}]$$

taken over all aleatory variables **V** conditionally on **U**=**u** may be considered as a scalar quantity representing this conditional aleatory uncertainty of the outcome Y.

Using expectation to represent conditional aleatory uncertainty must not be considered very restrictive since many of the standard distributional parameters characterizing aleatory uncertainty can be viewed as expectations of appropriately chosen outcome functions Y'. E.g. the value $F_Y(y)$ of a distribution function of a random variable Y at any given point y may be represented as expectation of the indicator variable $Y' = I_{\{Y \le y\}}$, i.e. Y'=1 if Y≤y and Y'=0 otherwise, from which follows that $EY' = F_Y(y)$.

In the following the standard concise notation

$$E[Y|\mathbf{U}]$$

will be used to denote the above conditional expectation $E[Y|\mathbf{U}=\mathbf{u}]$ considered as function of the epistemic uncertainties **U**, i.e. as a quantity subject to epistemic uncertainty from **U** alone.

The principal aim of an approximate epistemic sensitivity analysis of results from models subject to both epistemic and aleatory uncertainties will therefore be to determine appropriate sensitivity indices of the conditional expectation $E[Y|\mathbf{U}]$ with respect to the components $U_1,\ldots,U_n$ of **U** avoiding the time-consuming double-loop Monte Carlo sampling.

The following fact is the basis of the proposed method:

Many of the standard sensitivity measures of $E[Y|\mathbf{U}]$ *with respect to* $U_1,\ldots U_n$ *are uniformly proportional to the corresponding sensitivity measures of* $Y=h(\mathbf{U},\mathbf{V})$ *with respect to* $U_1,\ldots U_n$. *The proportionality constant* c *is, in most cases, given by*

$$c = \sqrt{\frac{\operatorname{var} EY|\mathbf{U}}{\operatorname{var} Y}}.$$

I.e. if SM_i denotes the (population) sensitivity measure of $E[Y|\mathbf{U}]$ with respect to epistemic parameter U_i, and SM'_i denotes the corresponding sensitivity measure of $Y= h(\mathbf{U},\mathbf{V})$ with respect to the same parameter, then

$$SM'_i = c \cdot SM_i$$

for all i=1,…,n. This holds for many types of sensitivity measures with the same constant c.

Consequently, this property implies that the sensitivity indices for
(a) the conditional expectation $E[Y|\mathbf{U}]$ and for
(b) the direct outcome $Y=h(\mathbf{U},\mathbf{V})$
provide *the same uncertainty importance ranking* with respect to parameters $U_1,\ldots U_n$.

This result holds for the sensitivity measures
- Correlation Coefficient (CC)
- Standardized Regression Coefficient (SRC)
- Correlation Ratio CR (="main effect" sensitivity index)

and with slight modifications also for
- Partial Correlation Coefficient (PCC)
- "total effect" sensitivity index ST
- "linearized"(or R^2-) Version of the "total effect" sensitivity index STL.

The proof of this fact becomes very simple if the concept of conditional expectation $E[Y|\mathbf{U}]$ is employed. It is worthwhile mentioning that the notion of conditional expectation is very useful also in the context of sensitivity analysis. Many results from the standard sensitivity analysis which look rather complex and difficult can very effectively be

represented, very clearly interpreted and very easily proved with the aid of the concept of conditional expectation.

The following basic properties of conditional expectation are useful in this context. They can be found in many textbooks and can also very easily be proved:

(1) $E(E[Y|\mathbf{U}]) = EY$
(2) $var(E[Y|\mathbf{U}]) = varY - E(var[Y|\mathbf{U}])$
(3) $E(E^2[Y|\mathbf{U}]) = E(Y \cdot E[Y|\mathbf{U}])$
(4) $E(E[Y|\mathbf{U}]|U_i) = E[Y|U_i]$
(5) $E[E[Y|\mathbf{U}] \cdot U_i] = E[E[Y|U_i] \cdot U_i] = E[Y \cdot U_i]$
(6) $cov(E[Y|\mathbf{U}],U_i)=cov(Y,U_i)$
(7) the linear regression of $E[Y|\mathbf{U}]$ with respect to $\mathbf{U}$ and the linear regression of Y with respect to $\mathbf{U}$ are identical, i.e. $RC(E[Y|\mathbf{U}],U_i) = RC(Y,U_i)$ with RC(…) being the corresponding regression coefficients.

Using these properties the above result can easily be proved. Here, e.g., the proofs for the correlation coefficient CC and the correlation ratio CR ("main effect" sensitivity index):

$$CC(E[Y|\mathbf{U}],U_i) = \frac{cov(E[Y|\mathbf{U}], U_i)}{\sqrt{var\, E[Y|\mathbf{U}] \cdot varU_i}} = \frac{cov(Y, U_i)}{\sqrt{var\, E[Y|\mathbf{U}] \cdot varU_i}} =$$

$$= \frac{cov(Y, U_i)}{\sqrt{varY \cdot varU_i}} \cdot \sqrt{\frac{var\, Y}{var\, E[Y|\mathbf{U}]}} = CC(Y, U_i) \cdot \sqrt{\frac{var\, Y}{var\, E[Y|\mathbf{U}]}} =$$

$$= CC(Y, U_i) \cdot 1/c \, .$$

$$CR^2(E[Y|\mathbf{U}],U_i) = \frac{varE[E[Y|\mathbf{U}]|U_i]}{varE[Y|\mathbf{U}]} = \frac{varE[Y|U_i]}{varE[Y|\mathbf{U}]} = \frac{varE[Y|U_i]}{varY} \cdot \frac{varY}{varE[Y|\mathbf{U}]} =$$

$$= CR^2(Y,U_i) \cdot 1/c^2 \, .$$

The proofs for the other sensitivity measures are similar.

3. SAMPLING METHOD FOR AN APPROXIMATE SENSITIVITY ANALYSIS

Owing to the preceding result it seems natural and reasonable to replace the above-mentioned but often impracticable double-loop sample-based sensitivity analysis for the conditional expectation $E[Y|\mathbf{U}]$ by the corresponding sensitivity analysis for the direct outcome $Y=h(\mathbf{U},\mathbf{V})$ with respect to the components $U_1,\ldots U_n$ of $\mathbf{U}$, alone. The Monte Carlo sampling procedure appropriate for such sensitivity analysis for the direct outcome Y, however, is a simple single-loop sampling with the epistemic parameters $\mathbf{U}$ and the aleatory variables $\mathbf{V}$ being sampled "simultaneously" according to their joint probability distribution $f(\mathbf{u},\mathbf{v})$. This joint probability distribution is given by the product of the marginal distribution $f(\mathbf{u})$ of $\mathbf{U}$ and the conditional distribution $f(\mathbf{v}|\mathbf{U}=\mathbf{u})$ of $\mathbf{V}$ given $\mathbf{U}=\mathbf{u}$, i.e. by the expression

$$f(\mathbf{u},\mathbf{v}) = f(\mathbf{v}|\mathbf{U}=\mathbf{u}) \cdot f(\mathbf{u}) \, .$$

In most applications the marginal distribution $f(\mathbf{u})$ of the epistemic parameters $\mathbf{U}$ will be given directly, while the conditional distribution $f(\mathbf{v}|\mathbf{U}=\mathbf{u})$ of the aleatory variables $\mathbf{V}$ may also be given in terms of intermediate results from the computational model.

Thus, the "simultaneous" sampling procedure with sample size N generates N joint epistemic & aleatory sample values

$$(\mathbf{u}_1,\mathbf{v}_1),\ldots\ldots,(\mathbf{u}_N,\mathbf{v}_N)$$

from which, eventually, the corresponding sample values

$$y_1,\ldots\ldots,y_N$$

of the direct outcome Y=h(**U**,**V**) are calculated via the computer code.

From all these sample values the above mentioned standard sensitivity measures with respect to the parameters $U_1,\ldots U_n$ for the outcome Y=h(**U,V**) can be computed. Since the proportionality constant $c=\sqrt{(varE[Y|\mathbf{U}]/varY)}$ is usually not known one cannot directly derive the sensitivity indices for E[Y|**U**] from the sensitivity indices for Y. However, according to the preceding section, the sample based parameter importance ranking obtained for Y may approximately be used as the importance ranking for the conditional expectation E[Y|**U**] asked for. Methods for approximating/estimating the proportionality constant c will be presented in section 5.

It is also clear that Simple Random Sampling (SRS) as well Latin Hypercube Sampling (LHS) or any other sampling method appropriate for the selected type of sensitivity measure may be used for such sample-based approximate sensitivity analysis.

4. ACCURACY CONSIDERATIONS

The accuracy of the approximate sensitivity analysis for the outcome Y depends on the (usually) unknown value of the (squared) proportionality constant

$$c^2 = \frac{\text{var } E[Y \mid \mathbf{U}]}{\text{var } Y}$$

which relates the sensitivity measures for E[Y|U] to the sensitivity measures for Y.

Clearly, $0 \leq c^2 \leq 1$ since varE[Y|**U**] ≤ varY due to the above property (2) of the conditional expectation. From the proportionality $SM_i=1/c\cdot SM_i'$ (i=1,…,n) it follows that the values of the sensitivity measures SM_i' for Y are uniformly lower than the corresponding sensitivity measures SM_i for E[Y|**U**]. If this constant c^2 is small, the sample-based approximate sensitivity analysis for Y may produce small or even statistically not significant values of the sensitivity measure for a parameter although the sensitivity of E[Y|**U**] with respect to this parameter one is actually interested in may be high. Nevertheless, c^2 is unknown and therefore it is important to analyze it more closely.

By definition, c^2 is easily identified as squared multiple correlation ratio (or "main effect" sensitivity index) [2],[3] of Y with respect to the whole parameter vector **U**. It can therefore be interpreted in several ways, e.g.

- as an indicator of the accuracy of the approximation of Y=h(**U**,**V**) by E[Y|**U**] as a function of **U** alone,
- as an indicator of the relative contribution of the epistemic uncertainties from **U** to the overall "joint" uncertainty in Y=h(**U**,**V**) from **U** and **V**,
- as the extent to which the overall "joint" uncertainty in Y coming from **U** and **V** is dominated by the epistemic uncertainty coming from **U** alone.

Consequently, the more "dominant" the epistemic uncertainties the higher the c^2 value, and, consequently, the higher the dependability of the proposed approximate sensitivity analysis.

In practical applications it may sometimes be immediately clear which type of uncertainty is dominant such that the reliability of the approximate sensitivity results may also be judged immediately. Nevertheless, an approximation of c^2 is needed on the basis of the reduced sampling effort without employing the impracticable double-loop approach.

5. APPROXIMATING THE PROPORTIONALITY CONSTANT c^2

Three alternative procedures are proposed to approximate res. to estimate the (squared) proportionality constant $c^2 = \frac{\text{var } E[Y \mid \mathbf{U}]}{\text{var } Y}$. Below the three procedures are ordered according to the amount of the additional computational effort necessary to determine the corresponding approximated res. estimated value of c^2.

(1) Procedure No.1 to approximate c^2 is based solely on the underlying sample values from the "joint" sampling of **U** and **V**, i.e. without additional model computations. It is given by

$$\hat{c}^2 = \frac{R^2(Y,\mathbf{U})}{R^2(Y,(\mathbf{U},\mathbf{V}))} ,$$

with

$R^2(Y,(\mathbf{U},\mathbf{V}))$ = multiple sample correlation coefficient of outcome Y with respect to the joint sample of (**U**,**V**).

$R^2(Y,\mathbf{U}))$ = multiple sample correlation coefficient of outcome Y with respect to the sample of **U** alone.

Both multiple correlation coefficients can easily be computed from the available sample values $(\mathbf{u}_1,\mathbf{v}_1),\ldots..,(\mathbf{u}_N,\mathbf{v}_N)$ and $y_1,\ldots,y_N$ according to the well-known formulae:

$$R^2(Y,\mathbf{U}) = \rho^t_{Y,\mathbf{U}}\, \mathbf{R}_\mathbf{U}^{-1}\, \rho_{Y,\mathbf{U}}$$

$$R^2(Y,(\mathbf{U},\mathbf{V})) = \rho^t_{Y,(\mathbf{U},\mathbf{V})}\, \mathbf{R}_{\mathbf{U},\mathbf{V}}^{-1}\, \rho_{Y,(\mathbf{U},\mathbf{V})}$$

with

$\rho_{Y,\mathbf{U}}$ = vector of empirical correlation coefficients between Y and the components of **U**

$\rho_{Y,(\mathbf{U},\mathbf{V})}$ = vector of empirical correlation coefficients between Y and the components of **U**,**V**

$\mathbf{R}_\mathbf{U}^{-1}$ = inverse of the empirical correlation matrix $\mathbf{R}_\mathbf{U}$ between the components of **U**

$\mathbf{R}_{\mathbf{U},\mathbf{V}}^{-1}$ = inverse of the empirical correlation matrix $\mathbf{R}_{\mathbf{U},\mathbf{V}}$ between the components of **U**,**V**.

All these quantities are computed from the underlying sample values $(\mathbf{u}_1,\mathbf{v}_1),\ldots..,(\mathbf{u}_N,\mathbf{v}_N)$ and $y_1,\ldots.,y_N$ generated by the single-loop joint sampling of **U**,**V** and the corresponding model computations of Y. The sample size N must exceed the joint number of variables in **U**,**V**.

The motivation behind this method is simply to approximate the conditional expectation (= regression of the 1st kind) by the linear regression (= regression of the 2nd kind).

(2) Procedure No.2 of approximating c^2 is based on two samples: (a) the underlying sample values $y_1,\ldots,y_N$ from the same "joint" sampling of **U** and **V** and (b) sample values from an additional (single-loop) sampling of aleatory variables **V** alone with the values of epistemic parameters **U** held fixed at their nominal values $\mathbf{u}_0$. It is defined by

$$\hat{c}^2 = \frac{s^2(Y) - s^2(Y \mid \mathbf{U} = \mathbf{u}_0)}{s^2(Y)} \ .$$

where

$$s^2(Y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

is the variance from the underlying sample values $y_1,\ldots,y_N$, and

$$s^2(Y|\mathbf{U}=\mathbf{u}_0) = \frac{1}{N'}\sum_{i=1}^{N'}(y'_i - \bar{y}')^2 .$$

is the variance from the other sample values $y'_1,\ldots,y'_{N'}$ generated by sampling the aleatory variables **V** alone while the epistemic parameters **U** are held fixed at their nominal values $\mathbf{u}_0$. In many applications this additional sample may already be available as the "nominal result" computed before starting uncertainty and sensitivity analysis.

The motivation behind this method is to approximate the term Evar[Y|**U**] appearing in the expression varE[Y|**U**] = varY - Evar[Y|**U**] for the numerator of c^2 by the term var[Y|**U**=$\mathbf{u}_0$].

(3) Procedure No. 3: While the first two methods should rather be considered as numerical approximations to the constant c^2=varE[Y|**U**]/varY, the third method may be viewed as an estimate of c^2 in the full statistical sense. It is based on the following basic and easy to prove property of conditional expectation:

If **V** *and* **V'** *are identically distributed and conditionally independent given* **U**, *i.e. the joint conditional distribution of* **V** *and* **V'** *given* **U** *is the product if the two marginal conditional distributions, formally:* f(**v**,**v**'|**U**=**u**) = f(**v**|**U**=**u**)·f(**v**'|**U**=**u**) *and if* Y = h(**U**,**V**) *and* Y' = h(**U**,**V**'*), then the (squared) proportionality constant* c^2 *can be expressed by:*

$$c^2 = \frac{\text{var E}[Y \mid \mathbf{U}]}{\text{var Y}} = \frac{\text{cov}(Y, Y')}{\sqrt{\text{varY varY'}}} = \rho(Y,Y') \ ,$$

i.e. c^2 *is the correlation coefficient between the variables* Y *und* Y'.

Consequently, one can estimate the proportionality constant c^2 by the sample correlation coefficient r(**y**,**y**') from the two-dimensional sample $(y_1,y_1'),\ldots..,(y_N,y_N')$ from the bivariate distribution of (Y,Y'). The corresponding well-known formula is

$$\hat{c}^2 = r(\mathbf{y},\mathbf{y}') = \frac{\sum_{i=1}^{N}(y_i - \bar{y})\cdot(y'_i - \bar{y}')}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{N}(y'_i - \bar{y}')^2}} ,$$

where

$y_1,\ldots,y_N$ are the sample values of Y from the underlying "joint" sample of **U** and **V**, i.e. $y_i=h(\mathbf{u}_i,\mathbf{v}_i)$,), i=1,…,N and

$y_1',\ldots,y_N'$ are the sample values of Y' from the "joint" sample of **U** and **V'** generated by independently sampling the aleatory variables **V'** alone, according to the conditional distribution with the epistemic parameters **U** held fixed at the same values as in the 1st sample, i.e. $y_i'=h(\mathbf{u}_i,\mathbf{v}_i')$, i=1,…,N.

The additional computational effort for this statistical estimate of c^2 is therefore N additional model computations (= 2nd single-loop sample of size N).

Obviously, these two single-loop samples may also be viewed as a realization of the above-mentioned nested double-loop sampling with the "inner" loop sample size being 2.

Remark 1:

The above statistical estimate $\hat{c}^2$ may also be considered as an extension of the familiar procedure [4] to estimate the so-called "main effect" and "total effect" sensitivity indices SM and ST in the case of *not independent* variables. Changing the notation and replacing **U** by $\mathbf{X}_2$ and **V** by $\mathbf{X}_1$ the "total effect" sensitivity index ST_1 for Y with respect to $\mathbf{X}_1$ may be defined as

$$ST_1 = \frac{E\,var[Y \mid \mathbf{X}_2]}{varY} = \frac{varY - var\,E[Y \mid \mathbf{X}_2]}{varY}.$$

It can be interpreted as "the relative amount of variance of Y that is expected to remain if the values of all variables except variables $\mathbf{X}_1$ will become known". Analogously, the "main effect" sensitivity index SM_1 for Y with respect to $\mathbf{X}_2$ may be defined as

$$SM_2 = \frac{var\,E[Y \mid \mathbf{X}_2]}{varY}$$

and interpreted as "the relative amount of variance of Y that is expected to be removed if the values of all variables $\mathbf{X}_2$ will become known". This representation holds for independent as well as for dependent variables $\mathbf{X}_1$ and $\mathbf{X}_2$ and is equivalent to the representation given in [4] in the case of *independent* variables (e.g. ST_1 := sum of all terms containing $\mathbf{X}_1$ of the "Sobol decomposition" of $Y=h(\mathbf{X}_1,\mathbf{X}_2)$ into a sum of uncorrelated terms of increasing dimensionality [2], [4]). It is immediately seen that

$$ST_1 = 1 - c^2$$
$$SM_2 = c^2$$

with $\mathbf{X}_1,\mathbf{X}_2$ playing the role of **V**,**U** in the above representation of c^2. It can also be easily seen that for independent variables the estimate presented in this paper and the estimate presented in [4] are nearly equivalent. Consequently, in the procedure [4] to compute the "main effect" and the "total effect" sensitivity indices it is not necessary to assume the input variables be independent. This procedure can be used for dependent variables, as well, provided the two samples of $\mathbf{X}_1$ are generated conditionally independently given $\mathbf{X}_2$.

Remark 2:

According to the above procedure a 2nd sample is generated to estimate (together with the 1st sample) the constant c^2 while to estimate the sensitivity indices only the 1st sample is needed. It appears, and is intuitively clear, too, that using the mean sample values $y_i^* = (y_i+y_i')/2$, i=1,…,N from both samples an improvement of the accuracy of the sensitivity results can be achieved compared to the results obtained with the values y_i , i=1,…,N, from a single sample. As before, since $E[Y^*|\mathbf{U}]=E[Y|\mathbf{U}]$ and $varY^*=(varY+varE[Y|\mathbf{U}])/2$, it can easily be shown that a similar proportional relationship holds between the sensitivity measures SM_i of $E[Y|\mathbf{U}]$ and the corresponding sensitivity measures SM^*_i of $Y^*=(Y+Y')/2$ with respect to parameter U_i , i.e.

$$SM^*_i = c^* \cdot SM_i$$

with the new proportionality constant c^{*2} given by

$$c^{*2} = \frac{\text{var } E[Y^* | \mathbf{U}]}{\text{var } Y^*} = \frac{2 \text{ var } E[Y | \mathbf{U}]}{\text{var } Y + \text{var } E[Y | \mathbf{U}]} = \frac{2\,c^2}{1+c^2} > c^2$$

As stated above, since $c^{*2} > c^2$, the uncertainty importance ranking (sensitivity results) based on the y_i^* values will provide a more reliable approximation to the importance ranking for $E[Y|\mathbf{U}]$ than the importance ranking based on the y_i values from a single sample.

A straightforward generalization to K conditionally independent samples, i.e.

$$Y^* = 1/K \sum Y^{(k)}$$

provides an improvement with the proportionality constant

$$c^* = \frac{K \text{ var } E[Y | \mathbf{U}]}{\text{var } Y + (K-1) \text{ var } E[Y | \mathbf{U}]} = \frac{K\,c^2}{1+(K-1)c^2}.$$

This, obviously, is equivalent to the above-mentioned nested double-loop sampling with the "inner" loop sample size being K.

6. SIMPLE ANALYTICAL EXAMPLE (LINEAR NORMAL CASE)

To illustrate some of the preceding results a simple (artificial) numerical example is presented where all quantities of interest can be determined analytically and compared with the results from the sampling procedures presented above. In this example a simple linear independent normal case is considered, i.e.

$$Y = h(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{n} a_i U_i + \sum_{j=1}^{m} b_j V_j$$

where all epistemic parameters $\mathbf{U}=(U_1,\ldots,U_n)$ and all aleatory variables $\mathbf{V}=(V_1,\ldots,V_m)$ are independent and have the standard Normal distribution N(0,1). The coefficients a_i , b_j are assumed to be known. Then it can easily be shown that

- $\text{var}Y = \sum a_i^2 + \sum b_j^2$,
- $E[Y|\mathbf{U}] = \sum a_i U_i$,
- $\text{var}E[Y|\mathbf{U}] = \sum a_i^2$,
- $c^2 = \frac{\text{var } E[Y | \mathbf{U}]}{\text{var } Y} = \frac{\sum a_i^2}{\sum a_i^2 + \sum b_j^2}$,
- $SM_i' = SM(Y,U_i) = \frac{a_i}{\sqrt{\sum a_i^2 + \sum b_j^2}}$, (i=1,…,n) ,
- $SM_i = SM(E[Y|\mathbf{U}], U_i) = \frac{a_i}{\sqrt{\sum a_i^2}}$, (i=1,…,n) ,

where SM denotes any type of sensitivity measure, since, due to linearity and independence all standard sensitivity measures of Y or of $E[Y|\mathbf{U}]$ with respect to U_i are equal.

Here it can directly be seen: the higher the contribution of the epistemic uncertainties from **U** to the overall joint uncertainty in Y, expressed by the constant c^2, the more precise the proposed approximation of the sensitivity measures for $E[Y|\mathbf{U}]$ by the sensitivity measures for Y.

For numerical calculations it was assumed that n=m=5, $\mathbf{a}=\mathbf{b}=(1,2,3,4,5)$. Consequently varY=110, $\text{var}E[Y|\mathbf{U}]=55$, $c^2 = 1/2$.

The following table summarizes the results obtained analytically and with the sampling methods described above. It shows the values of the sensitivity measures (Standardized Regression Coefficient, SRC) for E[Y|**U**] with respect to all five parameters $U_1,\ldots,U_5$ obtained in four different ways:

(1) analytically,
(2) from double-loop simple random sampling with sample size 100x100
(3) from single-loop simple random sampling with sample size 500
(4) from single-loop simple random sampling with sample size 200.

Standardized Regression Coefficients (SRC) for E[Y|**U**]

Index of Parameter	(1) analytic	(2) two-loop ss=100x100	(3) one-loop ss=500	(4)one-loop ss=200
1	0.1348	0.1369	0.140	0.071
2	0.2696	0.2946	0.259	0.265
3	0.4044	0.4262	0.387	0.397
4	0.5392	0.5340	0.584	0.629
5	0.6740	0.7225	0.703	0.658

The three alternative methods for approximating/estimating the proportionality constant c provide the results:

The proportionality constant c

sample size	method 1	method 2	method 3	exact value
500	0.7027	0.7032	0.7003	$0.7071 = \sqrt{0.5}$
200	0.7437	0.6895	0.6920	$0.7071 = \sqrt{0.5}$

Conclusion: The results of this simple example look promising and suggest that in real situations with complex and computationally expensive models where the double-loop sampling is prohibitive, the approximate sensitivity analysis presented in this paper may provide reasonable results. It may therefore be preferred to the alternative of not performing any sensitivity analysis.

REFERENCES

[1] Helton J.C. et al., 1998. Uncertainty and sensitivity results obtained in the 1996 performance assessment for the waste isolation pilot plant. SAND98-0365, Albuquerque: Sandia National Laboratories; 1998

[2] Saltelli A., Chan K., Scott M. (eds.). Sensitivity Analysis. John Wiley & Sons, 2000.

[3] McKay M.D., 1995. Evaluating prediction uncertainty. Tech. Rep. NUREG/CR-6311. U.S. Nuclear Regulatory Commission and Los Alamos National Laboratory.

[4] Homma, T., Saltelli A., 1996. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety* Vol. 52, 1-17

Sensitivity Analysis When Model Outputs Are Functions

Katherine Campbell, Michael D. McKay, and Brian J. Williams

Statistical Sciences Group
Los Alamos National Laboratory, Los Alamos, NM 87545 USA
Corresponding author: Brian Williams: Fax: +1 505 667 4470 E-mail: brianw@lanl.gov

Abstract: When outputs of computational models are time series or functions of other continuous variables like distance, angle, etc., it can be that primary interest is in the general pattern or structure of the curve. In these cases, model sensitivity and uncertainty analysis focuses on the effect of model input choices and uncertainties on the overall shapes of such curves. We explore methods for characterizing a set of functions generated by a series of model runs for the purpose of exploring relationships between these functions and the model inputs.

Keywords: functional sensitivity analysis**,** functional data analysis, basis functions

1. INTRODUCTION

The outputs of computational models are often time series or functions of other continuous variables like distance, angle, etc. Following Campbell [1], we propose that sensitivity analysis of such outputs be carried out by means of an expansion of the functional output in an appropriate functional coordinate system, i.e., in terms of an appropriate set of basis functions, followed by sensitivity analysis of the coefficients of the expansion using any standard method. The principal new problem, therefore, is choosing an appropriate coordinate system in which to apply the selected sensitivity analysis methods. We consider both pre-defined basis sets and data-adaptive basis sets, with their associated advantages and disadvantages. We devote only passing mention to some related, but important problems, such as increasing the interpretability of the results by appropriate preprocessing of the functional outputs (in particular, alignment or registration of curves), and by enforcing some degree of smoothness when data-adaptive bases are used.

We will use a simple made-up example for explaining ideas. Fig. 1 shows a sample of curves generated by varying the four parameters, *a*, *b*, *c* and *d* in the "model"

$$f(\theta) = 10 + a \exp\left(-\frac{(\theta - b)^2}{K_1 a^2 + c^2} \right) + (b + d) \exp\left(K_2 \, a \, \theta \right). \qquad (1)$$

We interpret these functions as model output from a problem where the independent variable θ is a polar angle ranging from -90º to 90º. The model was run 81 times, using a complete 3^4 factorial design for the four input parameters.

In analyzing this "model output" we are typically less interested in what affects the values at, say, 45º, than in questions such as: What shifts the curves up and down or moves them left or right? What makes the central peak wider or narrower? What makes the right-hand tail

higher or lower? We could, of course, pick some appropriate functionals for answering these questions. The last, for example, we might address by examining the sensitivity of the values at 90º to the four input parameters. In order to address questions such as peak width we could devise some surrogate measurement that could be computed on each curve and then study its sensitivity to the input parameters. However, such choices are highly problem specific.

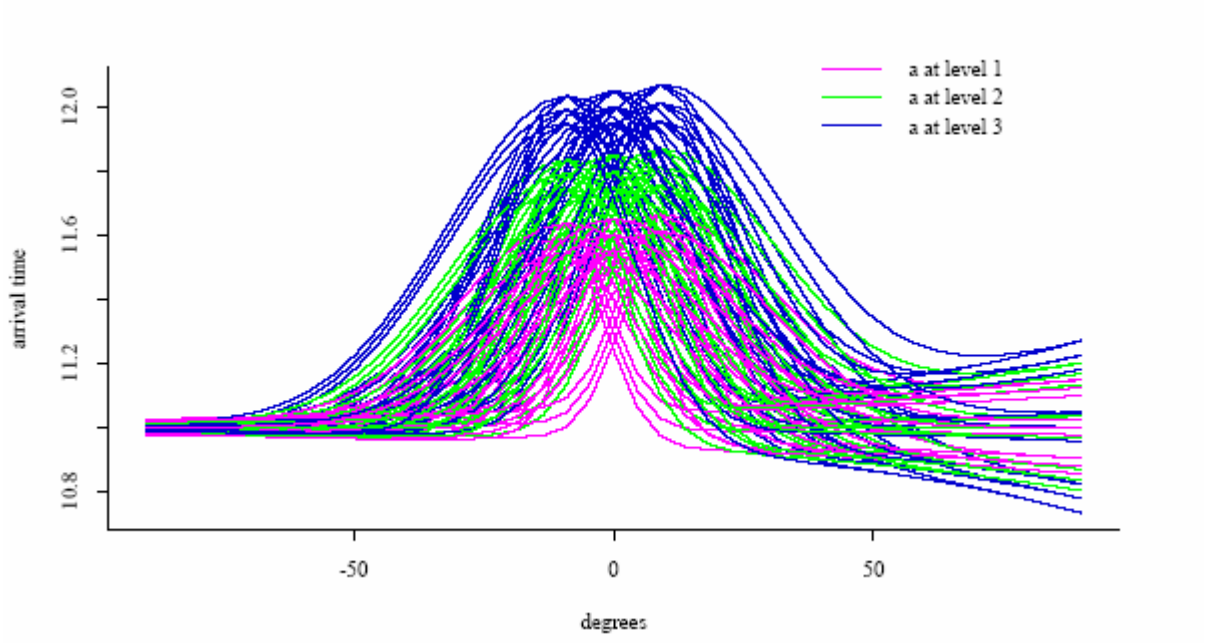


Figure 1. Functional output from 81 runs of the example model

2. TRANSFORMING FUNCTIONAL DATA

It might seem natural to regard functions provided on a grid of T points as T dependent variables for the purposes of sensitivity analysis. However, this approach can be unsatisfactory for many reasons:

- The T variables are highly correlated with one another, so this natural coordinate system is inefficient for statistical methods like discriminant analysis, sensitivity analysis, or almost anything other than multivariate statistical methods. Results are redundant from one value of θ to another.
- The pointwise results can be difficult to interpret for the underlying physical or modeling problem. In particular, information about the global functioning of the model or physical system contained in such curve features as location, scale and phase shifts, as well as in localized fluctuations including tail behavior, cannot generally be extracted from individual univariate analyses.

- Even though the data are the output of a computer model, the different runs may not have generated outputs at the same times or points θ. Alternatively, identical model output times may not be physically comparable because, as a function of the input parameters, the modeled process may be evolving faster in one run than another. So we may need to register the output curves (rescale time) in some physically more interpretable manner before proceeding with analysis.

All of these problems can be addressed by transforming the functional output in one way or another. For sensitivity analysis, the most useful approach is expanding the output functions in terms of some basis functions (after rescaling time, if necessary) and then applying the statistical method of interest—in our case, a sensitivity analysis method—to the coefficients of that expansion. Different types of bases can be considered. There are familiar, predefined bases such as Legendre polynomials or other orthogonal polynomials, trigonometric functions, Haar functions, or wavelet bases. Adaptive basis functions include principal components and partial least squares components. Ramsay and Silverman [2] provide a detailed treatment of functional data analysis methodology. In the remainder of this section, we highlight the techniques that are directly relevant to the application of Section 1.

If the columns of $\Phi_{T\times K}$ ($K \le T$) are a proposed set of basis functions, then the original functional output from N model runs, an $N \times T$ matrix Y, can be rewritten as

$$Y - \bar{Y} \approx H\Phi^T, \tag{2}$$

where $\bar{Y} = N^{-1}11^T Y$ with 1 the N-vector of ones, or

$$y_i(t) - \bar{y}(t) \approx \sum_{k=1}^{K} h_{ik}\varphi_k(t) \quad \text{for } 1 \le i \le N,$$

where the mean function $\bar{y}(t)$ is computed as the mean of the $y_i(t)$ for each t. Equality holds in (2) if and only if the row space of $Y - \bar{Y}$ is a subspace of the column space of Φ.

Most standard basis systems are orthonormal. For example, the Legendre polynomials are orthonormal with respect to Lebesgue measure on [-1, 1]. But the Legendre polynomials in sin(t), which are used in the example below, are not orthonormal with respect to ordinary Lebesgue measure $d\theta$, but only with respect to a weighted measure $\cos\theta d\theta$. Adaptive bases functions may be orthonormal by construction, or not. Orthonormality of the basis functions is a nice property, since then the total variance is naturally partitioned among the variances of the coefficients:

$$\begin{aligned}\sum_{i=1}^{N}\|y_i - \bar{y}\|^2 &= \sum_{i=1}^{N}\left(\sum_{t=1}^{T}(y_i(t) - \bar{y}(t))^2\right)\\ &\approx \sum_{k=1}^{K}\left(\sum_{i=1}^{N}h_{ik}^2\right)\\ &= \sum_{k=1}^{K}\|h_k\|^2 \quad .\end{aligned} \tag{3}$$

(Usually the basis functions are ordered so that the first few capture most of the total variance.) However, even when the basis functions are not orthonormal, the total variance

captured by the expansion in terms of the first k ($k \leq K$) basis functions can be computed, and orthonormality may be less important than some other features when it comes to sensitivity analysis.

3. LEGENDRE POLYNOMIAL BASES

Since the example is being interpreted as a set of functions of angles from -90º to +90º, the Legendre expansion in sin(t) is a natural choice among standard expansions. Fig. 2 shows how the coefficients $\{h_{ik}\}$ of the expansions of the (N =) 81 functional outputs depend on the parameters, for k = 1, 2, ..., 6 (= K). The Legendre polynomials are alternately symmetric and anti-symmetric around zero, as shown in the top row of Fig. 2. The first k polynomials define a k-dimensional subspace of the (T =) 41-dimensional space in which the output functions are vectors. The percentages at the top show how much of the total variance in the original family of functions lies in this subspace for k up to 6. Note for future reference that the six-dimensional subspace defined by the first six polynomials still includes less than 90% of the total variance.

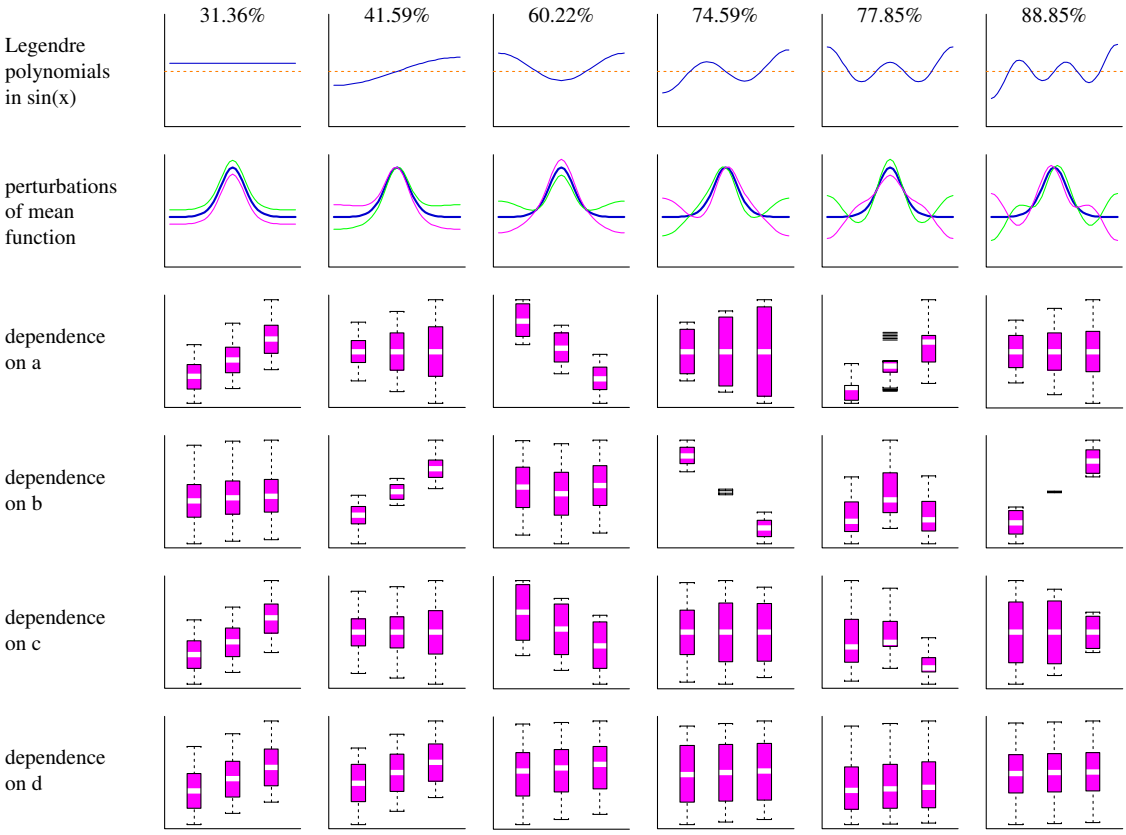


Figure 2. Dependence of the coefficients of the Legendre expansion on the parameters

In the second row, the Legendre polynomials are interpreted as perturbations of the overall mean of the 81 output functions. The mean function is the darker line. The mean plus and minus a multiple of the Legendre polynomial are the lighter lines.

The remaining rows contain box plots showing dependencies of the coefficients on the four parameters. Of course, we are not proposing sensitivity analysis by inspection-only as a serious method, but sensitivity analysis methodology is not the main goal of this paper. The figures are intended to suggest what more formal sensitivity analysis methods, such as described by McKay [3] and in Saltelli et al. [4], would indicate.

Variability in the coefficients of the Legendre polynomials of even order is controlled largely by *a*, although *c* and *d* influence the constant, zero-order term. The odd orders are controlled mostly by *b* with some influence of *d* on the first-order term.

An advantage of Legendre polynomials and other standard expansions is that they are well understood by many modelers. The other main advantage of using a consistent, non-adaptive basis system arises when a series of problems is being considered. The differences among corresponding analyses are then localized to the coefficients, instead of being partitioned out between the coefficients and the basis functions themselves.

The disadvantages arise in the case where the selected basis functions are not particularly well suited to the problem at hand. The Legendre polynomial basis, for example, is not a particularly good choice for a problem in which one of the main effects is neither symmetric nor antisymmetric, as in our example. The dispersion in the right-hand tail by comparison with the tight left-hand tail is not well captured by any single polynomial but spread out over several of them. The other disadvantage is that a relatively simple effect may be spread over several terms. For example, in this problem the effect of *b*, responsible for the left-right shift of the main peak, is spread out over all polynomials of odd order.

4. ADAPTIVE BASES COMPUTED BY PRINCIPAL COMPONENTS ANALYSIS

The principal components of *Y*, considered as *N* observations in a *T*-dimensional space, are themselves *T*-vectors, and are the eigenvectors of the $T \times T$ sample covariance matrix. They form an orthonormal basis for the *T*-dimensional space (or for a subspace of *T*-dimensional space, if $N < T$) that is specifically adapted to maximize the variance of the projection of the data onto the first basis vector, then onto the subspace spanned by the first and second basis vectors, etc. Thus expansions in the principal component basis for sensitivity analysis should at least achieve some information aggregation, avoiding one of the more serious problems with the Legendre polynomial, namely the allocation of a fairly simple effect (e.g., width changes or left-right shifts) to several components.

The principal component analysis (PCA) is shown in Fig. 3. For the family of curves in Fig. 1, the first principal component is basically an up-down shift, but unlike the first Legendre function this shift is not constant across all angles. The subspace spanned by this one function accounts for about 46% of the total variance in the family of curves, compared with about 31% for the Legendre polynomial of order zero. Like the zero-order Legendre coefficient, the coefficient of the first principal component depends on all four parameters. The second principal component for this example is a left-right shift accounting for another 34% of the total variance and controlled primarily by the *b* parameter. A similar amount of the

total variance was spread across the Legendre polynomials of odd orders. The third principal component is devoted explicitly to the right-hand tail and accounts for 11% of the total variance. It is clearly controlled by the *d* parameter, something that could not be extracted from the Legendre analysis.

These first three terms capture over 90% of the total variance, compared to seven terms required by the Legendre analysis. The fourth component, which accounts for another 5% of the total variance, is a symmetric kurtosis or tail-fattening component depending most strongly on *a* and *c*.

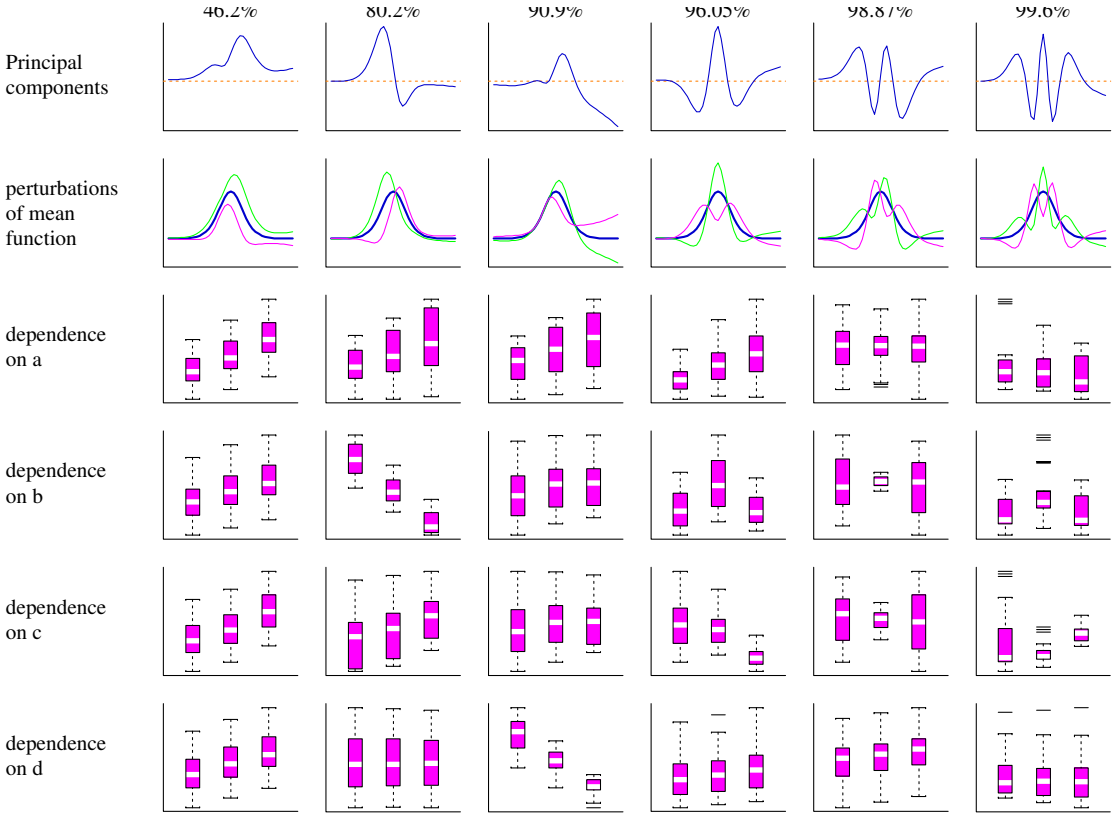


Figure 3. Dependence of the coefficients of PCA expansion on the parameters

5. ADAPTIVE BASES COMPUTED BY PARTIAL LEAST SQUARES

Partial least squares (PLS) regression was invented to handle near-collinearity among the independent variables, which is not usually a problem in analyzing computer experiments, assuming a reasonable experimental design. Thus, PLS is really a technique for decomposing the design matrix. (For a review of PLS regression, see Frank and Friedman [5].) However, PLS simultaneously provides a transformation of the dependent variables in such a way that the first PLS component of the dependent variables has the maximum variance that can be

predicted by a linear combination of the independent variables. The second PLS component is computed using the residuals from the prediction of the first, and has the maximum variance that can be predicted by a second, orthogonal component of the independent variables, etc. So one might think of PLS as "peeking" at the explanatory variables while doing something that is similar to a PC analysis of the dependent variable. Note that while the PLS components of the independent variables are orthogonal, the PLS components of the dependent variables are not, in general.

While there is no *a priori* guarantee that PLS results will be interesting for functional sensitivity analysis, we discuss them because they often seem to be fairly revealing. In particular, in the example they pull out some dependencies that were overshadowed by more important terms in both Legendre and principal component analyses.

The PLS components (Fig. 4) are somewhat more readily interpretable than the PCA components (Fig. 3). The first component is an up-down shift of the middle of the curve, depending as before on all four parameters. (The first PLS component should be the same as the first PCA component if the independent variables are standardized, which is to be recommended; it is only with the extraction of the second component that the algorithms diverge.) The second PLS component is a left-right shift, almost entirely a function of *b*, compared to the second PCA component which had more substantial contributions from *a* and *c* as well. The third PLS component is pure right-hand tail, dependent on *d*. The fourth is primarily a widthing term, although it also includes a small left-right shift component, and depends on *a* and *c*. As there are only four input parameters, the PLS algorithm can provide only four component vectors, but this four-dimensional subspace captures almost 96% of the total variability in this family of curves, which is almost as much as the first four PCA components. By comparison, the first four Legendre components captured only about 75% of the total variance.

The advantages and disadvantages of adaptive bases are pretty much the inverse of those for standard bases. The main advantage is good compaction or aggregation of the information; it is usually necessary to do sensitivity analysis on only the first few coefficients. The basis functions are also frequently more interpretable in physical terms. In a series of related problems, it may be interesting to study how the shapes of the component functions (as well as their coefficients) evolve. Of course, the down side to this is that shapes and coefficients are evolving simultaneously, which may lead to interpretation problems. In some cases it may make sense to pool all of the output functions for the series to extract a common set of PC or PLS components, so that the evolution of their coefficients through the series can be studied in the same way as the evolution of the coefficients of a fixed basis set, such as Legendre polynomials, could be examined.

6. OTHER CONSIDERATIONS

Penalty methods can be used to enforce a degree of smoothness on adaptive basis functions. Orthonormality is lost when this is done, but the results are probably more interpretable, and curve comparison across problems, or between model output and noisy data, becomes easier. Ramsay and Silverman [2] discuss the enforcement of smoothness in PCA (Chapter 7), and the technique is readily extended to PLS.

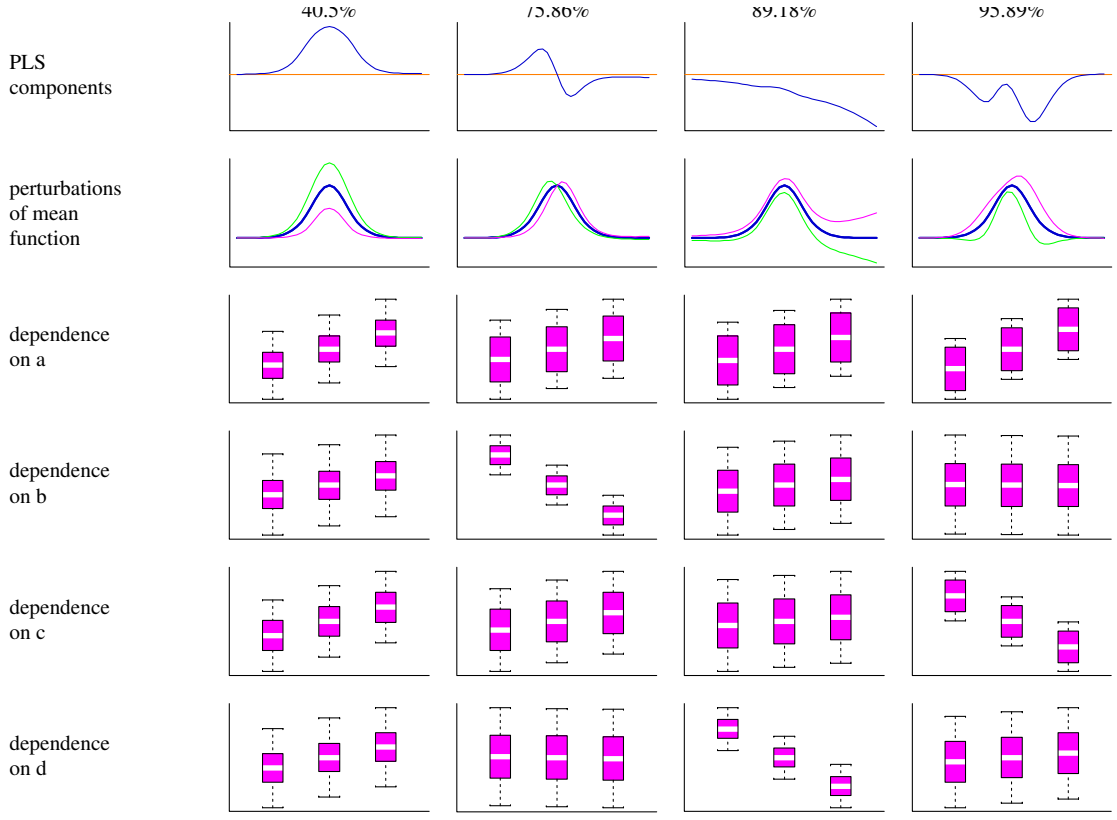


Figure 4. Dependence of the coefficients of PLS expansion on the parameters

Curve registration may be needed or advisable when the parameters affect the time- or space-scale or when the functions are not sampled at identical times in different runs. We would likely be interested in studying the sensitivity of the scaling and shifting to the input parameters, independently of the variability in the functional outputs after adjusting for these effects. Again, Ramsay and Silverman [2] address this problem in detail, proposing a series of methods from parametric location and/or scale change, through feature or landmark registration methods, to the estimation of general monotonic transformation.

7. SUMMARY

The purpose of this paper has been to suggest that sensitivity analysis for functional computer model outputs, correctly performed, is not significantly more difficult than for scalar outputs. The basic method is the expansion of the functional outputs in an appropriate functional coordinate system, i.e., in terms of an appropriate set of basis functions, followed by sensitivity analysis of the coefficients of the expansion using any standard method. The main art, then, is in choosing the appropriate coordinate system. We have considered both standard, pre-defined basis sets and data-adaptive basis sets. The example tends to favor the latter because of the aggregation and interpretability of the results, but the former may have value, depending on the problem or set of problems and the customer.

REFERENCES

1. Campbell K. “Functional sensitivity analysis of computer model output.” In: *Proc 7th Army Conf on Statistics*, Santa Fe, NM, 2001.
2. Ramsay JO, Silverman BW. *Functional data analysis*. Springer, New York 1997.
3. McKay MD. “Nonparametric variance-based methods of assessing uncertainty importance.” *Reliability Engineering and System Safety 1997*; *7*: 67-279.
4. Saltelli A, Chan K, Scott EM. (Eds.). *Sensitivity analysis*. John Wiley and Sons, Ltd. 2000.
5. Frank IE. Friedman JH. “A statistical view of some chemometrics regression tools.” *Technometrics* 1993;**35**: 109-135.

Decision-theoretic Sensitivity Analysis using Value of Information

Jeremy Oakley

Deparment of Probability and Statistics, University of Sheffield, The Hicks Building, Hounsfield Road, Sheffield, S3 7RH, UK;
E-mail: j.oakley@sheffield.ac.uk

1. INTRODUCTION

When a computer model is to be used to guide a decision, it is important for the decision-maker to acknowledge and investigate the uncertainty in the model. Typically, there will be uncertainty surrounding the true values of the input parameters in the model that should be used for the decision problem in question, and this then induces uncertainty in the output of the model. If the decision-maker considers their probability distribution for each unknown input in the model, they can then derive their probability distribution for the model output. The combination of their output distribution and an appropriate utility/loss function can then guide their decision.

In some cases, it may be possible to learn more about some or all of the uncertain input parameters before a final decision is made. In this case, it is then desirable to assess the importance of each uncertain input parameter in the model. Quantifying parameter importance is known as global or probabilistic sensitivity analysis. A measure of parameter importance that has been advocated previously is the variance-based measure (see 1). Variance-based measures consider the contribution of each uncertain input parameter to the variance of the model output. However, uncertainty about the model output as characterised by its variance is not necessarily equivalent to uncertainty about the optimum decision. Consequently, using variance-based measures to establish parameter importance in decision problems can in some cases produce misleading results, even as far as ranking the parameters in the wrong order of importance.

An alternative measure of parameter importance can be derived within the framework of utility theory. The idea is to determine whether different values of a particular input parameter lead to different optimum decisions, and if so, how much the expected utility/loss under alternative optimum decisions varies. Specifically, the expected utilty of learning the true numerical value of an uncertain input parameter before the decision is made can be calculated. This quantity is known as the partial expected value of perfect information (partial EVPI), and precisely quantifies the importance of an uncertain input variable. When the specific purpose of the model is to guide a decision within a clearly defined utility/loss structure, we advocate the partial EVPI as the single correct measure of an uncertain parameter's importance.

Further author information:
J.O.: Telephone: +44 (0)114 222 3853

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

In most practical situations, the decision-maker will not be able to learn the true value of an uncertain input parameter precisely, even if they desire to do so. The more likely possibility is that they may have the option of collecting more data to reduce their uncertainty about the unknown parameter. The expected value of perfect information framework can be extended to consider the expected value of collecting this data before making the decision; this is known as the expected value of sample information (EVSI). EVSI measures can then be used for deriving optimal sample sizes.

Both partial EVPIs and EVSIs can be computed using Monte Carlo methods. Unfortunately, to obtain these measures accurately, very large numbers of model evaluations are needed, potentially millions. For computationally expensive computer models, evaluating these measures may then require prohibitively lengthy computing times. However, in many cases it will be possible to exploit a feature of the computer model to dramatically speed up the computation; the function mapping inputs to output is often a smooth function. If the model is run at a particular set of input values and the output is observed, we will then also have information about the likely output at neighbouring sets of input parameter values.

When the time needed for a single run of the model is non-trivial, it can be highly advantageous to construct an *emulator*, a statistical approximation to the original computer model based on a fairly small number of different runs of that model. The emulator can then be used to give a fast approximation to the computer model regardless of the complexity of the model. An emulator is a regression model, and any regression technique can be employed. Our preferred option is the Gaussian process model. The Gaussian process emulator is a non-parametric approach that with the exception of continuity, makes no other assumptions about the functional form of the computer model. Gaussian processes have been used successfully before for efficient computation in other areas of sensitivity and uncertainty analysis. It will be demonstrated that the Gaussian process approach is of the order of 1000 times more efficient than Monte Carlo methods in terms of numbers of model runs, for computing partial EVPIs and EVSIs.

An application is given in the field of health economics. Economic models are used to estimate the cost-effectiveness of new treatments under consideration. A decision-maker will use the output of the model to help decide whether or not to approve the new treatment. There is always uncertainty regarding the values of the input parameters needed for the model; for example, it will not be known exactly how effective the treatment is, or what financial resources the patients on the treatment will use. There will be particular interest in conducting a probabilistic sensitivity analysis when using the model. It will often be possible to obtain more data regarding some of the model parameters, and hence reduce input uncertainty. Additionally, a certain class of models, known as patient simulation models, require an extensive simulation to produce the output for a single choice of input parameters. These models can be very computationally expensive, requiring in some cases in excess of an hour per run. In this scenario, emulator methods are essential for computation of EVPIs and EVSIs.

2. THE EXPECTED VALUE OF PERFECT INFORMATION

We now give a decision-theoretic measure of importance of an uncertain input variable. This measure is based on a standard result from decision theory (see for example 2), and was advocated by (3) and (4).

Suppose a decision-maker has to chose one decision d from a set of possible decisions $\mathcal{D}$. The decision-maker has a computer model to aid their decision, denoted by $y = f(\mathbf{x})$, where y is the model output and $\mathbf{x}$ are the model inputs. In the decision problem at hand, we suppose that there are 'true', uncertain values of the inputs that should be used in the model, and these uncertain values are denoted by $\mathbf{X}$ with distribution G. We then suppose that the utility of a decision d will be some function of the true output of the model, $f(\mathbf{X})$, and is denoted by $U\{d, f(\mathbf{X})\}$.

The decision maker then chooses the decision d to maximise their expected utility $E_{\mathbf{X}}[U\{d, f(\mathbf{X})\}]$. We can now define the expected utility of the optimum decision to be U^*, where

$$U^* = \max_d E_{\mathbf{X}}\{U\{d, f(\mathbf{X})\}\}. \tag{1}$$

Now suppose that the decision maker decides that they will learn the value of $\mathbf{X}$ before making their decision. Once they have learnt $\mathbf{X}$, their utility is then

$$\max_d U\{d, f(\mathbf{X})\}, \tag{2}$$

and so their expected utility of learning $\mathbf{X}$ (i.e., before they find out what $\mathbf{X}$ actually is) is

$$E_{\mathbf{X}}\{\max_d U\{d, f(\mathbf{X})\}\}. \tag{3}$$

The expected value of perfect information (EVPI) is then defined as the expected gain in utility:

$$E_{\mathbf{X}}\{\max_d U\{d, f(\mathbf{X})\}\} - \max_d E_{\mathbf{X}}\{U\{d, f(\mathbf{X})\}\}. \tag{4}$$

Now denote one of the uncertain input variables to be X_i. The same argument can be applied to derive the expected value of learning X_i before making the decision. Given X_i, we are still uncertain about the remaining input variables, $\mathbf{X}_{-i}$, and so we would choose the decision to maximise $E_{\mathbf{X}_{-i}|X_i}\{U\{d, f(\mathbf{X})\}\}$. The expected utility of learning X_i is then

$$E_{X_i}\left[\max_d E_{\mathbf{X}_{-i}|X_i}\{U\{d, f(\mathbf{X})\}\}\right], \tag{5}$$

and so the expected gain in utility, the partial EVPI of X_i is

$$E_{X_i}\left[\max_d E_{\mathbf{X}_{-i}|X_i}\{U\{d, f(\mathbf{X})\}\}\right] - \max_d E_{\mathbf{X}}\{U\{d, f(\mathbf{X})\}\}. \tag{6}$$

Here, we advocate the partial EVPI of X_i as a measure of importance of that variable in the model.

2.1. Computation

Partial EVPIs can be computed by Monte Carlo methods, but this can be computationally intensive and in some cases infeasible when a single run of the model takes a non-trivial amount of computing time. When the model is computationally expensive, a common approach is to use an *emulator*, a fast statistical approximation to the computer model based on regression (see 5; 6). This can be considerably more efficient than Monte Carlo when the output of the computer model is a smooth function of inputs. Full computational details for partial EVPI estimates using (Gaussian process) emulators are given in (7)

3. EXAMPLE: HEALTH ECONOMIC MODELLING AND THE GERD MODEL

One application area in which partial EVPIs are currently used is health economics. In health economics, the interest is in allocating health care resources as effectively as possible. The decision problem is to choose which out of set of competing treatments for an illness is the most cost-effective. Cost-effectiveness of a treatment is described with a single (financial) measure known as the *net benefit* of the treatment, and net benefits are often predicted using computer models. (This is because clinical trials typically only record information on clinical effectiveness). The models invariably require specification of parameters that are uncertain, and so there is interest in investigating the consequences of this input uncertainty. The utility of choosing a particular treatment is then given by the net-benefit of the treatment, and so it is possible to measure the importance of each input using partial EVPIs.

We give the example used in (7) that also illustrate the efficiency of the emulator approach in the computation of partial EVPIs. The model compares treatment strategies for gastroesophageal reflux disease. In this example, we suppose that a decision has to be made regarding the adoption of one of three treatment strategies:

1. Acute treatment with proton pump inhibitors (PPIs) for 8 weeks, then continuous maintenance treatment with PPIs at the same dose.

2. Acute treatment with PPIs for 8 weeks, then continuous maintenance treatment with hydrogen receptor antagonists (H2RAs).

3. Acute treatment with proton pump inhibitors PPIs for 8 weeks, then continuous maintenance treatment with PPIs at the a lower dose.

The model was presented in (8). In the scenario that we are considering, there are twenty-three uncertain inputs, relating to quantities such as probabilities of healing and recurrence of the symptoms with each treatment, and resources used by patients such as number of visits to a general practitioner. Distributions for all the uncertain inputs are described in (9). The output of the model can be converted into a utility for each treatment.

Using 600 runs of the model, we estimate the partial EVPI of each patient. The GERD model is computationally cheap, so we can determine the true partial EVPIs based on

massive Monte Carlo samples (several hundred million in this case). Although there is some inaccuracy in the estimates, we have identified all the influential inputs in the model, to within what we believe would be an acceptable order of magnitude. For comparison, we also estimate the partial EVPIs using a combination of Simpson's rule and Monte Carlo as described in section 2.

We give the actual values of the estimates and true values of the partial EVPIs for the six most important variables in table 1.

uncertain input parameter	true partial EVPI	Gaussian process estimate	Simpson/MC estimate estimates
hazard for healing on PPIs	1.286	1.194	3.465
no. of symptom weeks after surgery	2.271	2.500	4.229
Recurrence probability on PPIs (6-12) months	4.905	4.579	5.507
Recurrence probability on H2RAs (0-6) months	21.221	20.908	23.417
Recurrence probability on H2RAs (6-12) months	2.652	2.666	2.958
Recurrence probability on low dose PPIs (6-12) months	3.473	3.378	3.846

Table 1. True values, Gaussian process estimates and Simpson/Monte Carlo estimates of the partial EVPIs of the six most influential input variables. The Gaussian process estimates are based on 600 model runs, and the Simpon/Monte Carlo estimates are based on 410200 model runs.

These partial EVPIs can then be interpreted as (financial) values of learning the value of the corresponding parameter before choosing which treatment to use for the patient population. The figure represents dollars per patient, and so needs to be multiplied by the size of the patient population to give a final value.

References

[1] A. Saltelli, K. Chan, and M. Scott, editors. *Sensitivity Analysis*. Wiley, New York, 2000.

[2] A. O'Hagan. *Kendall's Advanced Theory of Statistics, Volume 2B, Bayesian Inference*. Edward Arnold, London, 1994.

[3] J. C. Felli and G. B. Hazen. Sensitivity analysis and the expected value of perfect information. *Medical Decision Making*, 18:95–109, 1998.

[4] K. Claxton. Bayesian approaches to the value of information: implications for the regulation of new health care technologies. *Health Economics*, 8:269–274, 1999.

[5] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statist. Sci.*, 4:409–435, 1989.

[6] A. O'Hagan, M.C. Kennedy, and J. E. Oakley. Uncertainty analysis and other inference tools for complex computer codes (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 503–524. Oxford: University Press, 1999.

[7] J. E. Oakley. Sensitivity analysis for computationally expensive economic models. Technical report, Department of Probability and Statistics, University of Sheffield, 2003.

[8] R. Goeree, B. O'Brien, R. Hunt, G. Blackhouse, A. Willan, and J. Watson. Economic evaluation of long term management strategies for erosive oesophagitis. *PharmacoEcon.*, 16:679–697, 1999.

[9] A. H. Briggs, R. Goeree, G. Blackhouse, and B. J. O'Brien. Probabilistic analysis of cost-effectiveness models: Choosing between treatment strategies for gastoesophageal reflux disease. *Med Decis Making*, 22:290–308, 2002.

A Second Order Differential Importance Measure for Reliability and Risk Applications

Enrico Zio, Luca Podofillini

Department of Nuclear Engineering, Polytechnic of Milan
Via Ponzio 34/3, 20133 Milan, Italy
enrico.zio@polimi.it, luca.podofillini@polimi.it

Abstract: The Differential Importance Measure (DIM) is a first-order sensitivity measure that ranks the parameters of the risk model according to the fraction of total change in the risk that is due to a small change in the parameters' values, taken one at a time. However, the DIM does not account for the effects of interactions among components. In this paper, a second-order extension of the DIM, named DIM^{II}, is proposed for accounting of the interactions of pairs of components when evaluating the change in system performance due to changes of the reliability parameters of the components. A numerical application is presented in which the informative contents of DIM and DIM^{II} are compared.

Keywords: Differential Importance Measure, Joint Importance, second order sensitivity measure.

1. INTRODUCTION

A limitation of the Importance Measures (IM) [1-3] currently used in reliability and risk analysis is that they rank only individual components or basic events whereas they are not directly applicable to combinations or groups of components or basic events [2]. In practice different basic events may, for example, represent different modes of failure or unavailability of a single component and in order to determine the importance of such component one has to consider all the related basic events as a group. Furthermore, many risk-informed applications deal with evaluating the risk change associated to changes in the plant technical specifications (surveillance and/or test frequencies, etc): such changes may indeed impact a group of components. To partially overcome this limitation, recently, the Differential Importance Measure, DIM, has been introduced for use in risk-informed decision making [3]. The DIM is a first-order sensitivity measure that ranks the parameters of the risk model according to the fraction of the total change in the risk that is due to a small change in the parameters' values, taken one at a time. The DIM bears an important property of additivity: the DIM of a group of components or basic events is the sum of the DIMs of the single components or basic events of the group. However, since DIM considers risk changes due to small changes of the parameters' values, it does not account for interactions among components.

The need for IMs capable of considering combinations of components arises also when planning a budget-constrained improvement in the reliability of a system design for example by replacing one of its components with a better-performing one, or by inspecting and maintaining it more frequently. Due to the budget constraints, the improvement may need to

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

be accompanied by the sacrifice of the performance of another, less important component. The interactions of these coupled changes to system design must be accounted for when assessing the importance of the system components. To this aim, second order sensitivity measures such as the Joint Reliability Importance (JRI) and Joint Failure Importance (JFI) measures have been introduced [4, 5].

In this paper, a second-order extension of the DIM, named DIM^{II}, is proposed for accounting of the interactions of pairs of components when evaluating the change in system performance due to changes of the reliability parameters of the components. The extension aims at supplementing the first-order information provided by DIM with the second-order information provided by JRI and JFI. Obviously, the need of resorting to information on second-order effects depends on the magnitude of the changes of the parameters values and on the non linearity of the system.

2. EVALUATING THE CHANGE IN THE SYSTEM PERFORMANCE

We consider a system of n components. Let O be a generic measure of the system performance (e.g unreliability, unavailability, risk, etc., depending on the application at hand). The performance O is a function of the components' unavailabilities (or failure probabilities) q_i, i=1, 2, …, n, i.e. $O=g_q(q_1, q_2, \ldots, q_n)$. A change in system performance due to arbitrary changes in the values of the q_i, i=1, 2, n can be expanded in McLaurin series as:

$$\Delta O = \sum_{i=1}^{n} \frac{\partial O}{\partial q_i} \Delta q_i + \sum_{i=1}^{n} \sum_{h>i=1}^{n} \frac{\partial^2 O}{\partial q_i \partial q_h} \Delta q_i \Delta q_h + \sum_{i=1}^{n} \sum_{h>i=1}^{n} \sum_{k>h=1}^{n} \frac{\partial^3 O}{\partial q_i \partial q_h \partial q_k} \Delta q_i \Delta q_h \Delta q_k + \ldots \quad (1)$$

Using the rare event approximation, the risk measure O can be written in terms of the probabilities of the n_{cs} minimal cutsets:

$$O \approx \sum_{j=1}^{n_{cs}} M_j \quad (2)$$

where M_j is the probability of the j-th cutset. Then, alternatively, the change in O due to generic changes of the parameters Δq_i, i=1, 2, …, n is [6]:

$$\Delta O = \sum_{i=1}^{n} S_i \Delta q_i + \sum_{i=1}^{n} \sum_{h>i=1}^{n} S_{ih} \Delta q_i \Delta q_h + \sum_{i=1}^{n} \sum_{h>i=1}^{n} \sum_{k>h=1}^{n} S_{ihk} \Delta q_i \Delta q_h \Delta q_k + \\ + \ldots + \sum_{i=1}^{n} \sum_{h>i=1}^{n} \ldots \sum_{r>s}^{n} S_{ih\ldots r} \Delta q_i \Delta q_h \ldots \Delta q_r \quad (3)$$

where $S_i = \sum_{j=1}^{n_{cs}} \frac{\partial M_j}{\partial q_i}$, $S_{ih} = \sum_{j=1}^{n_{cs}} \frac{\partial^2 M_j}{\partial q_i \partial q_h}$, and so on. Eq. (1) reduces to eq. (3) if the rare event approximation of eq. (2) holds. The right-hand part of eq. (3) contains as many terms as the largest number of components in any minimal cutset. The quantities S_i, S_{ih}, $S_{ih\ldots r}$ can be straightforwardly calculated as follows [6]: S_i is the sum of the contributions to O in eq. (2) of the minimal cutsets containing element i, with its unavailability set to 1; S_{ih} is the sum of the contributions to O of the minimal cutset containing elements i and h with their unavailabilities set to 1, $S_{ih\ldots r}$ is the sum of the contributions of the minimal cutset containing elements i, h,

…, *r* with their unavailabilities set to 1. Note from eq. (3) that the interaction terms S_{ih}, $S_{ih\ldots r}$ assume a value of zero if the components *i* and *h*, *i*, *h*, … and *r*, respectively do not appear together in one of the minimal cutsets. Thus, for example when the rare event approximation holds, for groups of components belonging to different blocks in series only the first-order terms in eq. (3) contribute to ΔO since they do not appear together in any minimal cutset. On the contrary, for components in parallel logic, contributions from the higher-order terms in eq. (3) are expected, since the components always appear together in a minimal cutset.

3. FIRST-ORDER IMPORTANCE MEASURES: BIRNBAUM AND DIM

The Marginal Reliability Importance (MRI) (often referred to as the Birnbaum IM) of component *i* is defined with respect to its unavailability q_i as [1, 4]:

$$MRI(i) = \frac{\partial O}{\partial q_i} \tag{4}$$

According to the MRI, components for which a variation in unavailability results in the largest variation of the system performance have the highest importance.

The MRI applies when the components' unavailabilities or failure probabilities q_i, i=1, 2, …, n are known explicitly. However, the quantities q_i are often expressed in terms of additional reliability parameters x_k, k=1, 2,…, n_p such as failure and repair rates, maintenance and inspection frequencies, etc: in turn, the system performance *O* can be expressed in terms of the parameters x_k, i.e. $O = g_x(x_1, x_2, ..., x_{n_p})$. Furthermore, the MRI applies to single components. However, the changes may affect a number of components at the same time. For example, a change in a maintenance frequency will affect the unavailabilities of all of the components that undergo that particular maintenance policy.

Recently, the Differential Importance Measure (DIM) has been introduced to quantify the importance of the parameters x_k entering the system performance model [3]. DIM considers the total variation of the output function *O* due to a small variation of its parameters, taken one at a time. If the variation of the parameter is small enough, the variation of *O* is the total differential *dO*:

$$dO = \sum_{k=1}^{n_p} \frac{\partial O}{\partial x_k} \cdot dx_k \tag{5}$$

The DIM of the parameter x_l, $DIM(x_l)$, is defined as the fraction of the total change in *O* which pertains to the change in the parameter x_l:

$$DIM(x_l) = \frac{dO_{x_l}}{dO} = \frac{\frac{\partial O}{\partial x_l} \cdot dx_l}{\sum_{k=1}^{n_p} \frac{\partial O}{\partial x_k} \cdot dx_k} \tag{6}$$

The DIM is additive in the sense that the DIM of a subset of parameters x_r, x_s, .., x_t, is [3]: $DIM(x_r \cup x_s \cup \ldots \cup x_t) = DIM(x_r) + DIM(x_s) + ... + DIM(x_t)$.

The DIM can be useful in risk-informed applications involving the quantification of risk changes in O due to proposed changes of a plant technical specification, e.g. a surveillance/test/maintenance frequency. Being a first order local sensitivity measure, the DIM can be used to forecast a finite change ΔO due to any change in the parameters' values only provided that these latter changes are small enough to be used in (6). Only in this case, in fact, the higher-order contributions to ΔO in eqs. (1) and (3), which describe the interactions due to simultaneous change in pairs of parameters, triplets, etc. can be neglected. The effects of these interactions are illustrated in the next Section with reference to pairs of components.

4. JOINT FAILURE AND RELIABILITY IMPORTANCES

To evaluate quantitatively the interaction between components, the concepts of Joint Failure Importance (JFI) and of Joint Reliability Importance (JRI) of pairs of components have been introduced as an extension to the single-component MRI [4, 5]. JFI is introduced when the considered system performance O is a measure of the system loss (i.e. unreliability, unavailability, risk, etc.) and it is expressed in terms of the components' unavailabilities q_i, i=1, 2, …, n. JRI refers to the case in which O is a measure of the system gain (i.e. reliability, availability, etc.) and is expressed in terms of the components' availabilities p_i=1-q_i. JFI and JRI for components i and h are defined as:

$$JFI(i,h) = \frac{\partial^2 O}{\partial q_i \partial q_h}; \qquad JRI(i,h) = \frac{\partial^2 O}{\partial p_i \partial p_h} = -JFI(i,h) \tag{7}$$

An interesting property of the joint importance measures is the possibility of determining the sign of JFI(i,h) and JRI(i,h) based on the relative logical position of components i and h within the system. In particular [4]:

JFI(i,h) ≥ 0 (JRI(i,h) ≤ 0) for components in parallel

JFI(i,h) ≤ 0 (JRI(i,h) ≥ 0) for components in series

More generally:

JFI(i,h) ≥ 0 (JRI(i,h) ≤ 0) if components i and h appear together in at least one minimal cut-set but not in any minimal path-set.

JFI(i,h) ≤ 0 (JRI(i,h) ≥ 0) if components i and h appear together in at least one minimal path-set but not in any minimal cut-set.

Joint importance measures are useful to quantify the interactions of components with respect to the system performance. Awareness of such interactions among components is useful when the analysts are interested in evaluating the effects on the system of modifications regarding two components or, in a more general sense, two parameters (e.g. failure rates, maintenance periods, etc). Indeed, when planning a modification of a reliability parameter of a component towards a better performance (e.g. replacing it with a better-performing one, inspecting or maintaining it more frequently) one is often forced, by budget constraints, to sacrifice the performance of another.

5. SECOND ORDER DIFFERENTIAL IMPORTANCE MEASURE

For finite changes in the components unavailabilities Δq_i, it may be relevant to evaluate also the second order contribution to ΔO. With reference to components i and h, the variation of O, ΔO_{ih}, due to the variations of the parameters Δq_i and Δq_h is:

$$\Delta O_{ih} = \frac{\partial O}{\partial q_i}\Delta q_i + \frac{\partial O}{\partial q_h}\Delta q_h + \frac{\partial^2 O}{\partial q_i \partial q_h}\Delta q_i \Delta q_h = MRI(i)\Delta q_i + MRI(h)\Delta q_h + JFI(i,h)\Delta q_i \Delta q_h \quad (8)$$

A second order DIM, DIM^{II} can thus be defined as:

$$DIM^{II}(i,h) = \frac{\Delta O_{ih}}{\Delta O^{II}} \quad (9)$$

Note that, as stated in Section 2, when the rare event approximation can be used in the system modeling, one can neglect the computation of the JFI of pairs of components if the they do not belong to the same minimal cutset.

6. NUMERICAL EXAMPLE

6.1. Comparing the information of DIM and DIM^{II}

Consider the system of Figure 1. As system performance O we consider its limit unavailability. The components' unavailabilities are $q_1=q_2=10^{-3}$, $q_3=q_4=q<<1$. In this numerical example we will compare the informative content of the measures DIM and DIM^{II} when assessing the effect on the system performance O of changes in the components' unavailabilities of pairs of components.

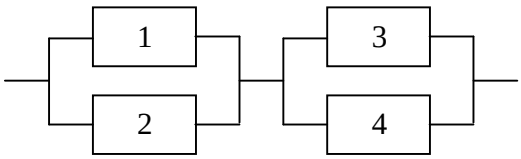


Figure 1. System reliability block diagram

Consider at first the behavior of DIM(1)=DIM(2) and DIM(3)=DIM(4) as functions of the parameter $q=q_3=q_4$ in the interval $(5\cdot10^{-4}, 2\cdot10^{-3})$ shown in Figure 2. The change in the parameters' values is $\Delta q_i=10^{-3}\cdot q_i$, for i=1, 2, 3, 4. For $q=10^{-3}$ (i.e. $q_1=q_2=q_3=q_4$), the four components have the same DIM, due to the symmetry of the system. Then, as expected, in the cases $q\neq q_1=q_2$ the most unavailable components are the most important according to the DIM. Indeed, DIM(1)=DIM(2) > DIM(3)=DIM(4) for $q_1=q_2=10^{-3}> q=q_3=q_4$ and DIM(1)=DIM(2) < DIM(3)=DIM(4) for $q_1=q_2=10^{-3}<q= q_3=q_4$.

From the additivity property [3], the first order DIMs for the pairs of components are:

$$DIM(1,2) = DIM(1) + DIM(2) = 2\cdot DIM(1) \quad (10)$$

DIM(3,4) = DIM(3) + DIM(4) = 2·DIM(3)

DIM(1,3) = DIM(1) + DIM(3) = DIM(2) + DIM(4) = DIM(2,4)

The behavior of the DIMs for the above pairs of parameters in the system is reported in Figure 3 as functions of q. Note that for $q=10^{-3}=q_1=q_2$, all measures DIM(1,2), DIM(1,3), DIM(2,4), are equal, due to the symmetry of the system and to the fact that the DIM considers the variation of one parameter value at a time. Still, when varying the values of the unavailabilities of two components simultaneously, one would expect a difference between the case of a pair of components in parallel, say (1, 2), and the case of a pair of components in series, say (1, 3). This difference can be traced by considering second-order interactions among the components, i.e. the DIM^{II}. In the cases $q \neq q_1=q_2$, the measures DIM(1,2)=2· DIM(1) and DIM(3,4)=2· DIM(3) duplicate the behavior of DIM(1)=DIM(2) and DIM(3)=DIM(4). Instead, DIM(1,3) =DIM(2,4) is independent on q. Indeed,

$$DIM(1,3) = \frac{DIM(1)+DIM(3)}{\sum_{j=1}^{n} DIM(j)} = \frac{DIM(1)+DIM(3)}{2\cdot(DIM(1)+DIM(3))} = 0.5 \tag{11}$$

Let us now compute the second-order sensitivity coefficients (eq. (7)):

$$JFI(1,2) = \frac{\partial^2 O}{\partial q_1 \partial q_2} = 1 - q_3 q_4 = 1 - q^2 \qquad JFI(1,3) = -q_2 q_4 = -10^{-3} q$$

$$JFI(3,4) = 1 - q_1 q_2 = 1 - 10^{-6} \qquad JFI(2,4) = -q_1 q_2 = -10^{-3} q$$

As expected [4], JFI<0 for the components 1 and 3 in series and JFI>0 for the components 1 and 2 and 3 and 4 in parallel. Furthermore, as anticipated in Section 2, the absolute value of the JFI(1, 3) of components 1 and 3 in series is smaller than that of the two parallel pairs (1,2) and (3, 4). Indeed the values of the unavailabilities q_i are such that the rare event approximation in eq. (2) certainly holds. This would suggest that, actually, one could neglect the contribution corresponding to the interaction term of the couple of components in series.

Figure 3 also reports the measures DIM^{II}(1, 2), DIM^{II}(1, 3) and DIM^{II}(3, 4) (symbols ◊, dots and *, respectively). Due to the small values of the variation of the parameters considered ($\Delta q_i = 10^{-3} \cdot q_i$ i=1, 2, 3, 4), the measures DIM^{II} do not differ appreciably from the DIM, since the contribution to the total change in the output performance O of the second-order terms are negligible. In the same Figure the DIM and DIM^{II} measures are reported for pairs of components in correspondence of larger values of the relative parameters' change, $\Delta q_i/q_i$, i=1, 2, 3, 4. As expected, the values of DIM and DIM^{II} differ progressively when higher values of $\Delta q_i/q_i$ are considered. As a general observation, DIM^{II} differs significantly from DIM for the pair (1,3) of components in series logic, whereas DIM^{II} reproduces the behaviour of DIM for the pairs (1,2) and (3,4) of components in parallel logic. This behaviour could seem unexpected since, as above stated, the interaction term JFI of components in series logic is negligible, whereas that of the components in parallel logic is large. In words, this fact can be explained as follows. In practice, due to the values of JFI(1,3)≈0 and JFI(1,2)≈1 (and JFI(3,4)≈1) we can write for DIM^{II}(1,3) and DIM^{II}(1,2):

$$DIM^{II}(1,3) \cong \frac{\frac{\partial O}{\partial q_1}\Delta q_1 + \frac{\partial O}{\partial q_3}\Delta q_3}{\Delta O^{II}}; \qquad DIM^{II}(1,2) \cong \frac{\frac{\partial O}{\partial q_1}\Delta q_1 + \frac{\partial O}{\partial q_2}\Delta q_2 + \Delta q_1 \Delta q_2}{\Delta O^{II}} \tag{12}$$

Due to the small value of the interaction term JFI(1,3)≈0, the numerator of $DIM^{II}(1,3)$ is equal to that of the DIM(1,3) for any value of $\Delta q_i/q_i$, whereas the denominator, which accounts for all of the JFIs, progressively increases for increasing values of $\Delta q_i/q_i$. As a consequence of this fact, the value of $DIM^{II}(1,3)$ is progressively shifted downwards for increasing values of $\Delta q_i/q_i$. Instead, as for the pair (1,2), both the numerator and the denominator of eq. (12) change their values from those of the numerator and the denominator of DIM(1,2), but the change is such that the ratio is approximately independent on Δq_i.

Let us first consider the case $q=10^{-3}=q_1=q_2$. While still DIM(1,2)=DIM(1,3)= DIM(3,4)= 0.5 by construction, $DIM^{II}(1,2) = DIM^{II}(3,4) > DIM^{II}(1,3)$ (Table 1). The ranking produced by the measure DIM^{II} suggests that increasing simultaneously the unavailabilities of the pairs of components in parallel logic (1,2) or (3,4) has a greater impact on the system unavailability than the same action performed on the pairs of components in series (1,3). This result is physically reasonable. An increase in unavailability of two components has more effect on the system unavailability if performed on components on the same node (i.e. in parallel) rather than on components on different nodes (i.e. in series). Indeed, with reference to the values of the ΔO_{ih}^{II} reported in Table 2, in the former situation the change in components unavailabilities is more critical since it impacts components on the same node, thus creating a system bottleneck. Instead, the latter situation is less critical for the system unavailability since the increase in components unavailability is shared by the two nodes. Table 1 also reports the values of DIM and DIM^{II} corresponding to $q=9\cdot10^{-4}$ and $q=1.2\cdot10^{-3}$, for the case $\Delta q_i/q_i$ =0.5. As for the case $q=9\cdot10^{-4}$, DIM(1, 2)>DIM(1, 3)>DIM(3, 4), whereas DIM^{II}(1, 2)>DIM^{II}(3, 4)>DIM^{II}(1, 3). Again, DIM considers the contribution to ΔO arising from a change in the unavailability of one of the components at a time. Therefore, the pair (1,2) results the most important according to this measure, since the two components 1 and 2, have the largest values of the first-order DIM (Figure 2), being more unavailable than components 3 and 4 ($q_1=q_2=10^{-3}>q_3=q_4=9\cdot10^{-4}$). The ranking provided by DIM^{II} is different and it reflects again that an increase in the unavailabilities of two components has more effect on the system unavailability if performed on components on the same node rather than on components on different nodes (refer also to the values of ΔO_{ih}^{II} reported in Table 2). This leads to the ranking inversion between the pairs (3,4) and (1,3). Similar considerations apply to the case of $q=1.2\cdot10^{-3}$.

Table 1. Values of DIM and DIM^{II} for the pairs of components (1,2), (1,3) and (3,4) for different values of q and $\Delta q_i/q_i$, i=1, 2, …, n

	$\Delta q_i/q_i$				DIM			DIM^{II}		
q	i=1	i=2	i=3	i=4	(1,2)	(1,3)	(3,4)	(1,2)	(1,3)	(3,4)
10^{-3}	0.5	0.5	0.5	0.5	0.50	0.50	0.50	0.50	0.40	0.50
$9\cdot10^{-4}$	0.5	0.5	0.5	0.5	0.55	0.50	0.45	0.55	0.40	0.45
$1.2\cdot10^{-3}$	0.5	0.5	0.5	0.5	0.41	0.50	0.59	0.41	0.40	0.59
10^{-3}	-0.5	-0.5	-0.5	-0.5	0.50	0.50	0.50	0.50	0.67	0.50
10^{-3}	0.5	-0.5	-0.5	+0.5	0.50	0.50	0.50	0.50	0.00	0.50

Table 2. Values of ΔO_{ih} and ΔO_{ih}^{II} for the pairs of components (i, h), for different values of q and $\Delta q_i/q_i$, i=1, 2, …, n

	$\Delta q_i/q_i$				ΔO_{ih}			ΔO_{ih}^{II}		
q	i=1	i=2	i=3	i=4	(1,2)	(1,3)	(3,4)	(1,2)	(1,3)	(3,4)
10^{-3}	0.5	0.5	0.5	0.5	10^{-6}	10^{-6}	10^{-6}	$1.25\cdot10^{-6}$	10^{-6}	$1.25\cdot10^{-6}$
$9\cdot10^{-4}$	0.5	0.5	0.5	0.5	10^{-6}	$9.05\cdot10^{-7}$	$8.10\cdot10^{-7}$	$1.25\cdot10^{-6}$	$9.05\cdot10^{-7}$	$1.01\cdot10^{-6}$
$1.2\cdot10^{-3}$	0.5	0.5	0.5	0.5	10^{-6}	$1.22\cdot10^{-6}$	$1.44\cdot10^{-6}$	$1.25\cdot10^{-6}$	$1.22\cdot10^{-6}$	$1.80\cdot10^{-6}$
10^{-3}	-0.5	-0.5	-0.5	-0.5	-10^{-6}	-10^{-6}	-10^{-6}	$-7.50\cdot10^{-7}$	-10^{-6}	$-7.50\cdot10^{-7}$
10^{-3}	0.5	-0.5	-0.5	+0.5	0	0	0	$-2.50\cdot10^{-7}$	0	$-2.50\cdot10^{-7}$

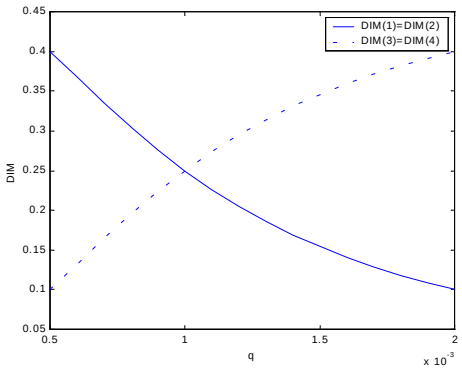


Figure 2. Values of DIM(1)=DIM(2) and DIM(3)=DIM(4) for different values of q, $\Delta q_i=10^{-3}q_i$

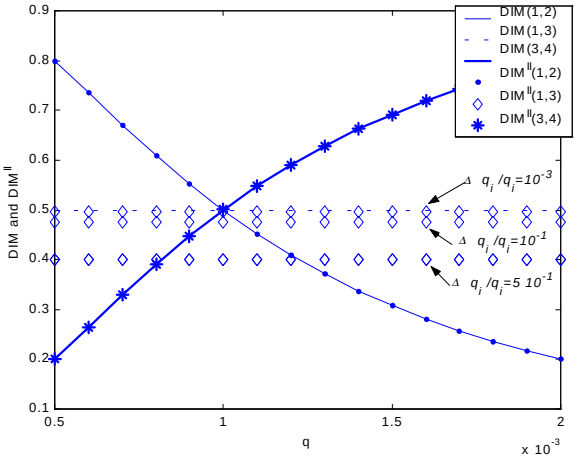


Figure 3. Values of DIM and DIM^{II} for the pairs of components (1,2), (1,3)=(2,4) and (3,4) for different values of q. Case of $\Delta q_i = 10^{-3}\, q_i$, i=1, 2, …, n

6.2. Use of DIM^{II} in risk-informed applications

In risk-informed applications, the information provided by DIM^{II} is handled by a decision-maker in different ways, depending on his/her goals.

Consider firstly the case of an analyst interested in reducing the costs associated to the system operation by replacing two components with two less expensive, but also less

performing, ones or by extending their maintenance frequencies. For example, with reference to the case of $q=10^{-3}=q_1=q_2$ and $\Delta q_i/q_i=0.5$, the performance of the pair of components (1, 3) can be sacrificed if required by budget constraints with minor consequences on the system unavailability, contrary to the case of acting on the pairs of components (1, 2) or (3, 4) in parallel logic, as $DIM^{II}(1,2) = DIM^{II}(3,4) > DIM^{II}(1,3)$. Note that this conclusion can be inferred only on the basis of the ranking produced by the second-order measure DIM^{II}, as DIM(1, 2)=DIM(1, 3)=DIM(2 ,4).

Consider now the case in which the analyst is interested in identifying the pairs of components to be improved to get the largest improvement in system performance. In the case of $q=10^{-3}=q_1=q_2$ and $\Delta q_i/q_i= - 0.5$, i=1, 2, 3, 4 the values of the DIM^{II} (reported in Table 1) suggest that the improvement efforts should be devoted to the pair of components (1, 3) in series, characterized by the largest value of DIM^{II}, and thus leading to the largest reduction in system unavailability (see also the values of ΔO_{ih}^{II} in Table 2). Again, this result is obvious from the physical viewpoint: the improvement has more beneficial effects on the system availability if performed on components on different nodes (i.e. in series) rather than on components on the same node (i.e. in parallel). Indeed, in the latter case, the improvement would be less effective due to the presence of the other non-improved node in series, which remains an unvaried system bottleneck.

Another situation that can occur in risk-informed decision-making arises from the fact that, in practice, the analyst has often to cope with a constrained budget that might forbid spending resources on two components of a pair. Thus, in this case the final decision of the analyst must be a trade-off between improving the availability of a component while worsening that of another, still with the goal of attaining the largest improvement in the system availability. In this case, the analyst is looking at changes in the components unavailabilities Δq_i and Δq_h with opposite signs (i.e. if $\Delta q_i>0$ then $\Delta q_h<0$ and viceversa) and, thus, it is preferable to act on pairs of components with JFI>0. In this case, if we refer again to the case $q=q_3=q_4= 10^{-3}=q_1=q_2$, for the generic pair ($i$, h), the net contribution of the first-order terms of eq. (1) equals zero since MRI(i)=MRI(h) and $\Delta q_i = -\Delta q_h$ and the system output variation ΔO_{ih}^{II} becomes:

$$\Delta O_{ih}^{II} = \frac{\partial^2 O}{\partial q_i \partial q_h} \Delta q_i \Delta q_h = JFI(i,h) \Delta q_i \Delta q_h \leq 0 \tag{13}$$

The values of the DIM(i, h) and DIM^{II}(i, h) and of the corresponding ΔO_{ih} and ΔO_{ih}^{II} for the pairs (i, h)=(1, 2), (1, 3) and (2, 4) are reported in Table 1 and Table 2 respectively with reference to the case $q=q_3=q_4= 10^{-3}=q_1=q_2$, $\Delta q_i/q_i=0.5$, i=1, 2, 3, 4. Evidently, those pairs of components with JFI>0, i.e. (1, 2) and (3, 4), are characterized by negative contributions ΔO_{ih}^{II}, corresponding to an increase in system unavailability. Indeed, if the unavailabilities of two components in parallel logic are changed in opposite directions, then, since the components with the lowest unavailability determines the unavailability of the pair, the overall system unavailability decreases. On the contrary, if the unavailabilities of two components belonging to different nodes in series are changed in opposite directions, then due to the weak interactions among the components (JFI≈0 in eq. (25)) the system unavailability remains basically unchanged.

CONCLUSIONS

This paper considers the differential importance measure, DIM, and the Joint Failure Importance measure, JFI, recently introduced in literature. The DIM is a first-order sensitivity measure that ranks the parameters of the risk model according to the fraction of the total change in the risk that is due to a small change in the parameters' values, taken one at a time, and, by construction, it does not account for second-order interactions among components. Instead, the JFI measure is a second order sensitivity measure, which considers the interactions of coupled changes to system design.

In this paper, a second-order extension of the DIM, named DIM^{II}, is proposed for accounting of the interactions of pairs of components when evaluating the change in system performance due to changes of the reliability parameters of the components. The extension aims at supplementing the first-order information provided by DIM with the second-order information provided by JRI and JFI.

A numerical application is presented in which the informative contents of DIM and DIM^{II} are compared. The results confirm that in certain cases when second-order interactions among components are accounted for, the importance ranking of the components may differ from those produced by a first-order sensitivity measure. Obviously, the need of resorting to information on second-order effects depends on the magnitude of the changes of the parameters values and on the non linearity of the system.

It is shown in the paper that in some applications it is possible to determine a priori whether the interaction term in DIM^{II} can be neglected even for large changes in the parameters, thus avoiding the computation of the JRI and JFI measures for all of the possible pairs of components. In particular, second-order interactions among components are negligible if the components do not appear together in the same minimal cutset. Furthermore, guidelines for the use of DIM^{II} in risk-informed decision-making are provided for different cases.

REFERENCES

1. Birnbaum L.W., On the importance of different elements in a multi-element system. *Multivariate analysis 2*, New York, Academic Press, 1969.
2. Cheok, M. C., Parry G. W., Sherry R. R., Use of importance measures in risk informed applications, *Reliab. Eng. and Sys. Safety* 1998; 60; 213-226
3. Borgonovo E., Apostolakis G. E., A new importance measure for risk-informed decision making. *Reliab. Eng. and Sys. Safety*, 2001; 72; 193-212
4. J. S. Hong and C. H. Lie, Joint reliability-importance of two edges in an undirected network, *IEEE Trans on Reliab* 1993 vol 42 no 1 pp 17-23
5. M. J. Armstrong, Joint reliability-importance of elements *IEEE Trans on Reliab* 1995 vol 44 no 3 pp 408-412
6. Vesely W.E., Kurth R. E. and Scalzo S. M., Evaluations of core melt frequency effects due to component aging and maintenance. *Report NUREG/CR-5510*, US Nuclear Regulatory Commission, Washington, DC, June 1990

A Response-Modeling Alternative to Surrogate Models for Support in Computational Analyses

Brian Rutherford

Sandia National Laboratories, P.O. Box 5800, MS 0829, Albuquerque, NM 87185, USA
E-mail: bmruthe@sandia.gov

Abstract: Often, the objectives in a computational analysis involve characterization of system performance based on some function of the computer response. In general, this characterization includes (at least) an estimate or prediction for the performance measure and an estimate of the associated uncertainty. Surrogate models can be used to approximate the response in regions where simulations were not performed. Most surrogate modeling approaches, however, are based on smoothing and uncertainty in the response is typically specified in a point-wise (in the input space) fashion. Together these aspects of the surrogate model construction might limit their capabilities.

One alternative is to construct a probability measure, $G(\mathbf{r})$ for the computer response, $\mathbf{r}$, based on available data. This "response-modeling" approach will permit probability estimation for an arbitrary event, $E(\mathbf{r})$, based on the computer response. In this general setting: $prob(E)=\int_r I(E(\mathbf{r}))dG(\mathbf{r})$ where I is an indicator function. Furthermore, one can use $G(\mathbf{r})$ to calculate an induced distribution on the performance measure, pm. For prediction problems where the performance measure is a scalar, the performance measure distribution F_{pm} is determined by: $F_{pm}(z)=\int_r I(pm(\mathbf{r})\le z)dG(\mathbf{r})$. We introduce response models for scalar computer output and then generalize the approach to more complicated responses that utilize multiple response models.

Keywords: computational simulation, experimental design, meta-model, prediction, reliability, response-modeling, response surface, surrogate models.

1. INTRODUCTION

Enhanced software methodology and improved computing hardware have advanced the state of simulation technology to a point where large physics-based codes can be a major contributor in many systems analyses. This shift toward the use of computational methods has brought with it new research challenges in a number of areas including model validation, (model-based) prediction and characterization of input, modeling and predictive uncertainty. It is these challenges that have motivated the work described in this paper.

The problem considered here is one of characterizing system performance based on results of a computer model. It is assumed that the model is expensive to run and consequently only a limited number of evaluations can be performed. It is assumed, further, that the model has been validated and hence the model has been determined to provide adequate results for the present application. For simplicity of presentation, we assume the model produces a single response for a given set of inputs. Different responses depending on uncertain model

parameters (calibration parameters) can, however, be accommodated through the methods discussed here. The tough issues of model validation and calibration are not addressed here.

We consider a number of specific system performance measures for illustration -- others are possible. These performance measures are based on the model responses that may be scalars or may be vectors or functions. In the introduction to the response models, we work with a scalar computer response. A scalar response is modeled by one response model. In the remainder of the paper, we generalize the approach using a functional response over time. The later applications illustrate methodology for using multiple response models to accommodate more general computer model outputs.

The response model described here is an atomic probability measure for the response calculated using a limited number of computational results. We use this measure to approximate probabilities associated with arbitrary events that are based on the response, although the major objective is to obtain a probability distribution function for the performance measure. We provide details on how the response models can be constructed and how they can be used in fairly general applications.

In this introductory section, we: (1.1) review some of the possible objectives in computational modeling; (1.2) discuss how surrogate models can assist in addressing these objectives, review one specific parametric form for surrogate models and mention some possible limitations in their use; and (1.3) discuss an alternative response-modeling strategy that overcomes some of these limitations. In the remainder of this paper, Section 2 provides a more detailed account of the response-modeling approach. A simple example is introduced that illustrates the construction of the measure G(**r**) for a scalar computer response. Section 3 extends the approach to more general types of computer responses providing two examples where multiple response models are used to characterize a functional output in time.

1.1. Objectives in Computational Modeling

Often, the objectives in a computational analysis involve the characterization of system performance based on some function of the response, **r**. We consider applications where **r**(**x**) is the computer response depending on the *p*-dimensional input **x.** The inputs may or my not be modeled probabilistically (with distribution $F(\mathbf{x})$). For given **r**, some common performance measures may be computed as:

$pm(\mathbf{r}) = \mathbf{r}(\mathbf{x}^*)$ (simple performance prediction problem);

$pm(\mathbf{r}) = \int \mathbf{r}(\mathbf{x}) dF(\mathbf{x})$ (average performance prediction problem);

$pm(\mathbf{r}) = \min_{\mathbf{x}} \mathbf{r}(\mathbf{x})$ (worst-case performance prediction problem);

$pm(\mathbf{r}) = \{\mathbf{x}^* : \mathbf{r}(\mathbf{x}^*) = \min_{\mathbf{x}}(\mathbf{r}(\mathbf{x}))\}$ (engineering design or optimization problem); or

$pm(\mathbf{r}) = 1 - \int I(\mathbf{r}(\mathbf{x}) \varepsilon R^*) dF(\mathbf{x})$ (reliability prediction problem)

where, R^*, the failure region, is some subset of the response space and I is an indicator function taking on the value 1 when the enclosed expression is true. In general, we desire both a prediction for the performance measure and an estimate of prediction uncertainty.

1.2. Surrogate Modeling for Computational Analyses

1.2.1. Surrogate models in computational analysis

Surrogate models can perform a number of functions in support of a computational analysis. Through interpolation extrapolation and/or integration, these models can be used to address complex problems involving experimental design, system analysis and prediction.

1.2.2. A common surrogate model form

One model that is commonly used as a surrogate response is the Gaussian Process model $\eta(\mathbf{x})$. A general form of the model is specified through its mean function $m(\mathbf{x})$ and spatial covariance function $\sigma^2\mathbf{C}(\mathbf{x},\mathbf{x}')$ for input vectors $\mathbf{x}$ and $\mathbf{x}'$. This model has been used successfully in numerous engineering applications, see [1] – [3] for typical assumptions and restrictions on the forms of m and $\mathbf{C}$.

1.2.3. Possible limitations to surrogate modeling

Depending on the performance measure, the smoothing involved in surrogate model construction and method of uncertainty characterization of the surrogate model might lead to difficulties. Consider the worst-case performance measure where $pm(\mathbf{r}) = \min_{\mathbf{x}} \mathbf{r}(\mathbf{x})$. The performance measure estimate based in the surrogate model will not give an accurate estimate of the minimum response value. Even an expected value for the worst-case response would be difficult to approximate using the information retained. Figure 1(a) illustrates the difficulty. Without information on the relationships of points in the lower regions of the curve, the probabilities are difficult to determine. Similarly, for the reliability performance measure, $pm(\mathbf{r}) = 1 - \int I(\mathbf{r}(\mathbf{x}) \varepsilon R^*) dF(\mathbf{x})$, while the surrogate model estimate might provide a reasonable estimate of reliability, the needed information is not retained in the surrogate model construct to accurately quantify uncertainty in this estimate.

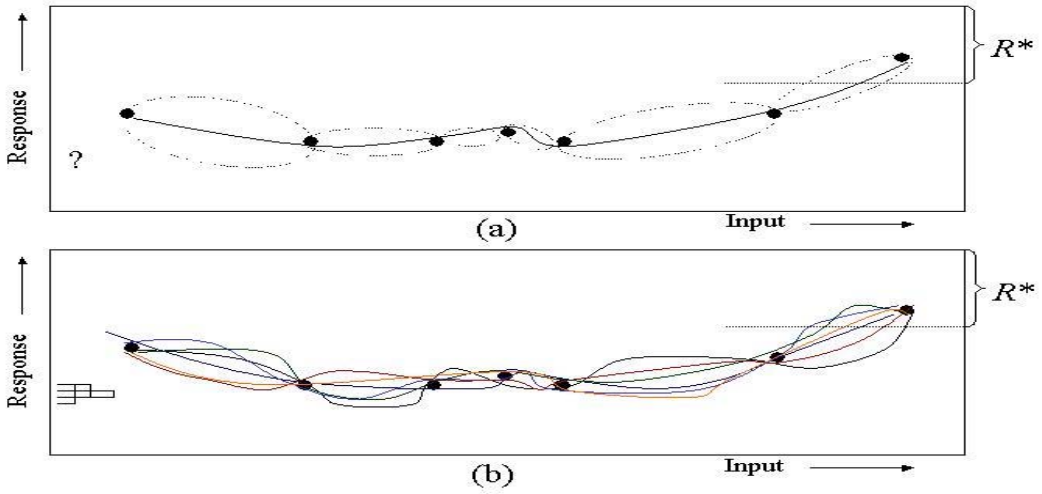


Figure 1. Hypothetical models for the response in a 1-input problem: (a) a surrogate model estimate with point-wise uncertainty bounds; and (b) a response-model with individual "realizations" forming an atomic measure over the response space. The histogram on the left indicates how a distribution might be constructed for a worst-case performance measure.

1.3. An Alternative Model for the Response

One alternative characterization for the system response is illustrated in Figure 1b and discussed in detail in Section 2 below. This "response-modeling" approach can avoid the problems discussed in the previous subsection. It consists of constructing an atomic measure over the response space that is based on assumptions concerning the appropriate model form and on the available computer response data. We refer to elements of the measure as "realizations" and assign them equal probability.

Once this measure has been established for the response space, we can approximate the probability associated with any event $E(\mathbf{r})$ through $prob(E)=\int_r I\left(E\left(\mathbf{r}\right)\right)dG\left(\mathbf{r}\right)$ and for the performance measure distribution through $F_{pm}\left(z\right)=\int_r I\left(pm(\mathbf{r})\le z\right)dG\left(\mathbf{r}\right)$. In Figure 1b we illustrate how this measure might be used on that 1-input problem. The histogram on the left of the figure provides an approximation to the density function for the worst-case performance measure -- a value that was difficult to estimate using the surrogate model. Similarly, we could demonstrate the uncertainty related to the reliability prediction by drawing a histogram of the reliability values computed using the individual realizations.

2. RESPONSE MODELING

2.1. Modeling Objectives

We construct a response-model as a discrete ensemble of realizations that could be interpreted as "probable" descriptions of the computer response as a function of the computer inputs. The realizations are constructed in the spirit of a Latin Hypercube sample [4] where they are generated to span the uncertainty range of the response while attempting to satisfy the consistency property stated in the next paragraph. The ensemble is used to approximate an uncertainty distribution for the fixed but unknown true response surface that captures the uncertainty in the response resulting from the knowledge being based on a limited number, n system evaluations. Formally, the response ensemble consists of a set of k realizations: $\mathbf{R}=\left(r^i(\mathbf{x}), i=1,...,k,\forall\mathbf{x}\right)$; G(**r**) assigns a probability $1/k$ to each realization $r^i(\mathbf{x})$. Using this formulation, the expressions above become: $prob(E)=1/k\sum_i I\left(E\left(r^i\right)\right)$ and $F_{pm}\left(z\right)=1/k\sum_i I\left(pm(r^i)\le z\right).$

Ideally, **R** is constructed in a manner consistent with the data $\mathbf{y}=y(\mathbf{x}_i)$, i=1,…,$n$ in the following sense: for any given event E based on the response, if the conditional probability $P\left(E\mid\mathbf{y}\right)=p$ then the expected number of realizations $r^i(\mathbf{x})$ satisfying E, (the expectation taken over repeated application of the response-modeling process) is kp.

The assertion $P\left(E\mid\mathbf{y}\right)=p$ requires assumptions for mathematical formalization. The response model is based on these assumptions addressing functional form and appropriate methods of construction. Decisions regarding the modeling assumptions are necessarily somewhat arbitrary. Some of the issues are addressed in [5] – [6], others are the topic of our current research. Some of the possible inaccuracies resulting from the assumptions tend to cancel each other out when making relative evaluations like comparing experimental design alternatives (our primary application of the response models). Figure 2 shows an example

response that is used throughout this section. The response is an analytical function that is simple to evaluate, but is used here, for illustration, in place of the expensive computational simulation model. A twenty-five-realization response model is constructed based on fifteen functional evaluations as indicated by the stars in Figure 2b. In the next subsection, we provide the details for constructing the response model.

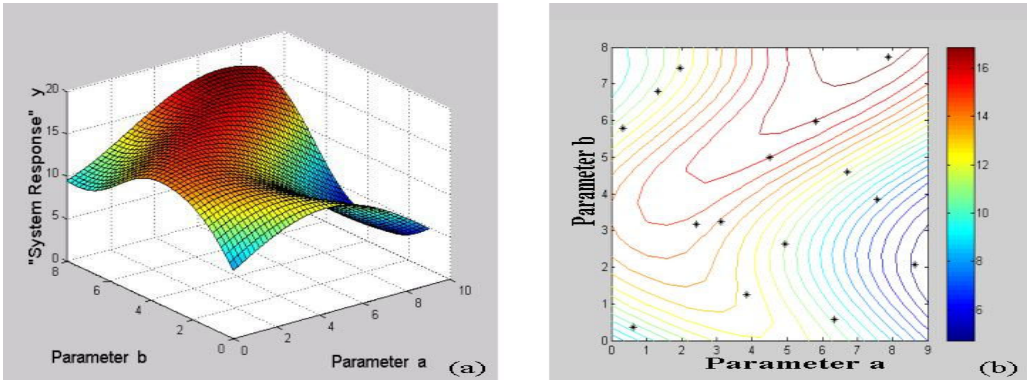


Figure 2. Analytical example response surface and contours based on two inputs.

2.2. Response Model Construction

Listed next are the steps used in construction of the realizations for the examples given in this paper. We utilize the Gaussian process model referenced in Subsection 1.2.2. We consider m to be a low order polynomial in $\mathbf{x}$ and restrict $\mathbf{C}$ to be of the form:

$$\mathbf{C}(\mathbf{x},\mathbf{x}') = \prod_{i=1}^{p} C\left(\left|x_i - x_i'\right|\right) = \prod_{i=1}^{p} e^{-\phi_i \left\|\mathbf{x}_i - \mathbf{x}_i'\right\|} \text{ for any } \mathbf{x} \text{ and } \mathbf{x}'$$

where $\left\|\cdot\right\|$ is the Euclidian norm and the ϕ_i are estimated from the data. In our case, $\mathbf{C}$ is the covariance for the residual response after fitting the low order polynomial as described in more detail below. It is more convenient for our application to decompose the Gaussian process model and use the form $\eta(\mathbf{x}) = P(\mathbf{x}) + \varepsilon(\mathbf{x})$ where $P(\mathbf{x})$ is a polynomial in $\mathbf{x}$ and $\varepsilon(\mathbf{x})$ is a zero-mean Gaussian process model. Steps (1) through (3) below describe construction of the polynomial component; steps (4), (5) and (6) describe construction of the Gaussian process term; step (7) combines the two elements.

1) Evaluate main effects, quadratic terms, and interactions, where possible, using the initial data, settling on an appropriate polynomial regression model.
2) Estimate the regression coefficients and their covariance structure.
3) Generate k sets of coefficients (assuming a multivariate normal for their joint distribution) using a Latin Hypercube design with the appropriate correlation structure imposed on the sets of coefficients using rank correlation procedures described in [7]. The remaining 4 steps are applied to each realization.
4) The residuals to the regression surface are transformed using the "Normal-scores transform" as recommended in [8].
5) The transformed residuals can then be used to estimate parameters of the spatial covariance function given above. We used a maximum likelihood procedure to

estimate the parameters for the combined set of residual (spatial difference) data from all realizations. Differences in residual magnitude and distribution are maintained through the transformation and back-transformation in 4) and 6).

6) The "sequential-Gaussian" conditional simulation procedure is used to generate the random function component $\varepsilon(\mathbf{x})$. More detail (of a mechanical nature) on the sequential Gaussian approach is given in [8] Chapter V. The algorithm generates a response surface over the grid in transformed space and then back-transforms the values according to a set of tables constructed during the transformation in 4). The conditioning data are the transformed residuals to the polynomial surface

7) The back-transformed random function term is added to the polynomial surface to complete the realization.

We illustrate some of these steps using the analytical example. Figure 3 shows the construction for two realizations. The regression component is in Figure 3a, the random function component in 3b and the completed realization in 3c. The pair of realizations illustrate possible differences within the ensemble. Figure 4 provides further illustration with three additional completed realizations for this example.

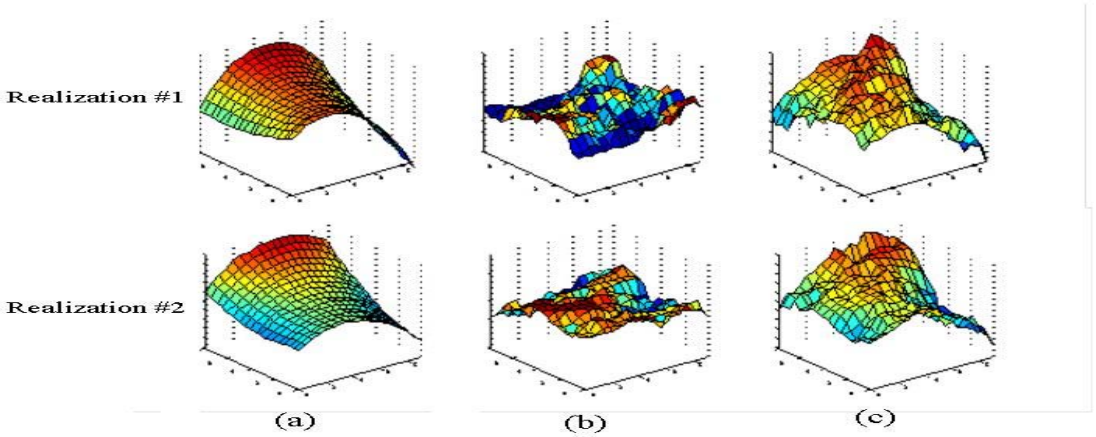


Figure 3. Components of the response realizations for the analytical example.

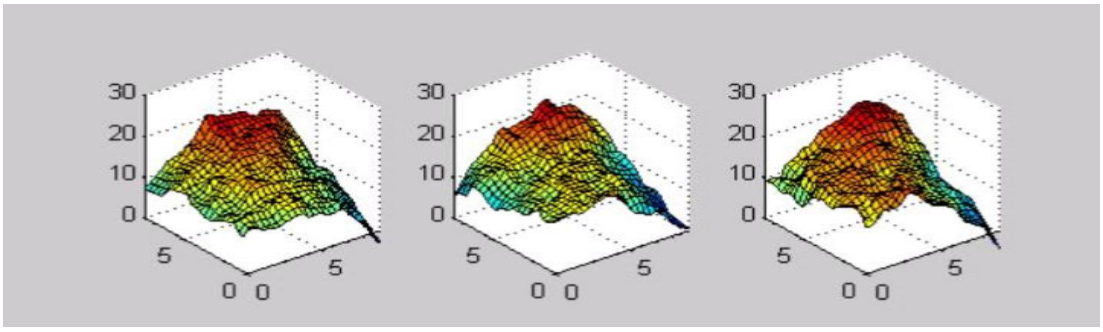


Figure 4. Three of twenty-five realizations constructed for the analytical example.

2.3. Response Model Application

Once the response model has been constructed, it can be used as described earlier to evaluate arbitrary events based on the response. The events of primary interest here relate to system performance. We demonstrate how probabilities for three different performance

measures can be assessed using the response model generated for the analytical example. The procedure for the first and last performance measure is to apply the formulas given early using the equally weighted realizations for $G(\mathbf{r})$. We illustrate the performance measure distribution here by showing the corresponding histograms.

Consider, first, the "best-case" performance measure where we will assume high values of the response are "good". Figure 5a shows a histogram of realization results for this quantity – from this we can easily approximate its distribution. In Figure 5b, the locations in the input space where these maximum values occurred are plotted. This plot addresses the optimization problem where we are interested in the input location yielding the maximum value. The next step might be to quantify in some way (some measure based on clustering metrics, for example) the spread of probable input locations. The maximum values in this example are confined to the discrete grid used to record values for $G(\mathbf{r})$. The final performance measure considered here is reliability where we assume the failure region R^* is that part of the response space exceeding 18. We need distributions concerning the inputs a and b to determine a probability of failure. For simplicity, we arbitrarily choose to assume both input parameters are uniformly distributed over their ranges. Figure 5c provides a histogram for reliability under these assumptions.

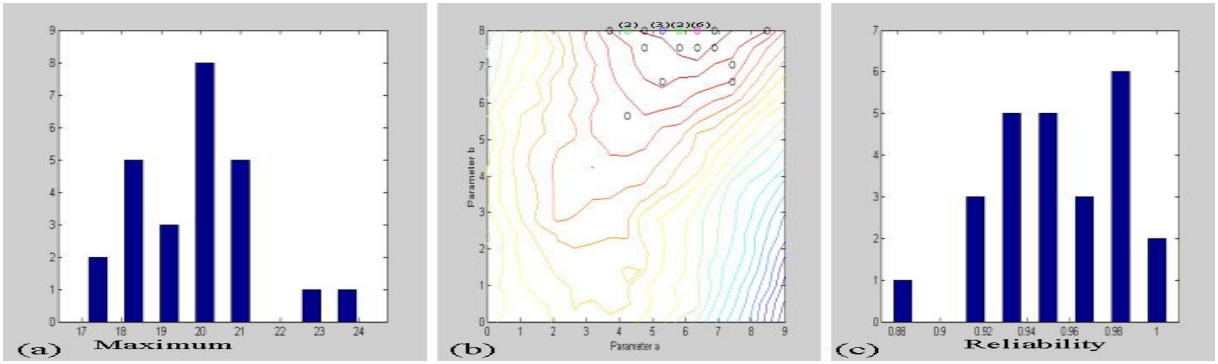


Figure 5. Performance measures for the analytical example.

3. EXTENSIONS TO FUNCTIONAL COMPUTER RESPONSES

Computer responses may be scalars or vectors but can also be functions of time and/or space. The examples that follow illustrate how these more complicated responses can be addressed using the response modeling approach. We refer to the relatively complicated empirical models used to create the probability measures for these responses as "behavioral models." They combine multiple components, including the response models and a "response assembly" model. For the examples considered here, the response assembly model creates a pulse over time based on a discrete set of "intermediate" parameter values that are modeled using one response model each. The Device #1 example illustrates a case where the response can be modeled adequately through a four-parameter circuit model. In this example the response models are used to approximate probability measures for these four parameters. The Device #2 example gives a case where there is no circuit model that showed the flexibility needed to accommodate the range of pulses in the data set. In this case, basis functions were established using principle components analysis and the response models were used to approximate probability measures for the basis function coefficients. Before

proceeding to the examples we outline briefly the process of constructing the behavioral models.

1. Determine a functional form that has the flexibility to accommodate the range of responses anticipated in the application (where possible). The parameters needed to fully specify a response are referred to as intermediate parameters (assume q of them).
2. Determine the "best fit" values of the intermediate parameters for the computer generated conditional response data. This setup will result in a set of p dimensional input -- q dimensional response (intermediate parameter) data.
3. Use the data above to construct q response models. The response models should be p-dimensional unless an analysis of the data indicates that some of the inputs are not important for some of the responses.

The behavioral models can now be used to make predictions. Any specific set of inputs will yield k intermediate parameter values from each of the q response models. These k^q sets of intermediate parameters can be used to generate a distribution of the response corresponding to the specified inputs. If the performance measure is constructed from the sets of responses, k^q values are available to approximate a distribution.

3.1. Device #1 Example

This example provides an illustration of how response modeling can be used for more complex responses and performance criteria. Figure 6 illustrates the behavioral model used in the analysis. The output pulses were modeled using a capacitance discharge unit (CDU) mapping voltage (V_0), inductance (I), capacitance (C) and resistance (R) to the current pulse. Specification of the performance measure, for this example depends on those aspects of the pulse considered critical for performance. Maximum current, for example, is one possible quantity of interest.

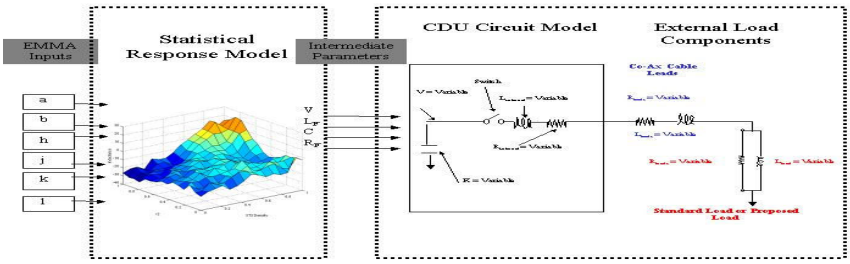


Figure 6. Behavioral model for the Device #1 example.

Following the outline specified above, four response models were constructed for the intermediate parameters (the four electrical parameters). Two objectives of this analysis were to be able to make predictions for arbitrary sets of inputs and to investigate performance throughout the 6-dimensional input space economically. Prediction uncertainty (expressed through 88% bounds) and computer generated values (the circles) for one of the pulses that was not used in constructing the response models are shown in Figure 7a. The confidence bounds apply "point-wise" for individual time values. Six of the twenty-seven points in three pulses used are outside the 88% confidence bounds -- almost twice the number that should fall outside

for a typical analysis on average. This result is not unlikely given the high correlation among points and the small sample size of these curves. It is possible, however, that the uncertainty is understated because of assumptions concerning the behavioral model form. These assumptions are difficult to evaluate and this source of uncertainty is not included in computation of the bounds.

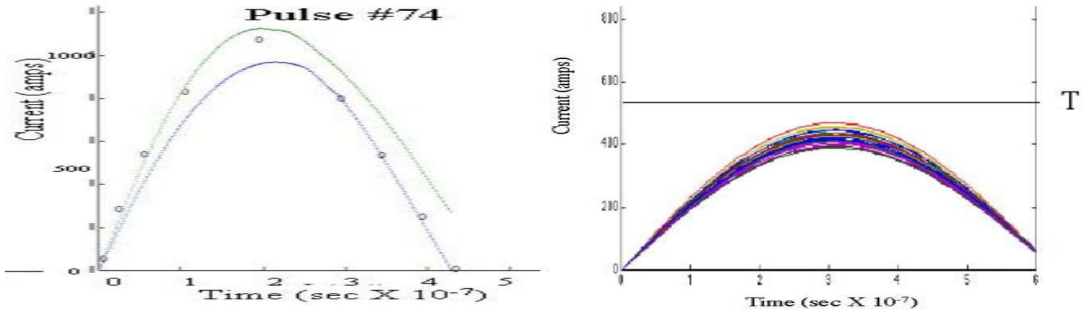


Figure 7. Modeling uncertainty for fixed inputs for the Device #1 example. Figure 7a shows an actual response (the circles) and 88% prediction bounds. Figure 7b shows the pulse yielding the lowest peak current and its related uncertainty.

To illustrate a possible scenario addressing the second objective, consider a hypothetical threshold established at $T = 550$ amps and assume we are concerned that pulses that do not achieve this threshold may indicate unacceptable reliability. The behavioral model was used to investigate the entire 6-dimensional input space in several hours. Figure 7b shows the ensemble of output pulses (indicating the prediction and uncertainty associated with the prediction) corresponding to the inputs yielding the worst performance according to this criteria.

3.2. Device #2 Example

In the Device #1 example, the current pulses were of shapes that could be accommodated using a circuit model. In this example, no simple formulation, flexible enough to characterize all current pulses, was available. We include this example to illustrate how a set of basis functions can be constructed to provided the flexibility for modeling arbitrary curves or surfaces using a response-modeling approach.

The initial data consisted of 136 runs of the computer model. The current pulses for these runs are shown in Figure 8a. The pulses were "discretized" (to 31 points along the time axis) and were shifted, "time registered", in a way that minimized their squared differences compared with the average pulse (see Figure 8b). The average pulse was then subtracted leaving the residual curves in Figure 8c.

The resulting sets of discrete values were evaluated through principle component analysis see [9] Chapters 5 and 6 for a complete description of these methods. Basis functions, using the principle components, were constructed as described in that text. Figure 9 illustrates the Device #2 behavioral model. Given values for the six inputs, the eight response models each generate 20 values that span the range of coefficients for the appropriate basis function or the time-registration parameter. Performance assessment for this example could be addressed in a similar way to that illustrated for the Device #1 example.

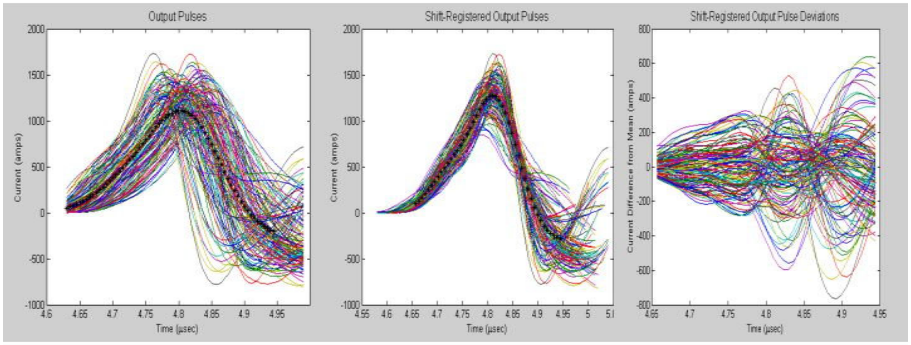


Figure 8. Current pulse responses for the Device #2 example.

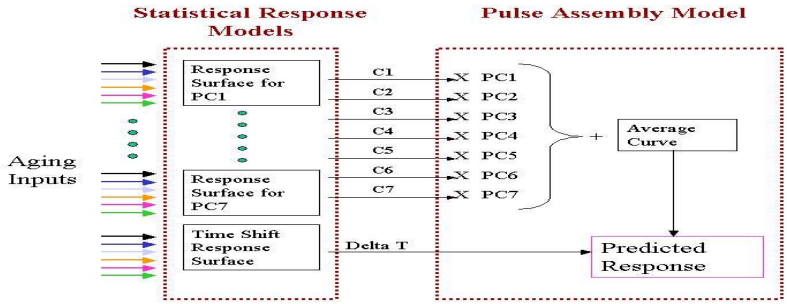


Figure 9. Behavioral model for the Device #2 example.

REFERENCES

1) Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. Design and Analysis of Computer Experiments, *Statistical Science*, Vol. 4, No. 4, 409-423, 1989.

2) O'Connell, M. A., and Wolfinger, R. D. Spatial Regression Models, Response Surfaces, and Process Optimization, *Journal of Computational and Graphical Statistics*, Vol. 6, No. 2, 224-241, 1997.

3) Kennedy, M. C., and O'Hagan, A. Bayesian Calibration of Computer Models (with discussion). *Journal of the Royal Statistical Society (Series B)*, 58, 425-464, 2002.

4) McKay, M. D., Beckman, R. J. and Conover, W. J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, *Technometrics*, Vol. 21, No. 2, 239-245 1979.

5) Rutherford, B. *An Algorithmic Approach to Experimental Design for the Computer Analysis of a Complex System*, SAND2000-0871, Sandia National Laboratories, Albuquerque, New Mexico, 2000.

6) Rutherford, B. *A Response-Modeling Approach to Uncertainty Characterization and Computer Experimental Design*, Whitepaper, Sandia National Laboratories, Albuquerque, New Mexico, 2003.

7) Iman, R. L. and Conover, W. J. A Distribution-Free Approach to Introducing Rank Correlation Among Input Variables, *Communications in Statistics, Part B - Simulation & Computation*, 11(3), 311-334 1982.

8) Deutsch, C. V. and Journel, A. G. *GSLIB Geostatistical Software Library and Users Guide Second Edition*, Oxford University Press, New York, 1998.

9) Ramsay, J., O. and Silverman, B., W. *Functional Data Analysis* Springer-Verlag, New York, 1997.

Local Analysis of Prameter Covariances Resulting from the Calibration of an Overparameterized Water Quality Model

U. Callies and M. Scharfe

GKSS Research Center, Max-Planck-Str. 1, 21502 Geesthacht, Germany
E-mail: ulrich.callies@gkss.de

Abstract: Mechanistic water quality simulation models are important tools for supporting environmental management decisions. Possibly the most severe problem with the usage of mechanistic models is that in most cases they cannot be fully identified from data due to model overparameterization. The calibration of overparameterized models results in covariances among model parameters, the neglect of which may lead to a significant overestimation of model output uncertainty. We discuss principal component analysis (PCA) of the posterior parameter error covariance matrix as a tool for the identification and proper representation of parameter covariances. Our study deals with a water quality model specifically designed to support the interpretation of algae biomass observations at one single station (Weir Geesthacht) on the Elbe river in Germany.

Keywords: Local sensitivity analysis; Hessian matrix; Principal component analysis; Model overparameterization

1. INTRODUCTION

The motivation for our modelling activity has been to test a hypothesis according to which observed negative correlations between temperature and chlorophyll *a* concentrations in summer at station Geesthacht on the Elbe River might indicate algae growth being limited by lack of silica. For this specific purpose a relatively simple model has been designed and fitted to observed chlorophyll *a* concentrations. However, even this simple model turns out not to be identifiable from the data used. One obvious reason why some model parameters or parameter combinations are poorly determined by the data, is that the observations have all been collected at the same location and do therefore not resolve profiles along the river. A possibly more important explanation, however, is that observations of one single state variable cannot disentangle details of the mechanistic processes. Different parameterizations linking external forcing (radiation, discharge, temperature) to model output (algal biomass concentrations) can be similarly effective in reproducing the data.

Being not parsimonious in the light of the existing data (i.e. being overparameterized) is a property of most detailed mechanistic models. Our study is intended to illustrate an approach for coping with this situation by explicitly accounting for parameter interaction structures that result from model overparameterization. Specific combinations of parameters may be much less uncertain than the individual parameters they are made up by. The opposite is also true: Some parameters may be collectively more uncertain than any of the individual parameters. PCA of the posterior parameter error covariance matrix allows to discriminate combinations of model parameters that are effectively controlled by the data from those parameter combinations that are irrelevant for model counterparts of observations. Often results from PCA can be interpreted in terms of the basic mechanisms represented in the model. We illustrate the method for the example from water quality modelling.

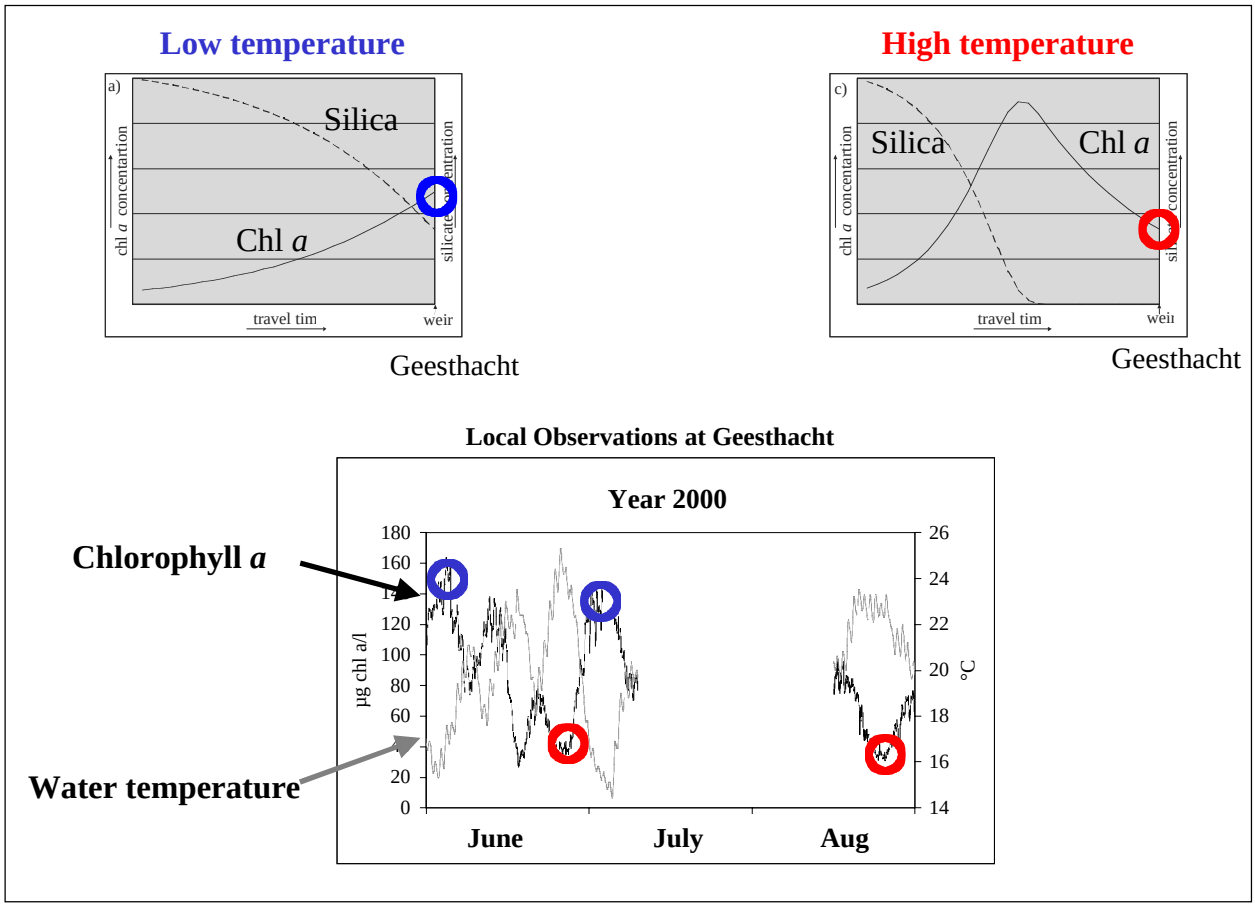


Figure 1: Suggested explanation for negative correlations between water temperature and chlorophyll *a* concentrations observed during certain summer periods at station Geesthacht. Upper panels: Modelled temporal evolution of chlorophyll *a* and silica concentrations in individual water packages assuming low and high temperature, respectively. Local time series can be produced by storing the final points of all trajectories that arrive at Geesthacht. Bottom panel: Observations at Weir Geesthacht. Temperature observations have been shifted by an estimated response time of two days.

2. METHODOLOGY

2.1. The Model

Our model involves the representation of two different algae species, green algae and diatoms, both of which are known to significantly contribute to the total amount of algal biomass in the river Elbe. Only growth of diatoms, however, depends on the availability of the nutrient silica. The general concept is to consider a series of individual water bodies travelling downstream towards station Weir Geesthacht. Starting from unrealistically low values, meaningful concentrations of algal biomass to be compared with observations evolve only by the end of the water parcel's journey after exponential growth over a limited period of

travel time (cf. Figure 1). This travel time is parameterized empirically as a function of river discharge.

At the beginning of its journey each water package is initialised by the same concentration of silica. Diatoms are assumed to cease growing and to start decaying as soon as this initial reservoir of silica has been used up (cf. upper right panel of Figure 1). The higher growth rates are the earlier the diatom maximum occurs and the more pronounced it is. If growth rates are large enough so that all available silica is assimilated already upstream of Geesthacht, further increasing growth rates (i.e. more favourable growth conditions) imply decreasing diatom populations at the end of the particle's journey. Thus, assuming that diatom growth rates increase with temperature this provides a plausible explanation for negative correlations between water temperature and chlorophyll *a* concentrations at station Geesthacht. The assumed relationship between Lagrangian trajectories of individual water packages and local observations at station Geesthacht is outlined in Figure 1.

For each of the two species an individual balance equation describes changes of algal biomass concentration, C_{alg}, with time, t, as a function of temperature dependent rates of algal growth, μ, respiration, ρ, and loss, σ:

$$\frac{d\,C_{\mathrm{alg}}}{d\,t} = \left[\mu(T) - \rho(T) - \sigma(T)\right] C_{\mathrm{alg}} \quad \text{with} \quad \mu(T) = \mu^{0}_{\max}\, q_1(T)\, F_{\mathrm{light}}\, F_{\mathrm{Si}} \tag{1}$$

The focus of the sensitivity and uncertainty study reported below is on six model parameters $x_1, \ldots, x_6$ that enter parameterizations of the algae growth rate, μ. Two parameters are the optimum growth rates $\mu^{0}_{\max}$ for green algae and diatoms, respectively, that occur when algae growth is not limited by lack of light or nutrients. Next we consider for each algae species the parameter k_{light} that specifies the light intensity, at which the algae growth rate reaches 71% of its maximum possible value. The light limitation factor F_{light} in Eq. (1) results from vertically averaging over the efficient water depth, D, which is another model input parameter that has been selected for our study:

$$F_{\mathrm{light}} = \frac{1}{D}\int_0^D \frac{I\, e^{-\lambda z}}{\sqrt{k^2_{\mathrm{light}} + I^2 e^{-2\lambda z}}}\, dz \tag{2}$$

Here I denotes the radiation intensity at the water surface and λ the total light attenuation coefficient due to the presence of mineral compounds and algal self shading.

The last parameter we chose is the fraction of silica, f_{Si}, in the cells of diatoms that governs the consumption rate of silica (with concentration C_{Si}) by the diatoms with concentration $C_{\mathrm{alg,d}}$ and growth rate μ_{d}:

$$\frac{d\,C_{\mathrm{Si}}}{d\,t} = -\, f_{\mathrm{Si}}\, \mu_{\mathrm{d}}\, C_{\mathrm{alg,d}} \tag{3}$$

The concentration of silica determines the size of the *Michaelis-Menten* type growth limitation factor, F_{Si}, in Eq. (1) with a half saturation constant K_{S} (relevant only for diatoms):

$$F_{\mathrm{Si}} = \frac{C_{\mathrm{Si}}}{K_{\mathrm{S}} + C_{\mathrm{Si}}} \tag{4}$$

All model parameters have been tuned manually to obtain reasonable simulations of chlorophyll *a* and silica concentrations. The resulting reference parameter values are shown in Table 1 together with rough estimates of prior parameter uncertainties.

To run the model external forcing represented by water temperature, discharge and radiation must be specified as a function of time. Hourly temperature observations are part of the data set from Weir Geesthacht. Daily observations of discharge at station Neu-Darchau (Elbe-km 536) are available from ARGE ELBE (http://www.arge-elbe.de). Hourly mean values of global radiation were provided by the GKSS Research Centre Geesthacht which is located few kilometres upstream of the weir.

Table 1: Reference parameter values and assumed uncertainties; Optimised values for 2000

	Name	Ref. Value	Range	StD	Value 2000	Unit
x_1	$k_{\text{light, g}}$	20	[10-30]	5.77	17.7	W/m^2
x_2	$k_{\text{light, d}}$	14	[10-18]	2.31	14	W/m^2
x_3	$\mu^0_{\text{max, g}}$	1.65	[1.3-2.0]	$2.02\ 10^{-1}$	1.63	1/d
x_4	$\mu^0_{\text{max, d}}$	1.6	[1.2-2.0]	$2.31\ 10^{-1}$	1.62	1/d
x_5	f_{Si}	0.2	[0.1-0.3]	$5.77\ 10^{-2}$	0.2	mgSi /mgC
x_6	D	2.35	[1.2-3.5]	$6.64\ 10^{-1}$	2.08	m

A comparison of observations and corresponding model results is presented in Figure 2. Considering the very simple model approach with water temperature, discharge and solar radiation being the only time dependent model inputs (no time dependent initial values or sources/sinks) the model reproduces a reasonable amount of observed variability. Note that also the observed very low levels of silica during summer are reflected by the model calculations.

For the sensitivity and uncertainty study reported below we assume that only chlorophyll *a* data were available for model calibration. Silica data are considered as an independent option for checking the appropriateness of the model mechanism.

2.2. Sensitivity and Uncertainty Analysis

Let $\vec{x}$ denote the vector being made up by those model input parameters, x_i, which are to be adjusted by fitting the model to data. If $\vec{m}$ denotes the vector of model outputs, m_t, at times t, a quadratic loss function, J_{obs}, may be used to assess the differences between model output and observations, d_t, scaled by an assumed observational error, σ_{obs}. In the case of model overparameterization the minimum of the loss function will be not well-defined due to small curvatures in certain directions. To make the optimisation problem well-posed we complement the loss function by a second component that penalizes all deviations of the actual parameter vector $\vec{x}$ from the reference vector, $\vec{x}_0$, which is specified in Table 1:

$$J = J_{\text{obs}} + J_{\text{prior}} = \frac{1}{2}\sum_{t=1}^{N}\frac{\left[m_t(\vec{x}) - d_t\right]^2}{\sigma_{\text{obs}}^2} \quad + \quad \alpha\frac{1}{2}\left(\vec{x} - \vec{x}_0\right)^{\mathrm{T}}\mathbf{V}_{\text{prior}}^{-1}\left(\vec{x} - \vec{x}_0\right) \tag{5}$$

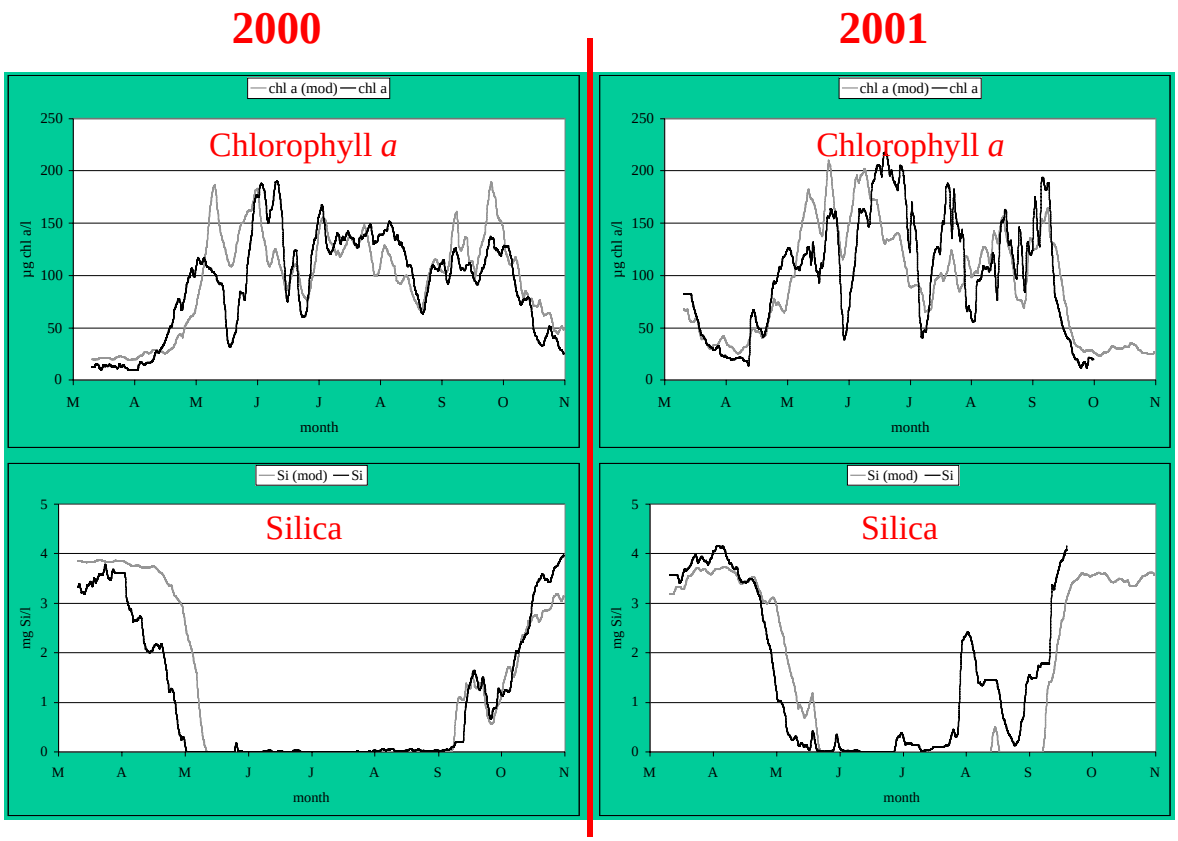


Figure 2: Observations (black) and corresponding model predictions for chlorophyll *a* (upper panels) and silica (lower panels) at station Geesthacht for two different years.

The diagonal prior parameter covariance matrix, $\mathbf{V}_{\text{prior}}$, is introduced to remove differing physical dimensions of the six parameters by proper scaling. For the present study we decided to measure changes of parameter values in terms of multiples of their estimated prior uncertainty (cf. Table 1). The scalar α enables one to adjust the overall sizes of the data and the penalty term relative to each other. Note that the size of the data term and therefore the appropriate choice of α must be a function of the number of observations. In the following we choose α=100 for N=78 data points comprising about one observation every third day between March and October 2000.

The relevance of parameter covariances for a moderately non-linear model's fit to data can be analysed by examining the curvature of the loss function at its minimum. Figure 3 illustrates the general idea. Directions of high curvature allow for only weak parameter variations without getting into conflict with the data, whereas in directions with low curvature parameters can be changed significantly without much affecting the value of the loss function. A complete description of the loss function's local curvature is provided by the Hessian matrix containing all second derivatives of the loss function. Principal component analysis (PCA) of the Hessian matrix allows to discriminate those directions in the 6-dimensional parameter space that can be calibrated by the available data from other directions, along

which effects of parameter changes tend to compensate each other. For a linear model the inverse Hessian can be identified as being the posterior parameter error covariance matrix [3].

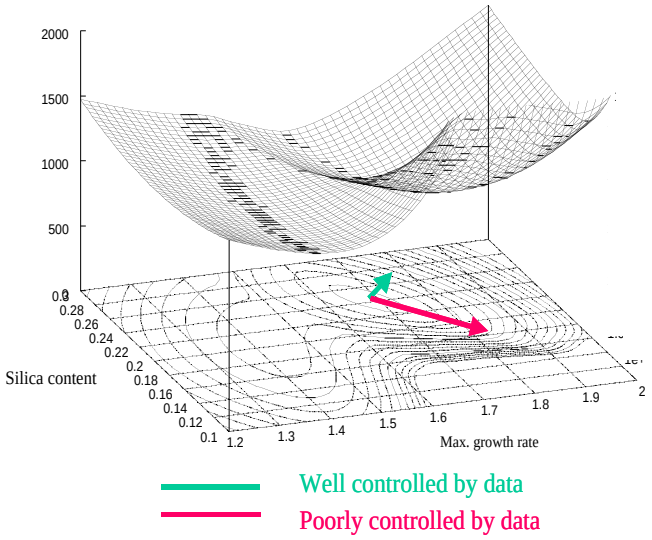


Figure 3: Loss function for variations of only two parameters; eigenvectors of the Hessian matrix defining directions of minimum and maximum curvature, respectively.

For the analysis reported below it is convenient to incorporate both the diagonal prior covariance matrix $\mathbf{V}_{\text{prior}}$ and the amplitude factor α into a coordinate transformation to non-dimensional parameters $\vec{x}'$:

$$\vec{x} = \left(\frac{1}{\alpha} \mathbf{V}_{\text{prior}} \right)^{1/2} \vec{x}' \tag{6}$$

For a linear model the loss function is quadratic. For a non-linear model the loss function should still be nearly quadratic in the vicinity of its minimum so that the following formula provides a good approximation of the Hessian and the inverse of the posterior parameter covariance matrix, $\mathbf{V}'_{\text{post}}$, [3]:

$$\left(\mathbf{V}'_{\text{post}}\right)^{-1} = \frac{\partial^2 J}{\partial \vec{x}'^2} = \frac{1}{\sigma^2_{\text{obs}}} \sum_{t=1}^{N} \left(\frac{\partial m_t}{\partial \vec{x}'} \right)^{\mathrm{T}} \left(\frac{\partial m_t}{\partial \vec{x}'} \right) + \mathbf{I} \tag{7}$$

A programming environment for the solution of control problems (Integrated Modelling and Analysis System (IMAS, [1]) including an automatic differentiation tool for computer programs has been used for an analytic calculation of the Jacobian matrix $\partial \vec{m} / \partial \vec{x}$ and for a gradient based minimization (Quasi-Newton algorithm) of the loss function. The inclusion of prior knowledge into Eq. (7) renders the inversion of the Hessian matrix possible. The relative

strength of background knowledge has been chosen sufficiently small to make the background term negligible for modes being well controlled by the data.

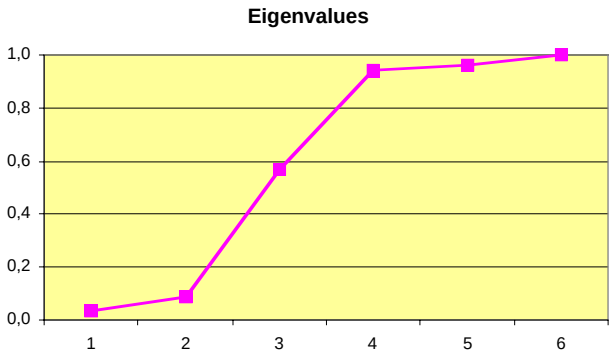


Figure 4: Eigenvalues of the posterior parameter covariance matrix analysed for year 2000.

3. RESULTS

The model has been fitted to data from the year 2000. The optimized parameter values listed in Table 1 do not deviate far from the reference values that were obtained as by careful manual model calibration. Figure 4 depicts the eigenvalues of the analysed posterior parameter error covariance matrix, which represent variances along directions in the parameter space being defined by the associated eigenvectors. According to Eq. (7) all eigenvalues of $\mathbf{V}'_{\text{post}}$ must be smaller than one as a consequence of the particular way of scaling. It turns out that only the first two eigenvalues are small enough to conclude that prior uncertainties have been overwritten by observational evidence. Three eigenvalues remain virtually unaffected by the data and stay close to the prior value one. One eigenvalue is influenced by both data and prior knowledge. Thus, PCA of the posterior parameter covariance matrix suggests that only two degrees of freedom in the six-dimensional parameter space are clearly relevant for a successful reproduction of chlorophyll *a* observations. It should be kept in mind, however, that these results are influenced by the way, in which parameters have been scaled. Scaling introduces some subjectivity but is a necessary prerequisite for PCA, which needs homogeneous physical dimensions [2].

A crucial question is whether or not the two parameter combinations, which are well controlled by the data, can be interpreted in terms of mechanistic processes represented in the model. Principal components are artificially defined variables and therefore do not necessarily have a physical interpretation. The two upper panels in Figure 5 depict the eigenvectors that correspond with the two smallest eigenvalues in Figure 4. To provide a better understanding of the mechanism of parameter calibration the bottom panel of Figure 5 presents an analysis of model prediction errors caused by uncertainties in the space of principal components. If the model is assumed to be linear, uncertainties of model parameters can be propagated independently and their effects on model output can be superimposed to each other.

It turns out that the two leading principal components that jointly explain about 89% of model output variance affect model predictions (i.e. are identifiable from the data) in distinct time intervals. This can be explained considering the different signs that diatom related

loading coefficients of the two eigenvectors have. According to both eigenvectors a strong (positive) impact on predicted chlorophyll *a* concentrations can be achieved by an intensification of the maximum growth rate of green algae, x_3, together with a reduction of their demand of light, x_1. However, according to the first eigenvector, which dominates model output uncertainty during periods when silica is lacking (cf. Figure 2), values of the corresponding parameters for diatoms, x_4 and x_2, are changed in the opposite directions. This reduces the overall diatom growth rate but nevertheless results in an increase of the final diatom biomass concentration (cf. upper right panel in Figure 1). If silica concentrations are sufficiently high, calibration of the second instead of the first eigenmode becomes crucial, which treats both algae species symmetrically.

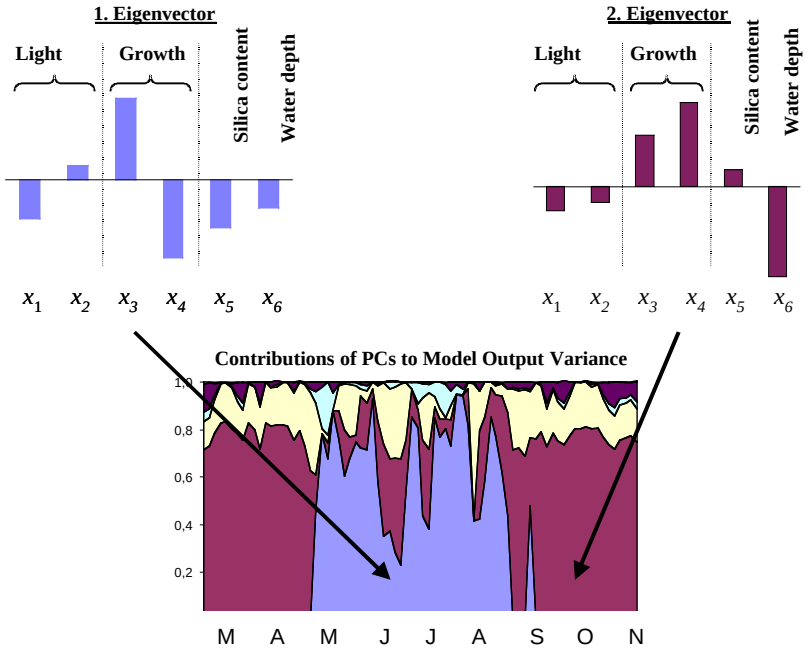


Figure 5: Cumulative plot of the relative contributions of the six principal components (amplitudes of the eigenvectors of the posterior parameter error covariance matrix) to model output variability. About 80% of model output variance can be attributed to the first two principal components.

The results of the local uncertainty analysis could be used for the implementation of a more systematic model calibration procedure. It suffices to calibrate the two or three leading principal components and to assign arbitrary values to the others. To illustrate the idea we performed a Monte Carlo experiment with 10,000 simulations. For each model run values of the six selected model parameters were drawn independently. Then those parameter combinations, which gave rise to the most successful model simulations (i.e. lowest values of the loss function), were projected onto the eigenvectors of the local parameter error covariance matrix. These projections, being standardized to have zero mean and unit standard deviation, are depicted in Figure 6. Projections onto the first three vectors are connected by solid lines. Projections onto the first two eigenvectors show small variability while anomalies

of the other projections may have large values even for the very best simulations. This proves that the choice of these amplitudes has little impact on the model's fit to data.

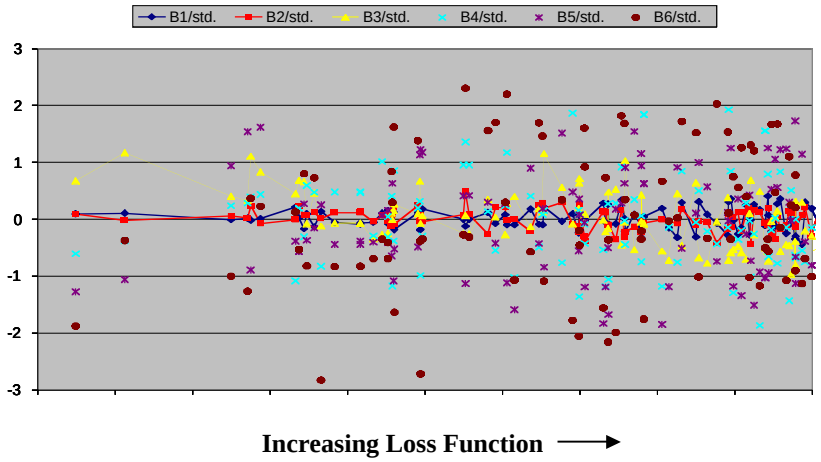


Figure 6: Monte Carlo generated parameter vectors giving rise to successful simulations are projected onto the six eigenvectors of the local parameter covariance matrix at the loss function's minimum. All projections are scaled by their respective standard deviations. Projections onto the first three eigenvectors are connected by solid lines.

4. DISCUSSION

Posterior parameter correlations reflect the fact that in the process of model calibration changing the values of different parameters had similar effects on model counterparts of the data. Taking into account such parameter interaction structures strongly mitigates uncertainties in model predictions [6]. PCA of the posterior parameter error covariance matrix gives a clear picture of how many degrees of freedom are really controlled by data. It helps to identify the nature of a model's overparameterization and its dependence on the kind of data being available for model calibration. The artificial new input parameters (principal components) do not necessarily have a physical meaning. In the present example, however, two parameter patterns being controlled by data could be related to the discrimination between different algae species in the model.

Vajda and Turányi [5] applied PCA for optimally reducing the mechanism of chemical reactions based, however, on a response function which measures model output variability but makes no reference to observed data. Using a small number of sensitive principal components as new independent model input parameters, thereby implicitly taking into account model parameter interactions, could much facilitate the adaptation of a model to new applications.

When a model is linear, uncertainties of model parameters can be propagated independently and their effects on model output can be superimposed. If a model is significantly non-linear, there are obvious limitations for a local sensitivity analysis, and a more general global method may be needed. As long as the local method is applicable, however, one can take advantage of the very basic definition of sensitivity in terms of the slope of model output at a given point in the parameter space [4]. Signs of model output

sensitivities are available rather than their pure sensitivity strengths, which facilitates an interpretation in mechanistic terms.

In the present application it turned out that the relevance of eigenmodes changed when the modelled concentration of silica dropped to zero. Accordingly any parameter change, which affects the modelled lengths of periods with lack of silica, will have an impact on the posterior parameter covariance matrix. This indicates the limitations of linear uncertainty analysis applied to the present example. Depending on a model concept with or without the inclusion of silica, model parameters must be calibrated differently, possibly giving rise to chlorophyll *a* simulations of similar quality. To consult observations of silica is the only way to resolve this ambiguity with regard to model formulation.

5. ACKNOWLEDGEMENTS

Parts of this work have been funded by the European Commission, IST Program, through contract IST-1999-11313 (IMPACT project). The authors thank Gerd Blöcker and Wilhelm Petersen for the provision of data and Friedhelm Schroeder for the provision of the basis version of the water quality model WAMPUM. The choice of the specific model approach was much influenced by ideas of Gerd Blöcker regarding the impact of silica ob algal biomass concentrations observed at Weir Geesthacht.

REFERENCES

1. A. Rhodin, U. Callies and D. P. Eppel. IMAS-Integrated Modeling and Analysis System, Release 1.0. Report GKSS 96/E/28, GKSS Research Centre, D-21494 Geesthacht, Germany, 1996.
2. A. Tarantola. *Inverse Problem Theory.* Elsevier, Amsterdam, 1987.
3. W. C. Thacker. The role of the Hessian matrix in fitting models to measurements. *J. Geophys. Res.,* 94, 6177-6196, 1989.
4. T. Turányi and H. Rabitz. Local Methods. In A. Saltelli, K. Chan and E. M. Scott, editors, *Sensitivity Analysis*, 81-99, Wiley, Chichester, 2000.
5. S. Vajda, S. and T. Turányi. Principal component analysis for reducing the Edelson-Field-Noyes model of the Belousov-Zhabotinsky reaction, *J. Phys. Chem*., 90, 1664-1670, 1986.
6. G. van Straten. Maximum likelihood estimation of parameters and uncertainty in phytoplankton models. In M. B. Beck and G. van Straten, editors, *Uncertainty and forecasting of water quality,* 157-171. Springer, Berlin, 1983.

Inference About the Plastic Behavior of Materials from Experimental Data

Kenneth M. Hanson

Los Alamos National Laboratory, MS-D413
Los Alamos, New Mexico 87545, USA
E-mail: kmh@lanl.gov

Abstract: Key to understanding the uncertainties in a physics simulation code is the full characterization of the uncertainties in the physics submodels incorporated in it. I demonstrate an approach to the determining the parameters of material models in a simulation code that combines the principles of physics and probabilistic Bayesian analysis. The focus is on the parameters and their uncertainties in the simulation-code submodels, as well as the numerical errors introduced in solving the dynamical equations. Bayesian analysis provides the underpinning for quantifying the uncertainties in models inferred from experimental results, which possess their own degree of uncertainty. The aim is to construct an uncertainty model for the submodels that is based on inferences drawn from comparing the code's predictions to relevant experimental results. As an example, I will present a preliminary analysis of a set of material-characterization experiments performed on tantalum to determine the parameters of the Preston-Tonks-Wallace model for plastic material behavior. I will indicate how data from a Taylor impact test may be used to update the parameters in the model by using Bayesian calibration in the context of a simulation code.

Keywords: plastic deformation model, Preston-Tonks-Wallace model, uncertainty analysis, Bayesian analysis, hierarchy of experiments, simulation uncertainty, model uncertainty, systematic uncertainty, Hopkinson-bar experiments, quasistatic-compression experiments

1. INTRODUCTION

The primary sources of data that are typically used to characterize the plastic behavior of a metal are obtained in quasi-static and Hopkinson-bar experiments. In quasi-static tests, a small cylinder of the material is typically squeezed at a constant, relatively slow rate and the change in its height is measured as a function of the load on the cylinder. These measurements are easily converted to stress and strain values. In Hopkinson-bar experiments, an elastic wave is transmitted through a thin cylinder of the material and its change in dimensions measured. Although these measurements require the use of a simulation code for precise interpretation, they are straightforwardly converted to a stress-strain curve at nearly constant strain rate. The strain rates attained in Hopkinson-bar experiments are around 10^3 per second, whereas in quasi-static tests they are typically about one per second or less.

Author's web URL: http://public.lanl.gov/kmh/

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

The analysis of these basic data is a fairly straightforward data-fitting problem, albeit a nonlinear one. The approach used here is quite standard. It is based on linearizing the response of the model output with respect to its input. The Jacobian, which characterizes the first-order sensitivities of the model, is used to minimize chi squared, that is, the mean-square differences between model predictions and the measured data, normalized to their variance. The Jacobian is also used to estimate the quadratic behavior of chi squared, and hence, the covariance matrix of the estimated parameters.

In the present example, it is necessary to introduce systematic uncertainties to account for sample-to-sample variations in material properties. The treatment of systematic uncertainties in analyzing experimental data is a topic that has not received enough attention in most analyses, let alone in the literature. The present analysis incorporates the systematic uncertainties in a straightforward way.

A major goal of any analysis is to transcribe uncertainties in the data into uncertainties in the fitted parameters. A useful self-consistency check on the results of the analysis involves propagating the uncertainties in the parameters (by means of a Monte Carlo procedure) to uncertainties in the stress-strain curves given by the PTW model. The uncertainty in the curves can be compared to the original data relative to their uncertainties to demonstrate that the model used in the analysis is consistent with the data. This test amounts to mapping the uncertainties in the data into the parameters and back again. In the context of the proposed framework [1–3], it is possible to design new experiments that can best provide data for reducing prediction uncertainty.

2. LIKELIHOOD ANALYSIS

Before describing the details of the present analysis, I will briefly review a standard approach to fitting a nonlinear model to data by the minimum chi-squared method [4]. It is assumed that one has a model to predict the values of the measured data. For each measured datum d_i, the models provides a value y_i in terms of the operating conditions of the experiment x_i and a parameter vector $\boldsymbol{a}$.

The likelihood is the probability of the measured data for a specified parameter vector $\boldsymbol{a}$. If the errors in the data are Gaussian distributed and statistically independent, the likelihood is given by

$$p(\boldsymbol{d}\,|\,\boldsymbol{a}) \propto \exp\left\{-\tfrac{1}{2}\sum_i \frac{[d_i - y_i(\boldsymbol{a})]^2}{\sigma_i^2}\right\} = \exp\left\{-\tfrac{1}{2}\chi^2\right\} \ , \tag{1}$$

where σ_i is the expected rms deviation of the measurement d_i, The corresponding value given by the model for a specified parameter vector $\boldsymbol{a}$ is designated by y_i. One recognizes the sum in the exponential in Eq. (1) as the familiar chi squared, χ^2, which quantifies the discrepancy between measurements and values predicted by a model.

The parameters that best fit the data are typically taken as those that maximize the likelihood, or equivalently, minimize χ^2. A standard tactic is to expand the model value y_i

around a particular value of the parameter vector $\boldsymbol{a}^0$ and at the value of the independent variable x_i in terms of a Taylor series,

$$y_i = y(x_i, \boldsymbol{a}) = y_i^0 + \sum_j \left.\frac{\partial y_i}{\partial a_j}\right|_{a^0} (a_j - a_j^0) + \ldots \,, \tag{2}$$

where $y_i^0 = y(x_i, \boldsymbol{a}^0)$. The complete set of derivatives make up the so-called the Jacobian matrix $\boldsymbol{J}$, which summarizes the results of a first-order sensitivity analysis. Dropping higher-order terms, chi-squared can be approximated as a quadratic function around its minimum,

$$\chi^2(\boldsymbol{a}) = \tfrac{1}{2}(\boldsymbol{a} - \hat{\boldsymbol{a}})^{\mathrm{T}} \boldsymbol{K} (\boldsymbol{a} - \hat{\boldsymbol{a}}) + \chi^2(\hat{\boldsymbol{a}}) \,, \tag{3}$$

where $\hat{\boldsymbol{a}}$ is the parameter vector at the minimum in χ^2, and $\boldsymbol{K}$ is the curvature matrix of $\chi^2(\boldsymbol{a})$ at $\hat{\boldsymbol{a}}$, which is commonly called the Hessian. The curvature can be written in terms of the Jacobian, evaluated at $\hat{\boldsymbol{a}}$, as

$$[\boldsymbol{K}]_{jk} = \left.\frac{\partial^2 \chi^2}{\partial a_j \partial a_k}\right|_{\hat{\boldsymbol{a}}} = \boldsymbol{J}\boldsymbol{J}^{\mathrm{T}} \,, \tag{4}$$

As noted above, when a flat prior is assumed, the posterior is proportional to the likelihood. In the quadratic approximation for χ^2 given in Eq. (3), the posterior is then a Gaussian

$$p(\boldsymbol{a} \,|\, \boldsymbol{d}) = \frac{1}{\det(\boldsymbol{C})(2\pi)^{n/2}} \, \exp\left\{-\tfrac{1}{2}(\boldsymbol{a} - \hat{\boldsymbol{a}})^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{a} - \hat{\boldsymbol{a}})\right\} \,, \tag{5}$$

which is written in a way to explicitly display the covariance matrix, $\boldsymbol{C}$:

$$\mathrm{cov}(\boldsymbol{a}) = \langle (\boldsymbol{a} - \hat{\boldsymbol{a}})^{\mathrm{T}} (\boldsymbol{a} - \hat{\boldsymbol{a}}) \rangle = \boldsymbol{C} = 2\boldsymbol{K}^{-1} \,. \tag{6}$$

The covariance matrix describes the degree of correlation among the uncertainties in the parameters. The analysis of nearly every type of experiment leads to off-diagonal terms in the covariance matrix, which must be stated for a full specification of the uncertainties in the parameters. The off-diagonal elements of the covariance matrix are often expressed in terms of the correlation coefficients, $\rho_{ij} = C_{ij}/\sqrt{C_{ii} C_{jj}}$.

Once the parameters and their uncertainties are obtained for a given set of experimental data, one may check their consistency with the input data by propagating random Monte Carlo draws from the inferred uncertainty distribution in the parameter space back into the data space. To draw random parameters from a Gaussian distribution with a specified covariance, one can employ a standard numerical technique: one draws a vector $\boldsymbol{r}$ with the same dimension as the parameter vector $\boldsymbol{a}$ from an uncorrelated unit-variance Gaussian distribution, and then calculates $\boldsymbol{a}^* = \boldsymbol{C}^{\frac{1}{2}} \boldsymbol{r}^* + \hat{\boldsymbol{a}}$. Here the matrix $\boldsymbol{C}^{\frac{1}{2}}$ is the square root of the covariance matrix, which may be calculated through singular-value or Cholesky decomposition [5]. The set of $\boldsymbol{a}^*$ will have the mean value of $\hat{\boldsymbol{a}}$ and the covariance $\boldsymbol{C}$.

3. MATERIAL-CHARACTERIZATION MODEL

The Preston-Tonks-Wallace (PTW) model [6] describes the plastic deformation of metals in terms of the dependence of plastic stress on plastic strain over a wide range of strain rates and temperature. The following summary of the PTW model is taken directly from Ref. 6.

In the PTW model, the plastic stress in a material is a function of the amount of strain ψ is has undergone, the strain rate $\dot{\psi}$, the material temperature T, and its density ρ. It is assumed that the plastic stress is independent of the history of the material. Furthermore, PTW ignores nonisotropic plasticity and material texture effects. Material fracture or failure is not included in PTW.

The PTW model is written in terms of three scaled dimensionless variables. The scaled stress variable is $\hat{\tau} = \tau/G(\rho, T)$, where τ is the flow stress, which is one-half the usual van Mises equivalent deviatoric stress σ, that is, $\tau = \sigma/2$, and $G(\rho, T)$ is the shear modulus, which is a function of the material density ρ and temperature T. The shear modulus is taken to be $G(\rho, T) = G_0(\rho)\,(1 - \alpha\,\hat{T})$, where $G_0(\rho)$ is the shear modulus at $T = 0$ and $\alpha > 0$ is a material parameter. The material temperature is scaled to its melt temperature T_m, which is a function of the material density ρ, $\hat{T} = T/T_m(\rho)$. For plastic flow, clearly $\hat{T} < 1$. The equivalent plastic strain is denoted by ψ. The strain rate $\dot{\psi}$ is scaled to an appropriate rate

$$\dot{\xi}(\rho, T) = \frac{1}{2}\left(\frac{4\pi\rho}{3M}\right)^{\frac{1}{3}}\left(\frac{G}{\rho}\right)^{\frac{1}{2}}\,, \tag{7}$$

where M is the atomic mass of the metal. $\dot{\xi}$ is the reciprocal of the time for a transverse sound wave to cross an atom. The strain rate always appears in the PTW formulas in terms of the ratio $\dot{\psi}/\dot{\xi}$.

For any fixed values of strain rate and temperature, the scaled stress $\hat{\tau}$ ranges between the lower and upper limits given by the yield stress $\hat{\tau}_y$ and the saturation value $\hat{\tau}_s$. The functional form for $\hat{\tau}$ depends on the strain ψ as follows

$$\hat{\tau} = \hat{\tau}_s + \frac{1}{p}\,(s_0 - \hat{\tau}_y)\ln\left\{1 - [1 - \exp(-p\,r)]\exp\left[-\frac{p\,\theta\,\psi}{(s_0 - \hat{\tau}_y)\,[\exp(p\,r) - 1]}\right]\right\}\,, \tag{8}$$

where p and θ are material-specific parameters and $r = (\hat{\tau}_s - \hat{\tau}_y)/(s_0 - \hat{\tau}_y)$. The parameter s_0 is explained below.

At strain rates below $10^8\,\mathrm{s}^{-1}$ for tantalum at room temperature, the plastic deformation process is controlled by thermal activation. The values for $\hat{\tau}_y$ and $\hat{\tau}_s$ are given by

$$\hat{\tau}_y = y_0 - (y_0 - y_\infty)\,\mathrm{erf}\left[\kappa\,\hat{T}\ln\left(\frac{\gamma\,\dot{\xi}}{\dot{\psi}}\right)\right]\,, \tag{9}$$

$$\hat{\tau}_s = s_0 - (s_0 - s_\infty)\,\mathrm{erf}\left[\kappa\,\hat{T}\ln\left(\frac{\gamma\,\dot{\xi}}{\dot{\psi}}\right)\right]\,, \tag{10}$$

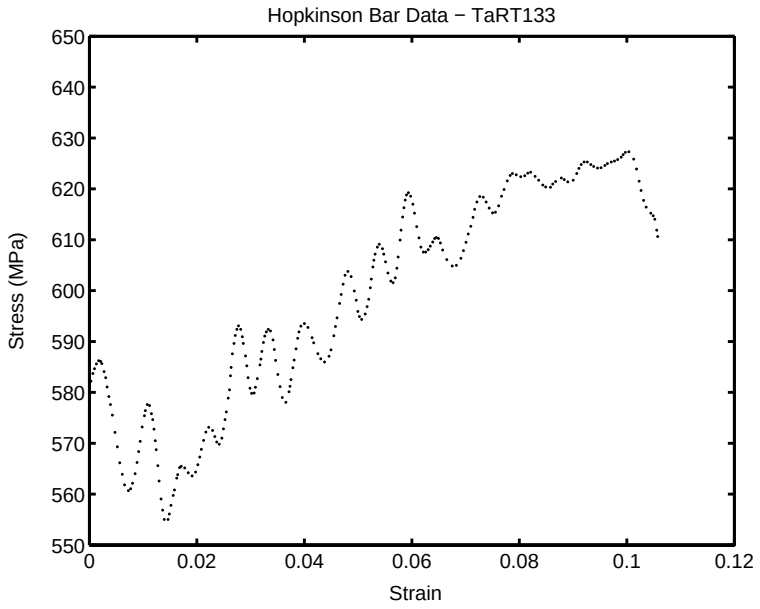


Figure 1. Plot of data obtained from Hopkinson-bar experiment done on tantalum at room temperature and a strain rate of 1300 s^{-1}.

where κ and γ are dimensionless material-related parameters. The error function, defined as $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-t^2\right) dt$, has the limiting values $\mathrm{erf}(0) = 0$ and $\mathrm{erf}(\infty) = 1$. Note that the logarithm is nonnegative because $\gamma\,\dot{\xi}/\dot{\psi} \geq 1$ in the low-strain-rate regime. Therefore, the argument of the error function is nonnegative. The parameters y_0 and y_∞ are the values that $\hat{\tau}_y$ takes at zero temperature and very high temperatures, respectively; s_0 and s_∞ have analogous meanings for $\hat{\tau}_s$.

The PTW model is designed to extend the range of normal plastic-flow models to very high strain rates, above $10^8\,\mathrm{s}^{-1}$, in which regime it relies on Wallace's theory of overdriven shocks in metals [7]. Because Taylor experiments, which are the goal of the present study, do not reach these very high strain rates, the formulas that apply in that regime will not be given. Suffice it to say that the PTW parameters β, y_1, and y_2 do not have an effect in the lower strain-rate region.

4. ANALYSIS OF MATERIAL-CHARACTERIZATION EXPERIMENTS

I now outline the analysis of the material-characterization experiments to estimate the PTW parameters and their uncertainty. In a sense, the approach is a straightforward chi-squared (or least-squares) analysis, but it has some noteworthy aspects, for example, the inclusion of systematic uncertainties.

Basic stress-strain data at moderate strain rates (about $10^3\,\mathrm{s}^{-1}$) are typically obtained in a Hopkinson-bar experiment in which an elastic wave is passed through a thin cylinder of the material under investigation. Strain gauges mounted on the support cylinders measure strain as a function of time. From these measurements, the stress-strain behavior of the material is inferred. The data from a Hopkinson-bar experiment on tantalum done at

room temperature (298°K) and a strain rate of 1300 s^{-1} are shown in Fig. 1. This figure shows a well-known feature of Hopkinson-bar experiments, the presence of wiggles in the measured stress as a function of strain, which are particularly evident at strains of 0.02 and below but also observable at higher strains. These oscillations, caused by elastic wave dispersion within the sample and apparatus, tend to reduce the accuracy of Hopkinson-bar data. The data below strains of 0.017 seem unreliable because of their higher amplitude of the oscillations and the fact that the stress rises as the strain approaches zero instead of falling. Likewise, the data above a strain of 0.1 seem to be corrupted by an artifact. In the present analysis, the data in these two end regions are excluded.

To make use of the data between 0.017 and 0.1, their uncertainties need to be quantitatively characterized. The approach taken here is to treat the fluctuations in these data as a random Gaussian process with zero mean. The underlying assumption is that the measured signal fluctuates around its true value. The process is described as a random signal drawn from a Gaussian distribution with a covariance between two strains ϵ_i and ϵ_j given by $\sigma^2 \exp(-|\epsilon_i - \epsilon_j|^2/\lambda^2)$. The fluctuations are analyzed by first fitting a quadratic function to the data and subtracting it from them. The value of σ is the rms deviation of the resulting data set, which is found to be 4.9 MPa, or 0.8 %. By maximizing the likelihood function for the data, the correlation length λ is determined to be 0.0019. To avoid giving inappropriate weight to these data in the subsequent analysis, the measured data points are thinned by a factor equivalent to one sample per correlation length, which corresponds to using only every fourth measurement in this case. A similar analysis of the other Hopkinson-bar data sets yields σ values between 3.1 MPa and 5.3 MPa, and correlation lengths between 0.0027 to 0.0055.

The same type of analysis is carried out for the quasi-static measurements. However, these data do not exhibit the same oscillatory fluctuations around a smooth curve and the meaning of this kind of analysis is less clear. The rms deviation of the quasistatic data from a smooth second- or third-order curve is about 0.35 MPa, or only about 0.04% to 0.09%. Each of these data sets are also thinned out to around ten data points to avoid giving them undue influence in the following analysis.

Particularly for the quasi-static data, experience suggests caution in accepting the results of the above analysis without critical assessment. A number of assumptions are typically made about the physics in these experiments, which may well limit the overall accuracy of the reported data to a few percent. The actual uncertainties in these data are ultimately dominated by the systematic effects discussed below.

The above analysis leads to the final set of tantalum data shown in Fig. 2, in which the standard error in each datum is displayed as an error bar.

In addition to the suite of data shown in Fig. 2, an auxilary set of four experiments were carried out at room temperature and 0.001 s^{-1} strain rate to ascertain the degree of reproducibility from one tantalum sample to the next. The samples were selected from different lots with various processing histories to get a good idea of the amount of variability one might see in commercially obtained tantalum. With just four samples to go on, the rms deviation in measured stress values is about 8%. Note that this is a systematic error that is common to all the data from a single tantalum sample.

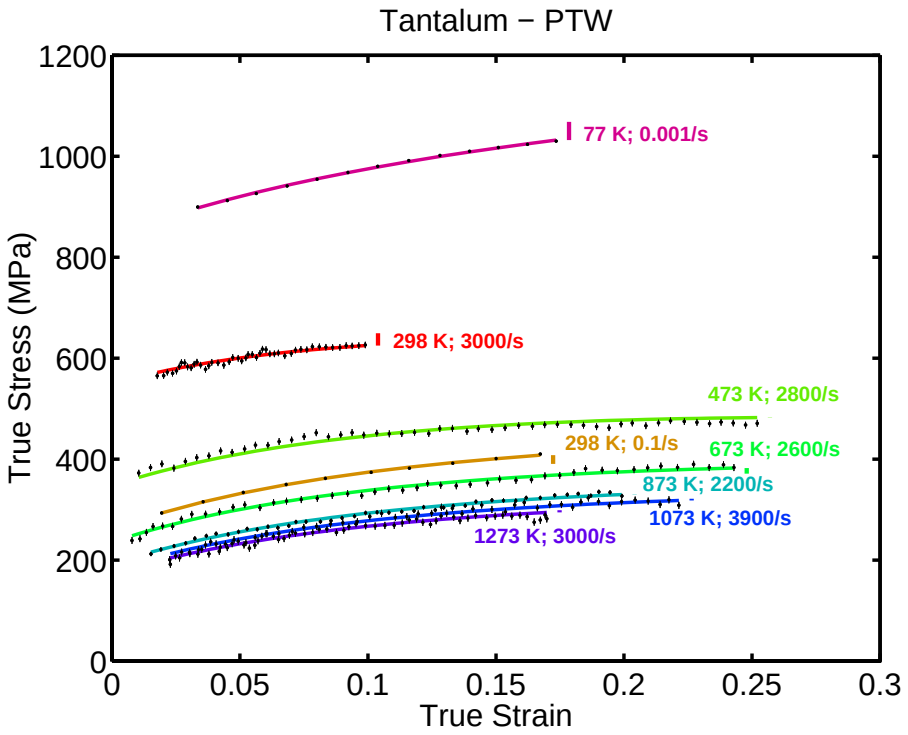


Figure 2. Comparison between data from material-characterization experiments for a variety of temperatures and strain rates and the PTW model fit to the data, shown as lines. The vertical bar to the right of each curve indicates the estimated systematic offset for that curve.

The tantalum specimens that yielded the data shown in Fig. 2 were obtained from a single lot so one would expect less sample-to-sample variation than in the above study. It seems reasonable to assign a systematic standard error of 3% to each of the displayed data sets. This assignment is supported by the degree of consistency between the data and the fitted model.

I include this sample-to-sample variability in the analysis by treating it as a systematic uncertainty. Because the observed differences between different samples amount to a small vertical shift of the curves, it is a good approximation to incorporate them in terms of an additive offset parameter for each curve. This effectively adds eight more parameters to the model, which need to be estimated as well as the parameters in the PTW model. To include this systematic effect in a full analysis of the eight data sets, the appropriate expression for the minus-log-likelihood is

$$-\log p(\boldsymbol{a}\,|\,\boldsymbol{d}) = \tfrac{1}{2}\sum_k \chi_k^2 + \tfrac{1}{2}\sum_k \frac{\Delta_k^2}{\sigma_k^2}\,, \tag{11}$$

where the index k identifies the data set obtained with a specific tantalum sample. The first term is a sum of χ^2 values for each data set at a specific temperature and strain rate. The second sum represents the prior probability for the offset parameters Δ_k, and σ_k is the rms deviation of the prior distribution on Δ_k, in the current analysis taken to be 3%

Table 1. PTW parameters for tantalum obtained from this preliminary analysis (fit4a) of the data shown in Fig. 2. All parameters are dimensionless.

Parameter	Value	Standard error
y_0	0.0123	0.0006
y_∞	0.00164	0.00004
s_0	0.0164	0.0007
s_∞	0.00308	0.00005
κ	0.91	0.08
γ	0.0000024	0.000002
θ	0.0145	0.0002

Table 2. Correlation coefficient matrix for the PTW model parameters obtained from the fit to the data shown in Fig. 2. The covariance matrix is estimated using Eqs. (4) and (6).

Parameter	y_0	y_∞	s_0	s_∞	κ	γ	θ
y_0	1	0.186	0.988	0.400	0.687	-0.464	-0.182
y_∞	0.186	1	0.208	0.913	0.142	0.022	-0.140
s_0	0.988	0.208	1	0.432	0.713	-0.496	-0.299
s_∞	0.400	0.913	0.432	1	0.443	-0.263	-0.257
κ	0.687	0.142	0.713	0.443	1	-0.935	-0.119
γ	-0.464	0.022	-0.496	-0.263	-0.935	1	0.087
θ	-0.182	-0.140	-0.299	-0.257	-0.119	0.087	1

of the mean stress value of the the kth data set. This term is needed to constrain the offset of the curves. Without it, the PTW parameters would be indeterminate.

The above model is fit to the data shown in Fig. 2 using the general approach described earlier to minimize the function given in Eq. (11). The Jacobian (sensitivity) matrix is estimated at each optimization iteration by the straightforward method of finite differences. In this fit, the following parameters are taken to be fixed: $G_0 = 654$ kilobars, $\alpha = 0.47$, $p = 4$, $A = 180.95$, $T_m = 3250$°K, and $\rho = 16.6$ g/cm^3. The value $p = 4$ is determined by auxiliary experiments that reach total strains of almost unity.[8] The following PTW parameters are not important because the strain rates are not high enough: y_1, y_2, and β. Adiabatic heating of the high strain-rate samples is taken into account, assuming a specific heat capacity of 0.15.

The fit to the data obtained in this preliminary analysis is shown in Fig. 2. The value of chi squared for 314 data points for this fit is 563, which corresponds to a chi squared per degree of freedom of approximately 1.9. The vertical bars to the right end of each curve

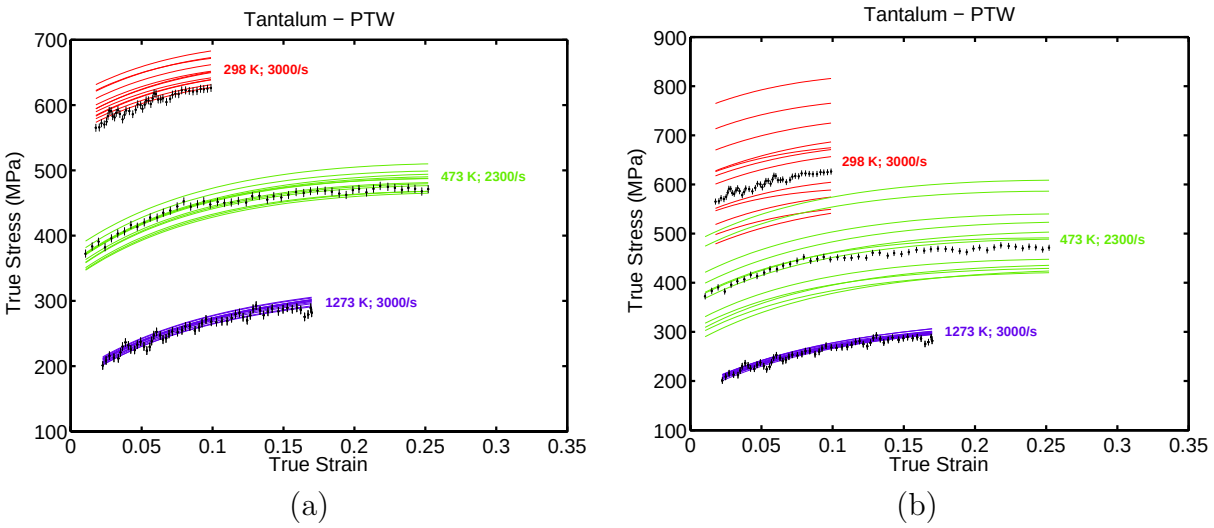


Figure 3. (a) A set of plausible stress-strain curves for three data sets (from top to bottom, 298°K, 1300 s^{-1}; 298°K, 0.1 s^{-1}; 1073°K, 3900 s^{-1}) obtained by drawing Monte Carlo samples from the uncertainty distribution in the PTW parameters as derived from the data shown in Fig. 2, and evaluating the PTW formulae. (b) Same as for (a), except that the correlations given in Table 2 are neglected. The ranges of variation for these curves are up to twice as larger as those observed in (a).

display the fitted value for the offset of that curve. The eight offset values range from –25 MPa to 14 MPa. The offsets contribution to chi-squared is 12.3 (the second term in Eq. 11), which is a reasonable value for eight parameters demonstrating consistency with their assumed rms deviation of 3%. The PTW parameters obtained from the fit and their rms uncertainties are given in Table 1. As important as the uncertainties in the individual parameters are, their correlation coefficients, presented in Table 2, are equally important. Use of the rms errors without consideration of the correlation coefficients would seriously misrepresent the results of this analysis, as will be demonstrated next.

The Monte Carlo technique described in Sect. 2 can be used to randomly draw PTW parameter vectors from their uncertainty distribution specified in Tables 1 and 2. Figure 3 shows a plot of the stress-strain curves that result from 12 such random draws for three data sets at different operating conditions. Because negative values of γ are not allowed in the PTW formalism and its relative error is so large, a log-normal distribution is used for that variable. From the comparison of these curves to the Hopkinson-bar measurements, one can conclude that the parameters and their uncertainties plausibly represent the data. However, the high value of chi squared mentioned above needs to be considered in a final analysis of these data. It indicates that either the uncertainties assigned to the data are too small or a discrepancy exists between the PTW model and the data. The Centroidal Voronoi Tessellation (CVT) [9, 10] algorithm is used here for drawing samples with more uniform spacing than a set of random samples would provide.

5. ANALYSIS OF TAYLOR DATA USING A SIMULATION CODE

The Taylor impact test represents an experiment of an intermediate level of complexity in the hierarchy of experiments chosen to elucidate the material model for tantalum. In the Taylor impact test, a cylindrical sample of material is propelled into a fixed, rigid surface. Taylor tests are often performed to confirm the plastic-behavior model of a material under severe strain conditions. Extremely high plastic strains and strain rates occur at the crushed end of the rod, resulting in severe local deformation [11, 12]. The experimental data usually consist of measurements of its final deformed profile.

The data from a Taylor experiment may be analyzed in much the same way as was done above for the material-characterization experiments. Systematic experimental uncertainties in the impact velocity, for example, may be included in a way similar to that used above to include the sample-to-sample variations. A contribution similar to the second term in Eq. (11) is necessary to account for the systematic offset for the specific sample used in the Taylor test. One viable approach to chi-squared minimization is to use the same methods as described above. When the simulation code is treated as a black box, the Jacobian matrix may be estimated by the method of finite differences. When the simulation code is available, however, the more sophisticated method of automatic differentiation may prove useful [13, 14].

Bayes theorem provides the proper means for combining the prior probability density function from the first analysis with the likelihood of the subsequent Taylor analysis [2]. The uncertainties from the above analysis may be included by adding to the expression given in Eq. (11) a term to represent the prior for the Taylor analysis, namely $\frac{1}{2}(\boldsymbol{a} - \hat{\boldsymbol{a}}_1)^{\mathrm{T}}\boldsymbol{C}_1^{-1}(\boldsymbol{a} - \hat{\boldsymbol{a}}_1)$, where $\hat{\boldsymbol{a}}_1$ is the PTW parameter vector estimated in the foregoing analysis and $\boldsymbol{C}_1$ is the estimated covariance matrix.

When this process is employed to simply adjust the values of the model parameters to match the new data, it is called calibration, which is different from what I am proposing. By basing this parameter-updating process on Bayes theorem and quantitative uncertainty estimates, the process of Bayesian calibration becomes one of continuing inference [13]. In a sense, the Monte Carlo technique for estimating uncertainties in simulation-code output described above is reversed; the uncertainties in the parameters are determined from the combined uncertainties in the measurements and the effects on the simulation of uncertainties in experimental set up. Since the inference procedure involves determining the uncertainties in model parameters, it provides the means for predicting the uncertainty in simulation output in other physical scenarios. Further details of the process are presented in Ref. 2.

The focus here has been on the parameters in the Preston-Tonks-Wallace model. The Bayesian methodology can address other questions, for example, comparison of two or more competing models to decide between them. It is generally applicable to answering all questions that one might pose about models [15].

Acknowledgments

I gratefully acknowledge Shuh-Rong Chen's assistance, both by supplying the experimental data presented here and by providing considerable guidance in regard to material-

science issues. The following people have kindly shared their insights with me: Francois Hemez, Rusty Gray, Paul Maudlin, Larry Hull, Tom Duffey, Mark Anderson, Bill Oberkampf, Tim Trucano, Scott Doebling, C. Shane Reese, Dave Higdon, Mike McKay, and Kathy Campbell.

REFERENCES

1. K. M. Hanson and G. S. Cunningham. Validation of hydrocode predictions (U). In Edward Caramana and Charlie McMillan, editors, *Proc. 8th Nucl. Expl. Code Devel. Conf.*, pages 457–462. Report LA-12963-C, Los Alamos National Laboratory, 1995. (separately as Report LA-UR-01-6671).
2. K. M. Hanson. A framework for assessing uncertainties in simulation predictions. *Physica D*, 133:179–188, 1999.
3. K. M. Hanson and François M. Hemez. A framework for assessing confidence in computational predictions. *Exper. Techniques*, 25:50–55, 2001.
4. P. R. Bevington and D. K. Robinson. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, Boston, 1992.
5. G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Univ. Press, Baltimore, 1996.
6. Dean L. Preston, Davis L. Tonks, and Duane C. Wallace. Model of plastic deformation for extreme loading conditions. *J. Appl. Phys.*, 93:211–220, 2003.
7. Duane C. Wallace. Irreversible thermodynamics of overdriven shocks in solids. *Phys. Rev. B*, 24:5597–5606, 1981.
8. S. R. Chen, private communication. (2004).
9. Qiang Du, Vance Faber, and Max Gunzburger. Centroidal Voronoi Tessellations: Applications and algorithms. *SIAM Rev.*, 41:637–676, 1999.
10. Lili Ju, Qiang Du, and Max Gunzburger. Probabilistic methods for centroidal Voronoi tessellations and their parallel implementations. *Parallel Computing*, 28:1477–1500, 2002.
11. G. T. Gray III *et al.* Bridging the experimental-modeling gap - the Taylor test: an integrated test used to validate constitutive models. *to appear in AMPTIAC Newsletter*, 2002.
12. P. J. Maudlin *et al.* On the modeling of the Taylor cylinder impact test for orthotropic textured materials: experiments and simulations. *Int. J. Plasticity*, 15:139–166, 1999.
13. K. M. Hanson, G. S. Cunningham, and S. S. Saquib. Inversion based on computational simulations. In G. J. Erickson, J. T. Rychert, and C. R. Smith, editors, *Maximum Entropy and Bayesian Methods*, pages 121–135. Kluwer Academic, Dordrecht, 1998.
14. A. Griewank. *Evaluating Derivatives: Principles and Techniques of Automatic Differentiation*. SIAM, Philadelphia, 2000.
15. D. S. Sivia. *Data Analysis - A Bayesian Tutorial*. Clarendon, Oxford, 1996.

COMMENTS ON GLUE

Renata Romanowicz and Keith Beven

Lancaster University, Lancaster, LA1 4YQ, United Kingdom
E-mail: R.Romanowicz@lancaster.ac.uk

Abstract: The paper presents an application of GLUE (Generalised Likelihood Uncertainty Estimation) methodology to the problem of estimating the uncertainty of predictions produced by environmental models. The methodology is placed in a wider context of different approaches to inverse modelling and, in particular, a comparison is made with Bayesian estimation techniques based on explicit structural assumptions about model error. Using a simple example of a rainfall-flow model, different evaluation measures and their influence on the prediction uncertainty and confidence limits are demonstrated.

Keywords: Bayesian inference, Deterministic environmental models, Generalised Likelihood Uncertainty Estimation (GLUE), Model determination.

1. INTRODUCTION

The Generalised Likelihood Uncertainty Estimation (GLUE) technique [1] was introduced partly to allow for the possible equifinality (non-uniqueness, ambiguity or non-identifiability) of parameter sets during the estimation of model parameters in over-parameterised models. The technique has been applied to a variety of environmental problems [2]. Its popularity results from the very few assumptions that it requires and the simplicity of the approach when used in practical applications. GLUE assumes that, in the case of large over-parameterised models, there is no inverse solution and, hence, that the estimation of a unique set of parameters, which optimise goodness-of fit-criteria given the observations, is not possible. The technique is based on the estimation of the weights or probabilities associated with different parameter sets, based on the use of a subjective likelihood measure to derive a posterior probability function, which is subsequently used to derive the predictive probability of the output variables. In [3] a statistically motivated, more formal equivalent of GLUE was developed. The idea was to explicitly derive the likelihood function based on the error between the observed outputs and those simulated by the model. This formal approach is equivalent to Bayesian statistical estimation: it requires assumptions about the statistical structure of the errors. GLUE is usually applied by directly likelihood weighting the outputs of multiple model realisations (either deterministic or stochastic, defined by sets of parameter values within one or more model structures) to form a predictive distribution of a variable of interest. Prediction uncertainties are then related to variation in model outputs, without necessarily adding an additional explicit error component. There is thus an interesting question as to whether an appropriate choice of likelihood measure can result in similar results from the two approaches.

There are a number of possible measures of model performance that can be used in this kind of analysis. The only formal requirements for use in a GLUE analysis are that the likelihood measure should increase monotonously with increasing performance and be zero for models considered as unacceptable or non-behavioural. Application-oriented measures are easily used in this framework. Measures based on formal statistical assumptions, when applied to all model realisations (rather than simply in the region of an “optimal” model)

should give results similar to a Bayesian approach when used within a GLUE framework [3], but the assumptions made (additive Gaussian errors in the simplest cases) are not always easily justified in the case of nonlinear environmental models with poorly known boundary conditions (see the discussion in [4]). In this paper, we shall explore the influence of the choice of observation based likelihood weights on the predictive uncertainty of the model.

2. RELATION BETWEEN GLUE AND STATISTICAL APPROACHES: DISCUSSION OF LIKELIHOOD MEASURES

There is a question as to how far GLUE can be consistent with formal statistical approaches since it weights the predictions of different feasible models without necessarily using an explicit error model, albeit that the likelihood weights are determined from a calculation that depends in some way on the model errors. In a Bayesian framework, the inverse problem is usually posed in the context of the observation equation:

$$y_t = g(u_0,...,u_t,\xi_0,...,\xi_t,\zeta_t,\theta,t) \tag{1}$$

or assuming an additive error model:

$$y_t = g(u_0,...,u_t,\xi_0,...,\xi_t,\theta,t) + \zeta_t \tag{2}$$

where y_t denotes the observed model output; u_t is model input, θ denotes vector of model parameters and the errors on the inputs, ξ_t, and the measurement error ζ_t are not known, and, in the general case, may be non-Gaussian (and might indeed be required to compensate for model structural error). The Bayesian approach allows detailed studies to be targeted at the most informative areas and variables (e.g. [5]). It also allows parameter and error estimates to be updated as more observations become available: for example, in the case of Gaussian normal assumptions (see e.g. [6]), the Bayesian estimator of the state variables in a linear, stochastic, dynamic system is the famous *Kalman Filter* (KF: [7]); while the Bayesian estimator of the constant or stochastic, time variable parameters in a linear-in-the-parameters regression model is the classic recursive least squares estimation algorithm. Environmental models are not normally that simple, and it might be difficult to formulate an appropriate error model. In that case there might be advantage in the use of a nonparametric representation of the errors ([8], [9], [10]).

Measures of goodness of fit applied in order to compare different nonparametric probability density functions (or log likelihood ratios) include:

(i) Mean Square Error (MSE),

$$MSE = E[\hat{f}(x) - f(x)]^2 = Var\{\hat{f}(x)\} - [E[\hat{f}(x)] - f(x)]^2, \tag{3}$$

which combines both variance and bias of the estimates;

(ii) averaged Predictive Squared Error (PSE), related to MSE:

$$PSE = E[\hat{f}(x) + \varepsilon - f(x)]^2 = MSE + \sigma^2; \tag{4}$$

where σ^2 denotes the prediction error variance; and

(iii) Mean Integrated Square Error (MISE):

$$MISE = E[\sum_{i=1}^{N} (\hat{f}(x_i) - f(x_i))^2]. \tag{5}$$

Instead of L_2 norm used in the above definitions, we can use L_∞ norm: $\sup_x |\hat{f}(x) - f(x)|$; L_1 norm: $\sum_{i=1}^{N} |\hat{f}(x_i) - f(x_i)|$ or even L_p norm: $\sum_{i=1}^{N} |\hat{f}(x_i) - f(x_i)|^p$. The influence of these different norms on the estimates of probability density function is discussed in [9].

[8] discusses the use of a Generalised Gaussian distribution in inverse modelling. It is defined as the normalised probability density f(x) with fixed L_p norm estimator of dispersion: $\sum_{i=1}^{n} n |x_i - x_0|^p f(x_i) = (\sigma_p)^p$ (discrete case), which has the minimum information content (widest spreading) and has the form:

$$f_p(x) = \frac{p^{1-1/p}}{2\sigma_p \Gamma(1/p)} \exp\left[-\frac{1}{p} \frac{|x - x_0|^p}{(\sigma_p)^p} \right] \tag{6}$$

where $\Gamma(.)$ denotes Gamma function and x_0 is the centre of f(x) in the L_p norm sense.

For p=2 $f_2(x)$ is the Gaussian function with mean x_0 and standard deviation σ_2. The Generalised Gaussian distribution covers a range of distributions from the symmetric exponential to a box-car distribution for $p = \infty$. In the case of general exponential distribution family, the log-likelihood based on (2) is equal to the sum of the functions of errors $x - x_0$ and depending on the error structure (in this case the parameter p), we shall get the criteria related to L_1, L_2 or L_p norm.

In [3], it is assumed that, for a general, nonlinear model, the distribution of errors is Gaussian with unknown mean and variance and the log-likelihood function has a sum of squared errors form, equivalent to the Nonlinear Least Squares approach. In this particular case, the equivalence of the measure of fit between the model output and observations and assumed error structure follows from the equivalence between likelihood and least square approaches for the mean of the distribution of independent errors ([11], [10]). Following this approach, the predictive distribution of output variables y_t, modelled by Eq.2, conditioned on the calibration data **z** is given by (discrete case):

$$P(y_t < y \mid \mathbf{z}) = \sum_{\theta} \sum_{\phi} P(y_t < y \mid \theta, \phi) f(\theta, \phi \mid \mathbf{z}) \tag{7}$$

where $f(\theta, \phi \mid \mathbf{z})$ is a posterior likelihood function for the parameters:

$$f(\theta, \phi \mid \mathbf{z}) \sim \exp(-\sum_{t=1}^{T} (z_t - g_t(\theta) - \mu)^2 / 2\sigma^2); \tag{8}$$

$\phi = (\mu, \sigma)$ denotes a vector of statistical model parameters and $P(y_t < y \mid \theta, \phi)$ is a standard normal distribution function N(0,1).

The GLUE methodology is closer in philosophical terms to a nonparametric approach. The critical difference is that posterior distributions for predicted variables are estimated directly from the outputs of a set of multiple acceptable or behavioural models, rather than from a model and an additive residual model. In [2] various likelihood measures are presented and applied to different environmental problems. These include measures similar in nature to

(6) (but also fuzzy measures and binary measures that can be used to exclude some models as non-behavioural). Romanowicz and Beven [3] have also shown how a formal error model may be used within the GLUE framework by evaluating likelihood weights over both model and error model parameters. GLUE allows that the likelihood surface may be very complex in form because of complex parameter interactions. It is the set of parameters that produces a behavioural model for a given input sequence, and there may be no well-defined posterior distribution for individual model parameters.

In the example that follows, we concentrate on two issues: (i) definition of a likelihood measure for use within GLUE with a well-defined scaling parameter based on the dispersion of the errors to control the width of prediction limits; (ii) the influence of the choice of different likelihood measures on the model predictive uncertainty.

3. THE MODELLING PROBLEM: APPLICATION OF RAINFALL-FLOW MODEL TO THE CAN VILA CATCHMENT, N. E. SPAIN

To illustrate the proposed methodology, we consider a rainfall-flow model for a set of rainfall-flow data. The study catchment area of Can Vila is situated in Spain, in the Valcebre catchment of the South-East Pyrenees ([12], [13]). The catchment is 0.56 km^2 in area and is partly covered by Pinus Sylvestris. Rainfall and flow measurements at 20 min. time steps were made available for winter 1995/96 and summer 1997 events. We shall use only winter observations in this study with hourly time step.

The data were modelled using two forms of the mechanistic rainfall-runoff model TOPMODEL ([14]). The SIMULINK version of TOPMODEL, described in [15], bases its calculations of the spatial patterns of hydrological response on the pattern of a topographic index for the catchment derived from a Digital Terrain Model (DTM). The time series data used by the model are the rainfall, runoff and evaporation averaged over the catchment. The model has a modular structure. The saturated zone model is assumed to be non-linear with the outflow $Q_b(t)$ calculated as an exponential function of a catchment average soil moisture deficit S_3 as:

$$\frac{dS_3}{dt} = Q_b(t) - Q_v(t)$$

$$Q_b(t) = Q_0 \exp(-S_3(t)/m) \qquad (9)$$

where Q_0=SK0 $e^{-\lambda}$ is the flow when $S_3(t)=0$. and $Q_v(t)$ denotes the recharge to the saturated zone. SK0 is a soil transmissivity parameter, m is a parameter controlling the rate of decline in transmissivity with increasing soil moisture deficit and λ is the mean value of the topographic index distribution in the catchment (see [16]). Other parameters control the maximum storage available in the root zone (LRZ) and the rate of recharge to the saturated zone (KS).

In the first step the MC sensitivity analysis was performed using the full version of TOPMODEL and January 1996 rainfall-flow data. Following an initial sensitivity analysis the parameter ranges were chosen to ensure that the range of the simulations covers the observations. 10000 simulations were then performed varying the four TOPMODEL parameters according to prior distributions shown in Table 1.

Table 1. Parameter distributions applied in MC analysis of TOPMODEL

	distribution	Min value	Max value	mean	std
SKO	uniform	10	500	251	141
m	uniform	0.003	0.03	0.017	0.0075
LRZ	Log-uniform	1.e-4	0.01	0.0147	0.023
KS	Log-uniform	1.e-15	0.01	0.0003	0.0012

4. ESTIMATION OF DISCHARGE PREDICTION LIMITS

Two methods were then applied to estimate prediction limits: (i) a formal likelihood function based on an assumed error model, and (ii) a non-formal GLUE approach with the likelihood weights proposed by Eq. 12-13.

In both approaches, following [3], we used a multiplicative error model to account for the usual tendency of rainfall-model errors to increase with increasing magnitude of the prediction. Thus:

$$\zeta_t = \log(\delta_t) = \log(Q_{obs,t}) - \log(Q_{sim,t}(\theta)) \quad (10)$$

where $Q_{obs,t}$ denotes the observation of flow at time t and $Q_{sim,t}(\theta)$ denotes the simulated flow for a given model run, depending on parameter set θ.

We then applied the error model (2) with the assumption for the vector error $\zeta_t \sim N(\mu, \Sigma)$; where μ denotes the unknown mean of the errors and $\Sigma = I\sigma^2$ is the covariance matrix. The observation sets for the conditioning of the estimates were chosen in such way that the correlation between the observations could be neglected. Eq. 7 can be used to estimate the predictive uncertainty of the model both for the calibration and the validation stages, under the assumption that the distribution of errors remains the same during the validation stage. Fig. 1 presents the predictions together with 95% confidence limits for the calibration (upper panel) and validation (lower panel) periods.

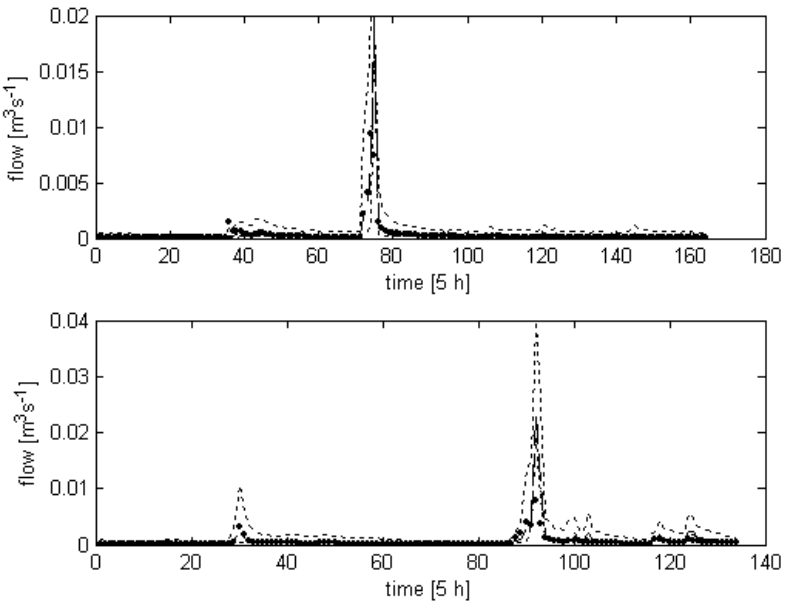


Figure 1. Flow predictions together with 95% confidence limits for Can Villa catchment; upper panel: calibration stage – December 1995, lower panel: validation event January 1996; dashed lines denote the 95% confidence limits, dots represent the observations; formal approach. Note change in discharge scale.

In the GLUE approach the prediction equation takes the form:

$$P(\hat{y}_t < y \mid \mathbf{z}) = \sum_i \{f(\theta_i \mid \mathbf{z}) \mid \hat{y}_t < y)\} \tag{11}$$

We look for the weights $f(\theta_i \mid \mathbf{z})$, which will account for both prediction and parameter/structure related errors. By analogy with (7-8) we assume the form:

$$f(\theta_i \mid \mathbf{z}) = \exp(-\sum_{t=1}^{T} (\log(Q_{t,sim}(\theta_i) - \log(Q_{t,obs}))^2 / \sigma^2) \tag{12}$$

In the formal approach, with an explicit error model, σ^2 is the variance of the prediction error based on the observations. The optimal value of this variance may be derived from the likelihood function (8).

In the non-formal GLUE approach, however, we can treat σ^2 as an additional scaling parameter reflecting our lack of knowledge of the true information content of the residuals ζ_t in constraining the uncertainty in the model predictions. One possible form for this scaling is to take the sum of the variances of the simulated flows over all the behavioural models and all time steps as an estimate, such that:

$$\sigma^2 = \sum_{t=1}^{T} \text{var}(\log(Q_{t,sim}(\theta)) \tag{13}$$

This will increase the dispersion of the resulting posterior likelihoods (relative to the formal case) to account for the predictive model uncertainty without making additional assumptions about the model error structure. This scaling could also be made time-dependent, since the cumulated variance of simulated flows at each time step can be calculated over some specified process memory, in a way similar to allowing a variable kernel in kernel density estimation ([9]) but without making any assumptions about the form of the likelihood surface.

The resulting predictions together with 95% confidence bands for the calibration period are shown in Figure 2.

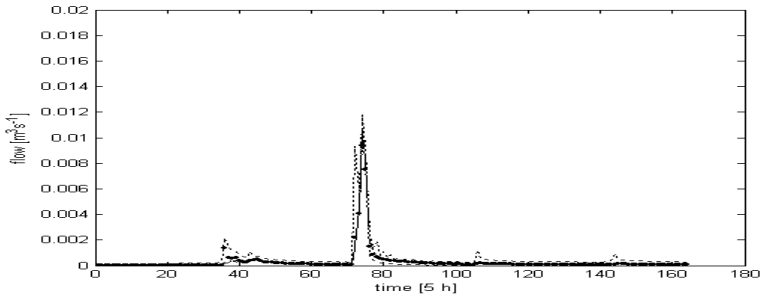


Figure 2. Flow predictions together with 95% confidence limits for Can Villa catchment; December 1995, dashed lines denote the 95% confidence limits, dots represent the observations; non-formal GLUE approach.

Comparison with the Fig. 1 shows that, in this case, the simplified GLUE method gives smaller over-prediction and better represents the observations. Results for a further evaluation period, using the same set of behavioural models and likelihood weights determined for the calibration period, are shown in Fig. 5, upper panel. In this case the GLUE method also gives very good results.

5. INFLUENCE OF THE CHOICE OF OBSERVATION SETS ON THE MODEL PREDICTIVE UNCERTAINTY: COMPARISON OF DIFFERENT GOODNESS OF FIT CRITERIA

The availability and quality of observations is often a major constraint on the identifiability of environmental models. In addition, different prediction problems might require different types of model evaluation. In the case of rainfall-flow models, there is usually sufficient amount of observations available but input errors and model structural errors can give rise to complex error structures for any model run, including heteroscedasticity, nonstationarity and correlation. We can attempt to model these complexities (as in [3]) but experience suggests that less formal methods can still provide useful prediction bounds. In what follows we shall compare the uncertainty predictions obtained when different (reduced) observation sets are chosen for the conditioning of the predictions and different norms are used to evaluate the likelihood weights. We shall use the non-formal GLUE approach in this comparison with L_2 norm in (12) replaced by L_1 norm and also we shall compare the use of observations from the whole time period with conditioning only on time steps with peak discharge observations (global MISE and local MSE goodness of fit criterion) used for the conditioning (as these criteria may use different norms).

As an example, Fig. 3 shows the results of conditioning on the two highest peak values only in terms of the resulting cumulative density functions (cdfs) for the parameters integrated over all the behavioural parameter sets for likelihood weights based on an L_1 norm (dashed

lines). These are compared with the cdfs obtained using likelihood weights based on L_2 norm and the same type of conditioning (dotted lines) and the likelihoods with scaling (12) conditioned on the whole range of observations (solid lines).

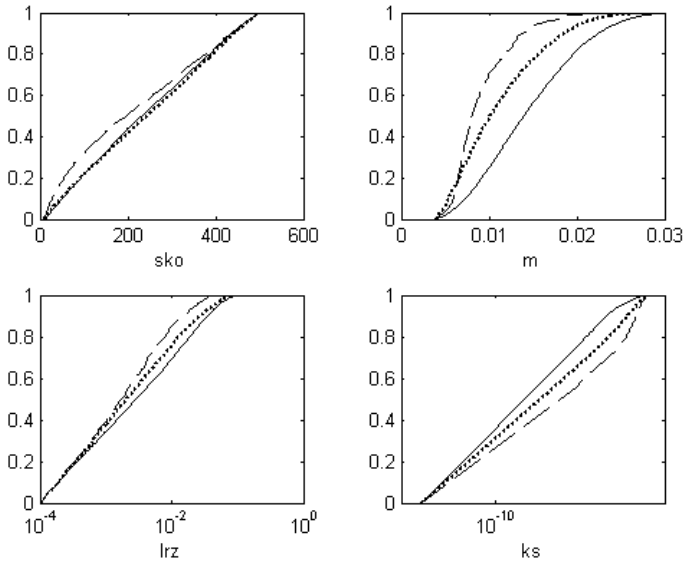


Figure 3. Comparison of the cdfs for the parameters obtained using two criteria for the observation based weights; continuous lines correspond to the weights (12) conditioned on the whole dataset (MISE L_2 criterion); dashed lines show cdfs for the 2 peak values of flow, norm L_1; dotted lines show cdfs for the 2 peak values, L_2 norm.

The results from Fig. 3 show significant differences of posterior distributions of parameters when different observation sets and likelihood weights are used. However, the results were less sensitive to the use of different norms (L_1, or L_2). The resulting prediction limits shown in Fig. 4 are also affected by the choice of the observation sets, with confidence limits for the conditioning on the 2 peak observations of flow better following the peak values but over predicting low flows.

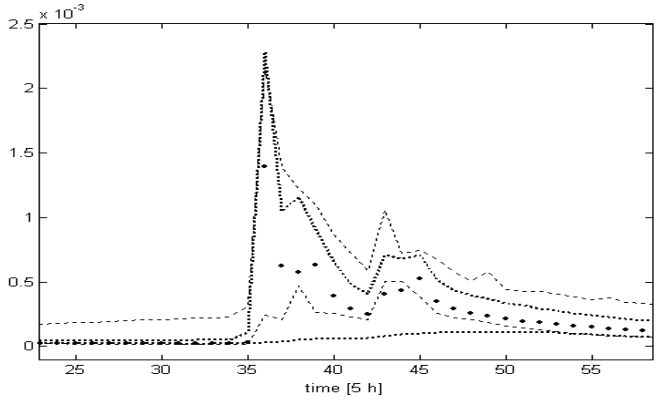


Figure 4. Comparison of confidence levels obtained using two criteria for the calibration period, December, 1995; the thick dashed lines correspond to 95% confidence limits obtained from MISE L_2 criterion (5); thin dashed lines show 95% confidence limits for 2 peak values of flow, L_2 norm, and dots denote the observed flow.

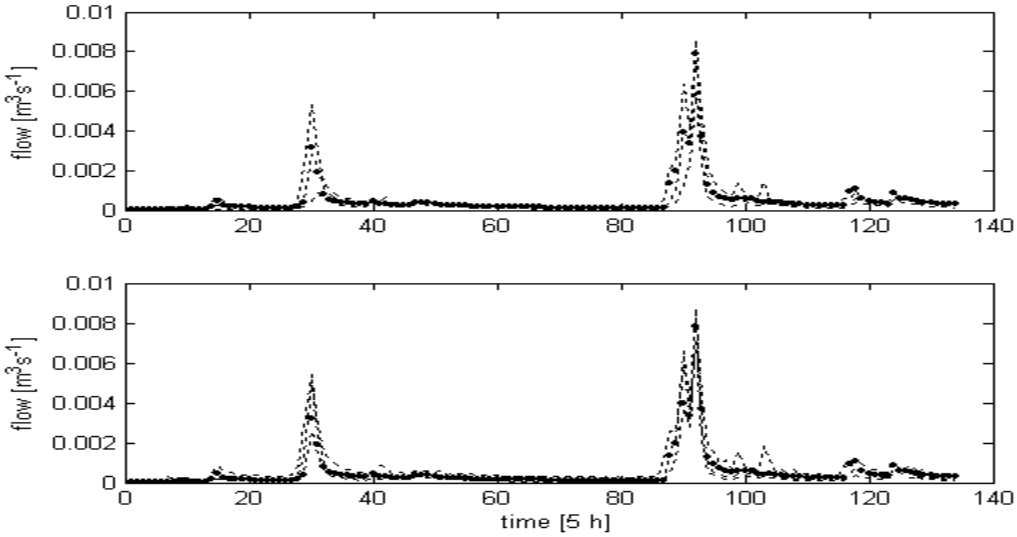


Figure 5. Validation stage: predicted flow (median) (--) and 95% confidence limits (:); big dots denote the observations; upper panel: weights derived from the sum of square errors; lower panel: weights derived from two peak values; January 1996. Note change in discharge scale from Fig. 1

Fig. 5 shows the application of the derived sampling scheme and likelihood weights to a new validation period in January 1996. Comparison of these results with the predictions for the same validation period obtained using formal approach (Fig. 1, lower panel) shows that the non-formal approach provides reasonably good predictions for high flows and is much better for low flows, where the formal approach overestimated the observed values.

6. CONCLUSIONS

Due to the stochastic nature of the variables influencing the deterministic model of the physical process at study, the predictions of the model output should also be considered to be stochastic. We may have some information about the probability distribution of different model variables, but usually environmental models are highly nonlinear and poorly defined, so it may be difficult, or even impossible to obtain the solution of the related inverse problem. This paper is meant as the bridge between formal and non-formal approaches to estimation of hydrological models. We presented a short discussion of statistical methods and their applicability to nonlinear, multidimensional and uncertain processes and pointed out that may be some justification for using a less formal approach such as GLUE.

Our results indicated that use of different criteria for evaluation of likelihood weights influences the shape of the resulting posterior distributions of the parameters but does not influence so much the uncertainty bands for the predictions. This is consistent with past experience reported, for example, in [2] and [17]. In order to obtain the control over the uncertainty limits we should use a suitable scaling parameter for the likelihood weights as well as a suitable choice of the observations for the conditioning of the probabilities for a particular application.

In future work we shall show the influence of the choice of time variable observation window on the uncertainty of model predictions and the way in which the results from different model structures can be integrated within the GLUE framework.

Acknowledgements

This study constitutes the preliminary modelling results in a collaboration with Francesc Gallart, Pilar Llorens, Jérôme Latron, of the Insitute of Earth Sciences Jaime Almera, Barcelona, Spain who are thanked for supplying the data.

REFERENCES

1. K. J. Beven and A. Binley. The future of distributed models: model calibration and uncertainty prediction, *Hydrological Processes.* 1992, 6: 279-298.
2. K. J. Beven, J. Freer, B. Hankin and K. Schulz. The use of generalised likelihood measures for uncertainty estimation in high order models of environmental systems. In *Nonlinear and Nonstationary Signal Processing*, W J Fitzgerald, R L Smith, A T Walden and P C Young (Eds). CUP, 115-151, 2000.
3. R. J. Romanowicz K. J. Beven and J. Tawn. Evaluation of Predictive Uncertainty in Nonlinear Hydrological Models Using a Bayesian Approach. In: *Statistics for the Environment 2, Water Related Issues* ,eds V. Barnett and K. F. Turkman, 297-315, Wiley, Chichester, 1994.
4. K. J. Beven and P. C. Young. Comment on Bayesian Recursive Parameter Estimation for Hydrologic Models by M Thiemann, M Trosset, H Gupta and S Sorooshian, *Water Resources Research,* 2003: 9(5), doi: 10.1029/2001WR001183.
5. G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, Chichester, 1992.
6. P. C. Young. Applying Parameter Estimation to Dynamic Systems. parts l, and 2 *Control Engineering*, 1969: 16, 10/11, 119-125/118-124.
7. R. E. Kalman. A new approach to linear filtering and prediction problems. ASME Transactions Journal Basic Engineering 1960: 83-D, 95-108.
8. A. Tarantola. *Inverse Problem Theory*. Elsevier; 1987
9. D. W. Scott. *Multivariate Density Estimation*. Wiley, New York, 1992.
10. A. B. Owen. *Empirical likelihood*. New York: Chapman&Hall/CRC; 2001.
11. R. J. Romanowicz, K. J. Beven and J. Tawn. Bayesian calibration of flood inundation models, in M G Anderson and D E Walling (Eds) *Floodplain Processes*, 333-360. Wiley, Chester, 1996.
12. F. Gallart, J. Latron , P. Llorens and D. Rabada. Hydrological functioning of Mediterranean mountain basins in Vallcebre, Catalonia: Some challenges for hydrological modelling, *Hydrological Processes*. 11 (9): 1263-1272, 1997.
13. F. Gallart, P. Llorens, J. Latron and D. Regues. Hydrological processes and their seasonal controls in a small Mediterranean mountain catchment in the Pyrenees, *Hydrology and Earth System Sciences*. 6 (3): 527-537, 2002.
14. K. J. Beven and M. J. Kirkby. A physically based variable contributing area model of basin hydrology, *Hydrological Sciences Bulletin*. 1979: 24, 1, 43-69.
15. R. J. Romanowicz. A MATLAB implementation of TOPMODEL, *Hydrological Processes*, 1997: 11, 1115-1129.
16. K. J. Beven. *Rainfall-runoff modelling: the primer*. Wiley, Chichester, 2001.
17. K. J. Beven and J. Freer. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems. *J. Hydrology*. 2001; 249: 11-29.

A Bayesian Analysis of Complex Dynamic Computer Models

Stefano Conti, Clive W. Anderson, Marc C. Kennedy and Anthony O'Hagan

Department of Probability and Statistics, The Hicks Building,
University of Sheffield, Sheffield S3 7RH, U.K.

Abstract: In many fields of science, sophisticated mathematical models are devised and implemented within large computer codes in order to simulate and predict complex real-world phenomena. These models are known for being exposed to various sources of uncertainty taking place at their building and validation steps, so that they are routinely subject to reliability tests by means of uncertainty and/or sensitivity analysis. Since such diagnostics typically require a large number of training code runs, for CPU-intensive models an approach based around preliminary *emulation* of a code's response, followed by application of the aforementioned techniques to the emulator, can be more practical and efficient. This paper extends results already established within a Bayesian set-up for deterministic models (see e.g. [1]) to dynamic multi-response computer codes, for which some of the outputs at one stage of a simulation become inputs to the subsequent stage. Advantages and difficulties in the implementation are here discussed, and a test-bed application to the Sheffield Dynamic Global Vegetation Model, developed within the UK Centre for Terrestrial Carbon Dynamics, is also presented.

Keywords: Bayesian inference, computer experiments, hierarchical models, sensitivity analysis, uncertainty analysis

1. INTRODUCTION

A nowadays widespread practice in the scientific community is the utilisation of large computer codes embedding sophisticated mathematical models descriptive of complex aspects of reality. The exploratory and predictive ability of any computer simulator is often hampered by substantial model preparation and computational requirements. Whereas computational burden is not remarkably cumbersome, nonetheless various uncertainties can still significantly compromise the performance of a computer model. Among recognised sources of uncertainty affecting the processes of model building and validation are (see [1] for a thorough discussion on the subject): **parameter uncertainty**, originating from unknown quantities tuning the code; **model inadequacy**, due to necessarily imperfect fit to the observed data; **residual variability**, related either to intrinsic randomness or unrecognised features of the real-world phenomenon; **parametric variability**, arising from quantities conveniently left unspecified; **observation error**, caused by inaccuracies at the hard data recording stage; and **code uncertainty**, related to the complex nature of the simulator.

Further author information: (Send correspondence to Stefano Conti, e-mail: s.conti@sheffield.ac.uk)

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

Several methodologies aimed at ascertaining the reliability and effectiveness of a given computer model are available off-the-shelf from the classical statistical literature: an exhaustive reference is provided by [2]. Unfortunately standard uncertainty/sensitivity analysis tools often require a large number of training code runs, hence proving unsuitable for validation of computationally expensive models. In order to circumvent this problem, a strategy based around preliminary emulation of the code's outcome (*meta-modelling*) suggested by [3] has been widely adopted. This procedure would typically be followed by application of the aforementioned techniques to the emulator, which in fact is treated like a cheaper alias of the original code.

In this context interesting results were obtained via a Bayesian semi-parametric representation of deterministic single-response codes, that is models returning the same scalar output when repeatedly fed with the same input configuration. In previous works in the field (refer e.g. to [4]) a Gaussian process prior for the code's output was shown to be a convenient, flexible and reasonable tool, especially for tackling the problems of model calibration and rectification.

Sometimes special features of the phenomena of interest may translate into aspects of the computer model that could complicate standard emulation. This is usually the case with dynamic computer models, typically designed for time-evolving processes. In particular, such codes have the distinctive feature that some of the inputs required at each stage of a simulation are actually outputs from previous stages. This affects the structure of the input space, in that variables required for the code's operation can be classified as: *constants*, which describe enduring characteristics of the examined events; *time-varying*, related to aspects of the process's evolution over time; and *intermediate* or *final outputs*, which in turn may or may not be recycled by the model into subsequent simulations. In this context meta-modelling of the code's outputs needs to be suitably adjusted in order to accommodate any relationship featured by variables evolving over time.

The paper discusses such adjustments and is organised as follows. In Section 2 a generalisation of the emulator as developed in [4] to multi-response codes is detailed. Section 3 is devoted to adapting the emulator to encompass dynamic computer models as well. An application to a sub-module of an environmental model is outlined in Section 4, while Section 5 summarises concluding remarks.

2. MODELLING MULTI-RESPONSE COMPUTER MODELS

Consider a deterministic computer model which takes inputs $\boldsymbol{x}$, typically lying in some (possibly high-dimensional) input space $\mathscr{X}$, and returns outputs $\boldsymbol{y}$. The process of computing vectors $\boldsymbol{y}$ from $\boldsymbol{x}$ can be formalised via a function $\boldsymbol{f}\colon \mathscr{X} \longmapsto \mathbb{R}^q$, where the input space is usually a subset of the Euclidean space $\mathbb{R}^p$ with typically $p \geq q$. In light of the usually high degree of complexity of the empirical processes being modelled via $\boldsymbol{f}(\cdot)$, it is normally unaffordable to explore the whole input space $\mathscr{X}$ just by repeatedly running the computer program. As a consequence, although in principle deterministic the code is in fact prone to at least code uncertainty (see Section 1). Hence $\boldsymbol{f}(\cdot)$ is regarded as a stochastic function on $\mathscr{X}$, though still assumed to be reasonably smooth. Therefore, within the Bayesian framework an appropriate prior process needs to be assigned to the

random function $\boldsymbol{f}(\cdot)$. In line with [4] we suppose that knowledge of the simulator can be summarised by the semi-parametric Gaussian process representation

$$\big[\boldsymbol{f}(\cdot) \mid B, \Sigma, R\big] \sim \mathscr{N}_q\big(\boldsymbol{m}(\cdot), c(\cdot,\cdot)\Sigma\big) \quad , \tag{1}$$

where $\forall\, \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathscr{X}$

$$\begin{aligned} \boldsymbol{m}(\cdot) &= B^{\mathrm{T}}\boldsymbol{h}(\cdot) \\ c(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \exp\big\{-(\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\mathrm{T}} R (\boldsymbol{x}_1 - \boldsymbol{x}_2)\big\} \end{aligned} \quad .$$

Here $\boldsymbol{h}\colon \mathscr{X} \longmapsto \mathbb{R}^m$ is a vector of arbitrary regression functions $\boldsymbol{h}(\boldsymbol{x})$, common to every component $f_i(\cdot),\ i = 1\ldots, q$ of $\boldsymbol{f}(\cdot)$, $B = [\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_q] \in \mathbb{R}_{m,q}$ is a matrix of regression coefficients, $\Sigma = [\sigma_{ij}] \in \mathbb{R}_{q,q}$ a dispersion matrix with generic entry $\sigma_{ij} = \mathbb{C}\mathrm{ov}\big[f_i(\cdot), f_j(\cdot)\big]$ and $R = \mathrm{diag}\{r_i\} \in \mathbb{R}^+_{q,q}$ a diagonal positive-definite roughness matrix. Gaussian processes constitute the natural counterpart of the Normal model usually invoked for finite-dimensional estimands, and enjoy the same flexibility and tractability when utilised for addressing problems related to functional inference. Separability of the covariance structure between the code inputs and its outputs is here assumed for simplicity. It is also worthwhile mentioning that diagonality of R, here imposed for parsimony, implies that the correlation structure between any pair $f(\boldsymbol{x}_1)$ and $f(\boldsymbol{x}_2)$ is insensitive to any interaction among inputs.

Running the computer code on a pre-selected *design set* $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \subset \mathscr{X}$ yields simulations organised in the data matrix $D = \big[f_i(\boldsymbol{s}_r)\big] \in \mathbb{R}_{n,q}$. The design set can be selected in accord to some space-filling experimental design criterion: see for instance [5, 6, 7] and annotated bibliography. Due to the learning mechanism intrinsic to the Bayesian paradigm, as more model runs become available the posterior distribution of $\boldsymbol{f}(\cdot)$ becomes more concentrated near the input configurations, which in turn are exactly interpolated.

In light of the assumptions listed above, the joint distribution of the code responses D conditional on nuisance parameters B, Σ, R is the matrix-Normal distribution

$$[D \mid B, \Sigma, R] \sim \mathscr{N}_{n,q}\big(HB,\ \Sigma \otimes A\big) \quad ,$$

where $H^{\mathrm{T}} = \big[\boldsymbol{h}(\boldsymbol{s}_1) \cdots \boldsymbol{h}(\boldsymbol{s}_n)\big] \in \mathbb{R}_{m,n}$, $A = [c(\boldsymbol{s}_r, \boldsymbol{s}_l)] \in \mathbb{R}_{n,n}$ and $\otimes$ denotes the Kronecker product. Letting now $\boldsymbol{t}^{\mathrm{T}}(\cdot) = \big[c(\cdot, \boldsymbol{s}_1) \cdots c(\cdot, \boldsymbol{s}_n)\big] \in \mathbb{R}^n$, standard Normal theory and some matrix calculus manipulations lead to the following conditional posterior distribution for the computer simulator:

$$\big[\boldsymbol{f}(\cdot) \mid B, \Sigma, R, D\big] \sim \mathscr{N}_q\big(\boldsymbol{m}^\star(\cdot), c^\star(\cdot,\cdot)\Sigma\big) \quad , \tag{2}$$

where

$$\begin{aligned} \boldsymbol{m}^\star(\cdot) &= B^{\mathrm{T}}\big[\boldsymbol{h}(\cdot) - H^{\mathrm{T}}A^{-1}\boldsymbol{t}(\cdot)\big] + D^{\mathrm{T}}A^{-1}\boldsymbol{t}(\cdot) \\ c^\star(\boldsymbol{z}_1, \boldsymbol{z}_2) &= c(\boldsymbol{z}_1, \boldsymbol{z}_2) - \boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_1)A^{-1}\boldsymbol{t}(\boldsymbol{z}_2) \end{aligned} \quad .$$

A possible way to obtain the posterior process of $\boldsymbol{f}(\cdot)$ conditional on the roughness matrix R alone is by integration of (2) with respect to the posterior distribution of the

nuisance parameters B, Σ. Since any substantial information about such parameters will hardly ever be elicited from the code developers, a prior choice found to be both reasonable and manageable is the Jeffreys non-informative independence distribution $\pi^{\mathrm{J}}(B, \Sigma \mid R) \propto |\Sigma|^{-\frac{q+1}{2}}$. Standard Bayesian calculations from (2) and $\pi^{\mathrm{J}}(\cdot)$ yield

$$[\boldsymbol{f}(\cdot) \mid \Sigma, R, D] \sim \mathscr{N}_q\big(\boldsymbol{m}^{\star\star}(\cdot), c^{\star\star}(\cdot,\cdot)\Sigma\big) \quad , \tag{3}$$

where

$$\boldsymbol{m}^{\star\star}(\cdot) = D^{\mathrm{T}}A^{-1}\boldsymbol{t}(\cdot) + \hat{B}^{\mathrm{T}}_{\mathrm{GLS}}\big[\boldsymbol{h}(\cdot) - H^{\mathrm{T}}A^{-1}\boldsymbol{t}(\cdot)\big] \tag{4a}$$

$$\begin{aligned} c^{\star\star}(\boldsymbol{x}_1, \boldsymbol{x}_2) &= c^{\star}(\boldsymbol{x}_1, \boldsymbol{x}_2) \\ &\quad + \big[\boldsymbol{h}(\boldsymbol{x}_1) - H^{\mathrm{T}}A^{-1}\boldsymbol{t}(\boldsymbol{x}_1)\big]^{\mathrm{T}}\big(H^{\mathrm{T}}A^{-1}H\big)^{-1}\big[\boldsymbol{h}(\boldsymbol{x}_2) - H^{\mathrm{T}}A^{-1}\boldsymbol{t}(\boldsymbol{x}_2)\big] \end{aligned} \tag{4b}$$

and $\hat{B}_{\mathrm{GLS}} = \big(H^{\mathrm{T}}A^{-1}H\big)^{-1}H^{\mathrm{T}}A^{-1}D$ is the GLS estimator of B. Provided that $n \geq m + q$ so that all ensuing posteriors are proper, the conditional posterior Student's $\mathscr{T}$ process

$$[\boldsymbol{f}(\cdot) \mid R, D] \sim \mathscr{T}_q\big(\boldsymbol{m}^{\star\star}(\cdot), c^{\star\star}(\cdot,\cdot)\hat{\Sigma}_{\mathrm{GLS}}; n - m\big) \tag{5}$$

is finally obtained, in which $\hat{\Sigma}_{\mathrm{GLS}} = (n-m)^{-1}(D - H\hat{B}_{\mathrm{GLS}})^{\mathrm{T}}A^{-1}(D - H\hat{B}_{\mathrm{GLS}})$ denotes the GLS estimator of Σ.

Direct utilisation of (5) and of precursory results for drawing inferences about the simulator $\boldsymbol{f}(\cdot)$ must still be preceded by estimation of the unknown roughness matrix R. A full Bayesian treatment of the roughness parameters, notoriously difficult to estimate (see in particular [8, 9] for insights), is here rejected on the grounds of computational tractability. Nonetheless a plug-in approach based upon the posterior mode of $(r_1, \ldots, r_p)$ arising from a diffuse, albeit proper, prior was found to yield satisfactory results. Once this task is accomplished expressions (4a)-(4b) furnish respectively a cheap code interpolator and, when combined with $\hat{\Sigma}_{\mathrm{GLS}}$, a measure of its accuracy. Furthermore, integration of the posterior $\mathscr{T}$ process (5) relative to appropriately selected distributions over (possibly portions of) the input space $\mathscr{X}$ constitutes the basis of customary uncertainty analysis techniques. For a complete review of such methodologies, and their adaptations and implications within a Bayesian set-up, see [2, 4, 10].

3. DYNAMIC META-MODELS

3.1. From Static to Dynamic Emulation

Dynamic computer models come into play when it is desired to reproduce and examine the evolutionary nature of a time-varying process. As mentioned in Section 1, in order to reproduce dynamic patterns computer models customarily utilise outputs from each stage of a simulation as inputs to subsequent stages. This is in essence achieved by computing the state vector $\boldsymbol{y}_t$ relative to a time step $t = 1, \ldots$ from inputs comprising both constant tuning values $\boldsymbol{x}$ and outputs $\boldsymbol{y}_{t-1}$ from the previous time period $t - 1$*. Taking into

*For many physical processes to impose a Markovian dependence of dynamic outputs over time, although not correct, may still produce an acceptable representation.

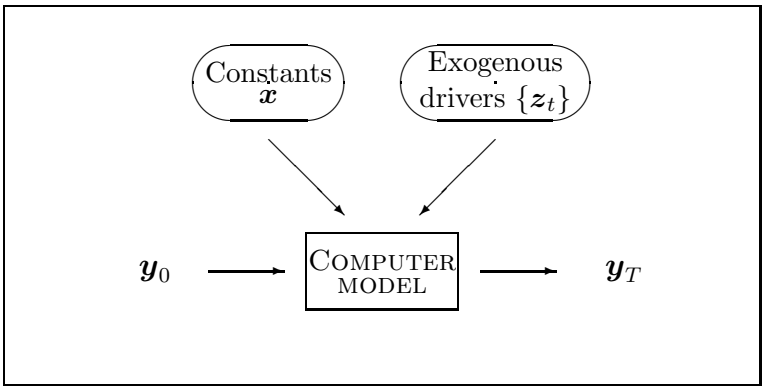


Figure 1. Single-step meta-modelling

account other time-evolving exogenous code drivers $\{\boldsymbol{z}_t\}$, the model's functioning can be then represented via the recursive relation

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z}_t, \boldsymbol{y}_{t-1}) \\ &= \boldsymbol{f}\big[\boldsymbol{x}, \boldsymbol{z}_t, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z}_{t-1}, \boldsymbol{y}_{t-2})\big] \\ &= \cdots \end{aligned} \quad .$$

If the time span of interest is delimited by endpoints $t = 0$ and $t = T$, then in principle emulation of $\boldsymbol{f}(\cdot)$ can be attained over such interval just in a single-run fashion (see Figure 1): under this perspective the simulator is imagined to take a set of input values, comprising initial system descriptors $\boldsymbol{y}_0$, and to return a collection of outputs inclusive of $\boldsymbol{y}_T$. The main appeal of such procedure clearly lies in enabling straightforward application of standard statistical analysis tools already existing for static codes. An important disadvantage however is that in this set-up the input space $\mathscr{X}$ comprises constants $\boldsymbol{x}$, the

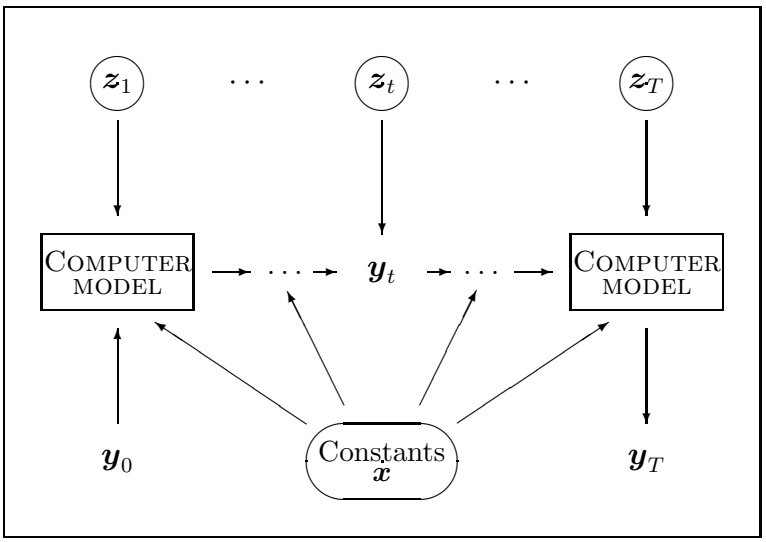


Figure 2. Recursive meta-modelling

initial state vector $\boldsymbol{y}_0$ and all drivers $\{\boldsymbol{z}_t\}$. On the other hand the output generated by $\boldsymbol{f}(\cdot)$ consists of the whole collection of runs $\{\boldsymbol{y}_t\}$. Therefore applications of this strategy is severely hindered by typically unwieldy dimensions for both the model input and output spaces. Additional relevant flaws undermining single-step emulation can be also recognised in the need to rebuild the posterior process (5) from scratch whenever the temporal window of interest changes and in deferment of any occasional model rectification exercise until the end of the simulation.

An approach which intuitively retains the code's evolutionary pattern is alternatively depicted in Figure 2. Here the proposed idea is to run a single-step emulator in a recursive fashion, until coverage of the time span $(0, T)$ is attained: given the state vector $\boldsymbol{y}_t$ acquired at any time $t = 0, \ldots, (T-1)$, subsequent emulation of $\boldsymbol{y}_{t+1}$ is then computed only on the basis of $\boldsymbol{y}_t$ and of current drivers $\boldsymbol{z}_{t+1}$. As a consequence input and output spaces become more tractable and interactive data assimilation is now feasible. Main drawbacks of the method in this case are: need to extend theoretical results from static to dynamic codes; tighter accuracy requirements at each time step for ensuring adequate overall meta-modelling; and gradual fading of computational advantages over direct Monte Carlo simulation for increasing values of T.

3.2. Dynamic Emulation Theory

Assume for simplicity that no exogenous drivers are required for running $\boldsymbol{f}(\cdot)$. In addition, suppose with no loss of generality that the variables which the code rolls forward over time are the last $q \leq p$ input entries; that is, for any t and any suitable $\boldsymbol{x} \in \mathbb{R}^{p-q}$, $\boldsymbol{y}_t \in \mathbb{R}^q$ we have $\boldsymbol{y}_{t+1} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}_t)$.

An issue that should promptly be highlighted is that under assumption (1) a recursive emulator of nonlinear codes can at best be approximately Gaussian: if $\boldsymbol{y}_t = \boldsymbol{f}(\bar{\boldsymbol{x}})$ for some $\bar{\boldsymbol{x}} \in \mathscr{X}$, then for any $\boldsymbol{x}$ such that $(\boldsymbol{x}, \boldsymbol{y}_t) \in \mathscr{X}$ the statement

$$\big[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{Y}_t) \mid \Sigma, R\big] \sim \mathscr{N}_q(\cdot, \cdot)$$

will not strictly hold[†]. Keeping an assumption of approximate Normality for modelled outputs on practical grounds obviously entails careful assessment of its plausibility at each stage of the whole emulation process. Subject to this condition it then becomes feasible to explore the first and second order properties of the posterior distribution of $\boldsymbol{f}(\boldsymbol{y}_t)$, given previous-step outputs $\boldsymbol{y}_{t-1}$ and nuisance parameters Σ, R. In fact, recalling results (4) approximate expressions have been derived in closed form for

$$\mathbb{E}\big[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{Y}_t) \mid \Sigma, R, D\big] = \mathbb{E}\big[\boldsymbol{m}^{\star\star}(\boldsymbol{x}, \boldsymbol{Y}_t) \mid \Sigma, R, D\big] \tag{6a}$$

$$\begin{aligned} \mathbb{C}\mathrm{ov}\big[\boldsymbol{f}(\boldsymbol{x}_1, \boldsymbol{Y}_{t_1}), &\boldsymbol{f}(\boldsymbol{x}_2, \boldsymbol{Y}_{t_2}) \mid \Sigma, R, D\big] \\ &= \mathbb{C}\mathrm{ov}\big[\boldsymbol{m}^{\star\star}(\boldsymbol{x}_1, \boldsymbol{Y}_{t_1}), \boldsymbol{m}^{\star\star}(\boldsymbol{x}_2, \boldsymbol{Y}_{t_2}) \mid \Sigma, R, D\big] \\ &\quad + \mathbb{E}\Big\{c^{\star\star}\big[(\boldsymbol{x}_1, \boldsymbol{Y}_{t_1}), (\boldsymbol{x}_2, \boldsymbol{Y}_{t_2})\big] \mid \Sigma, R, D\Big\}\Sigma \end{aligned} \tag{6b}$$

by applying the law of iterated expectations and relying upon properties of multi-Normal distributions. After integrating out the unknown parameters in Σ and R, expressions (6)

[†]An intuitive counterexample is provided by the case $f\colon \mathbb{R} \longmapsto \mathbb{R}^+,\ f(x) = x^2$.

in turn play an analogous role to their "static" counterparts (4) when attempting usual uncertainty or sensitivity analyses of $\boldsymbol{f}(\cdot)$ in a dynamic fashion. As regards the $(q\frac{q+1}{2})$-dimensional problem of marginalising Σ, in light of the dual origin of the Student's $\mathscr{T}_q$ distribution (see for instance [11], pg. 23) this was found to be efficiently dealt with by means of a simple univariate integration.

4. THE SHEFFIELD DYNAMIC GLOBAL VEGETATION MODEL: A TEST-BED

The Centre for Terrestrial Carbon Dynamics (*CTCD*) is a consortium of British academic and governmental institutions, established for the purpose of progressing scientific understanding of the role played by terrestrial ecosystems in the carbon cycle, with particular emphasis on forest ecosystems. The ultimate goals of the project are: to gauge carbon fluxes and their uncertainties at different space/time resolutions; to devise methodological, data and instrument advances for reducing these uncertainties; to deliver relevant findings in accessible formats to the scientific community and ultimately to policy makers. These tasks are pursued with the support of a variety of environmental models designed for simulating carbon patterns over different geographical and climatic scenarios. Unfortunately, such models suffer from coarse reproduction of some underlying physical processes and loose connections to driving data.

Bayesian statistical methods are being employed within the Centre for the assessment of relevant model (and data) developments required for reducing the uncertainty around them. In this setting, statistical challenges other than pure uncertainty and sensitivity analysis which presently require special care are: prediction, i.e. estimation of (possibly functionals of) model outputs at unavailable input configurations; screening, that is identification of which code inputs exert most significant influence on the outputs; and code verification, or detection of bugs in the actual implementation of the program.

Among the simulators devised and deployed within CTCD a central role is played by the Sheffield Dynamic Global Vegetation Model, daily version (henceforth *SDGVMd*). SDGVMd is aimed at illustrating possible responses of ecosystem processes to atmospheric CO_2 concentration and climate changes by modelling interactions at a regional to global scale between ecosystem carbon, water fluxes and vegetation. Inputs to SDGVMd comprise broad soil, vegetation and climate descriptors; outputs of the model include various measures of a site's carbon budget and miscellaneous environmental quantities. Additional challenges specifically offered by SDGVMd comprise a high-dimensional input space and the existence of embedded sub-modules operating at different time scales. A complete description of SDGVMd and the modules it incorporates can be found in [12].

For the purpose of illustrating the broad range of possibilities offered by Gaussian process-based meta-modelling, the soil module of SDGVMd (*CENTURY*: for details refer to [13]) was extracted and subject to emulation. In essence CENTURY manages the soil carbon (C) calculations within SDGVMd by recursively solving a set of independent PDEs, each being indexed by temperature, relative humidity and precipitation drivers and describing the monthly evolution of 8 different C pools. It should be stressed that

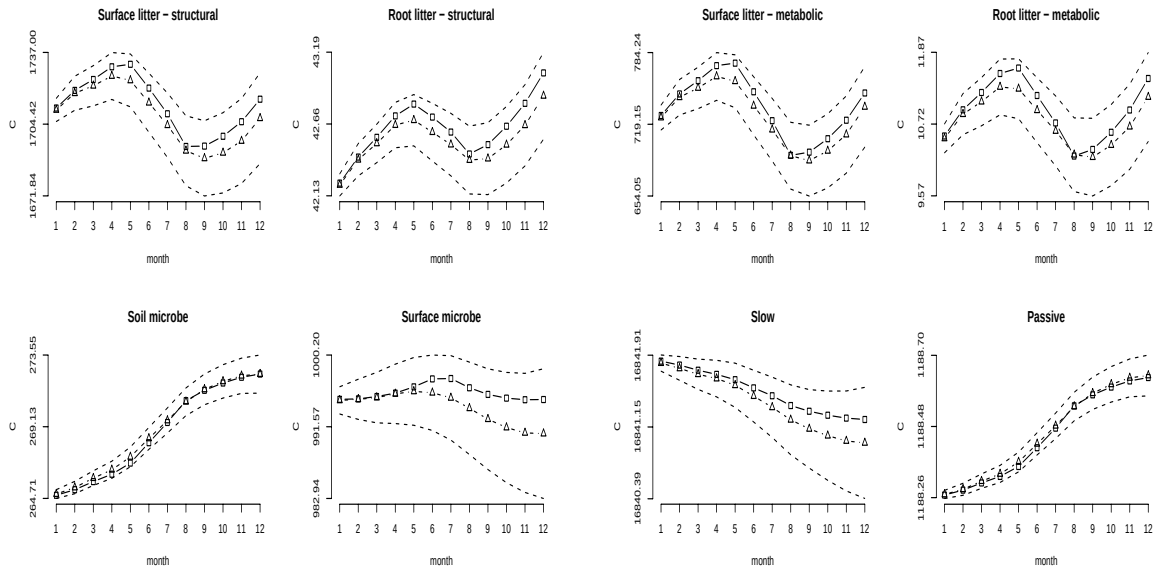


Figure 3. Simulated (—□—) vs. emulated ($\cdots\triangle\cdots$) C pools within 95% credible bounds ($---$).

in fact CENTURY is not a CPU-intensive model; nonetheless this feature conveniently enables straightforward evaluation of its emulator's performance.

Having thus recognised as an appropriate representation of CENTURY a function $\boldsymbol{f}: \mathbb{R}^{11} \longmapsto \mathbb{R}^{8}$, an interpolation exercise over a time period of 12 months was subsequently carried out. The input space was covered via a maximin Latin hypercube design of size $n = 200$; thereafter roughness parameters were estimated by the joint posterior mode based on vague i.i.d. Log-Logistic priors on $(r_1, \ldots, r_p)$. The prior mean was chosen to be linear, i.e. $\boldsymbol{h}(\boldsymbol{x}) = (1, x_1, \ldots, x_p)$, again for convenience. Figure 3 compares CENTURY's exact simulations for each C pool with their corresponding approximate posterior values from (6a), embedded within approximate 95% credible bounds.

A few comments are in order. It should be noticed how in most cases estimated interpolators appear to satisfactorily capture the underlying original outputs, apart from perhaps a couple of C pools (Surface microbe and Slow) where some drift can be observed to emerge over time. Additional effort should be placed into achieving somewhat narrower credible bounds for the interpolators, but provisional results look overall encouraging.

5. CONCLUSIONS

The paper focuses on two main goals. First, it deals with extensions to multi-response computer simulators of theoretical results already established for Bayesian meta-modelling of single-response codes. Second, it attempts to adapt the general methodology to encompass dynamic computer models within the same formal framework. Single-step and recursive emulation schemes were introduced and contrasted; preference towards the latter methodology was based on computational manageability. The proposed statistical

machinery was then tested on the soil compartmental model embedded within SDGVMd: although the implemented exercise was limited to pure code interpolation, the analysis's outcome confirmed that interesting insights can be gained from applying the principle of Bayesian Gaussian process-based emulation to more sophisticated settings.

ACKNOWLEDGMENTS

The authors wish to thank Caitlin Buck for helpful suggestions. This research is financed by the UK Natural Environment Research Council, contract number F14/G6/105.

References

[1] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(3):425–464, 2001.

[2] A. Saltelli, K. Chan, and E. M. Scott, editors. *Sensitivity Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2000.

[3] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statist. Sci.*, 4(4):409–435, 1989. With comments and a rejoinder by the authors.

[4] A. O'Hagan, M. C. Kennedy, and J. E. Oakley. Uncertainty analysis and other inference tools for complex computer codes. In *Bayesian statistics, 6 (Alcoceber, 1998)*, pages 503–524. Oxford Univ. Press, New York, 1999.

[5] M. E. Johnson, L. M. Moore, and D. Ylvisaker. Minimax and maximin designs. *Journal of Statistical Planning and Inference*, 26:131–148, 1990.

[6] J. R. Koehler and A. B. Owen. Computer experiments. In *Design and Analysis of Experiments*, volume 13 of *Handbook of Statist.*, pages 261–308. North-Holland, Amsterdam, 1996.

[7] J. A. Royle and D. Nychka. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Technical report, Department of Statistics, North Carolina State University, 1997.

[8] M. Schlather. Introduction to positive definite functions and to unconditional simulation of random fiels. Technical report, Department of Mathematics and Statistics, FAculty of Applied Sciences, Lancaster University, 1997.

[9] J. O. Berger, V. De Oliveira, and B. Sansó. Objective Bayesian analysis of spatially correlated data. *J. Amer. Statist. Assoc.*, 96(456):1361–1374, 2001.

[10] J. E. Oakley and A. O'Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4):769–784, 2002.

[11] D. Gamerman, editor. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall. Chapman & Hall, London, 1997.

[12] M. R. Lomas, F. I. Woodward, and S. Quegan. The role of dynamic vegetation models. Technical report, Department of Animal and Plant Sciences, University of Sheffield, Sheffield UK, 2002.

[13] W. J. Parton, J. M. O. Scurlock, D.S. Ojima, T. G. Gilmanov, R. G. Scholes, D. S. Schimel, T. Kirchner, J.-C. Menaut, T. Seastedt, E. Garcia Moya, A. Kamnalrut, and J. I. Kinyamario. Observations and modelling of biomass and soil organic matter dynamics for the grassland biome worldwide. *Global Biogeochemical Cycles*, 7:785–809, 1993.

Validation and Error Estimation of Computational Models

Ramesh Rebba, Sankaran Mahadevan and Shuping Huang

Department of Civil and Environmental Engineering,

Vanderbilt University, Box 1831-B, Nashville, 37235, USA

E-mail: rebba.ramesh@vanderbilt.edu, sankaran.mahadevan@vanderbilt.edu, shuping.huang@vanderbilt.edu

Abstract

This paper develops a Bayesian methodology for assessing the confidence in model prediction by comparing the model output with experimental data when both are stochastic. The prior distribution of the response is first computed, which is then updated based on experimental observation using Bayesian analysis to compute a validation metric. A model error estimation methodology is then developed to include model form error, discretization error, stochastic analysis error (UQ error), input data error and output measurement error. Sensitivity of the validation metric to various error components and model parameters in discussed. A numerical example is presented to illustrate the proposed methodology.

Keywords: Bayesian statistics, error estimation, sensitivity, uncertainty, validation, verification.

1. Introduction

Complex natural phenomena are increasingly sought to be modeled through sophisticated computational models, with very few or no full-scale experiments, thus reducing the time and cost of engineering development relying upon the understanding of these phenomena. However, such models incorporate many assumptions and approximations and hence need to be subjected to rigorous, quantitative verification and validation (V & V) before they can be applied to practical problems with confidence.

There are a number of physical, statistical and model uncertainties in the prediction apart from the various direct sources of numerical error. A probabilistic approach to V&V under uncertainty involves quantification of the statistical distribution of model prediction and then comparing it with experimental measurement that also follows a statistical distribution. Note that this could also be viewed as studying the joint distribution of the experiment and model. Various methods are available to carry out probabilistic analysis to quantify the uncertainty in the model output, given the statistical distributions of the input variables, such as Monte Carlo simulation [1] or response surface methods [2, 3]. The choice of method depends on the nature of model used for predicting the output, and the needs with respect to accuracy and efficiency.

Verification refers to the assessment of accuracy of the solution with respect to known solutions, or by some other means, such as a posteriori error estimation. This activity helps to identify, quantify and reduce the errors in the computational model [4, 5]. Several finite element discretization error estimators have been developed in the literature [6-8]. Error estimates for uncertainty quantification methods (Monte Carlo and response surface methods) are also available [2, 5].

Validation involves comparison of model prediction with experimental data [4]. The widely used method of "graphical validation" or viewgraph-based judgment (i.e., by plotting graphs of

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

prediction and observation) is inadequate although it is better than a qualitative comparison [9, 10]. A rigorous quantitative model validation metric should include both prediction and measurement errors, and other uncertainties. Several metrics have been developed to include parametric uncertainty [11]. One such metric normalizes the difference between model predictions and experimental values and computes a relative error norm for discrete and continuous domain problems. Another metric includes the uncertainty in the experimental value due to limited data through statistical distributions and classical hypothesis testing [12, 13].

Two types of validation metrics are developed in this paper, based on the Bayesian approach. The first metric considers test data based on a simple fail/pass criterion, while the second metric compares model prediction with observed response measurement, both being continuous variables. The second approach explicitly incorporates the variability in the experimental data and the magnitude of its deviation from the model prediction. Once the model is validated, it may be calibrated to improve its predictive capability. A prediction error estimation methodology is developed for this purpose in this paper; this includes model form error, discretization error, stochastic analysis error (UQ error), input data error and output measurement error. The overall error is a nonlinear combination of these various errors. Sensitivity analysis of the validation metric to different physical and statistical parameters of the model output and measurement error variance can be very useful for model improvement or calibration and resource allocation. Section 2 develops the Bayesian validation metrics, and Section 3 describes the proposed methodology for model error estimation and sensitivity analysis of the validation metric. An illustrative numerical example is provided in Section 4.

2. Validation metric

2.1. Bayes factor

Consider two models (or hypotheses) M_i and M_j. Their prior probabilities of acceptance are denoted by $P(M_i)$ and $P(M_j)$. By Bayes' rule, when an event/data is observed, the relative posterior probabilities of two hypotheses are obtained as [14, 15]:

$$\frac{P\left(M_i \middle| \text{observation}\right)}{P\left(M_j \middle| \text{observation}\right)} = \left[\frac{P\left(\text{observation} \middle| M_i\right)}{P\left(\text{observation} \middle| M_j\right)}\right]\left[\frac{P\left(M_i\right)}{P\left(M_j\right)}\right] \tag{1}$$

The term in the first set of square brackets on the right hand side is called the "Bayes factor" [16]. If the Bayes factor is greater than 1.0 then it can be inferred that the data favors the model M_i more than model M_j. If only a single model M is proposed, then the model could be either accepted as correct or rejected as incorrect. Thus the Bayes factor in Eq. (1) may also be written as $P\left(\text{observation} \middle| M \text{ is correct}\right) / P\left(\text{observation} \middle| M \text{ is not correct}\right)$. When an observation is made, then the Bayes factor estimates the ratio of relative likelihoods of the null hypothesis (i.e., data supports the proposed model) and alternate hypothesis (i.e., data does not support the proposed model). The Bayes factor metric is further developed below for two situations: 1) reliability model 2) response computation.

2.1.1. Validation with pass/fail test data

Let x_o and x be the predicted failure probability and true failure probability respectively of an engineering system. The value x_o is predicted by model M. This can be considered as a point null

hypothesis (H_o: $x = x_o$). To estimate the Bayes factor in Eq. (1), we need to consider an alternative hypothesis (H_1: $x \neq x_o$).

If n experiments are undertaken, and k failures (e.g., stress greater than an allowable value) are observed out of n tests, then the probability of observing the data given that the true probability is equal to x comes from a binomial distribution as

$$P(k \mid x, n) = {}^{n}C_{k} x^{k} (1-x)^{n-k} \tag{2}$$

Under the null hypothesis, this probability, $P(data \mid H_o: x = x_o)$ can be exactly estimated by simply substituting x_o in Eq. (2). Assume that there is no prior information about x under the alternative hypothesis. Therefore, a uniform distribution in [0, 1] is assumed for $f(x \mid H_1)$, the prior density under the alternative hypothesis [17]. Then the Bayes factor is computed as

$$B(x_o) = \frac{P(data \mid H_0 : x = x_o)}{P(data \mid H_1 : x \neq x_o)} = \frac{C_k^n x_0^k (1-x_0)^{n-k}}{\int_0^1 C_k^n x^k (1-x)^{n-k} \, f(x \mid H_1) \, dx} = (n+1) C_k^n x_0^k (1-x_0)^{n-k} \tag{3}$$

It is easy to identify the above expression in Eq. (3) to be the probability density function (PDF) of a beta distribution with parameters $k + 1$, $n - k + 1$. It is well known that the posterior PDF of x follows the beta distribution, when the prior PDF has uniform distribution. In more detail, if the prior has a uniform PDF in [0, 1] i.e., $f(x) = 1$, then the posterior PDF is

$$f(x \mid data : n, k) = (n+1) C_k^n x^k (1-x)^{n-k} \tag{4}$$

Note that this result is the same as in Eq. (3), which is the Bayes factor $B(x)$ evaluated at the probability x (see Fig. 1). Therefore, the Bayes factor can be viewed here as the posterior density of x evaluated at the predicted value x_o.

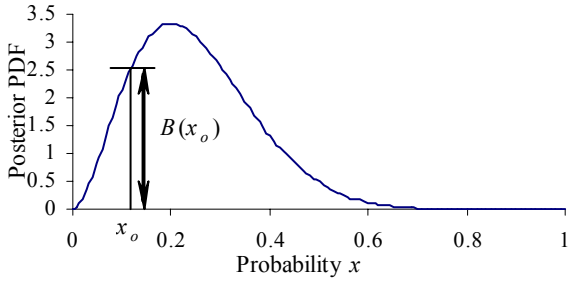


Fig. 1. Posterior density function for the probability of failure

The above result, that the Bayes factor is the posterior PDF at the predicted value x_0, is only for the case with uniform prior and binomial pass/fail data. For tests conducted in other situations, only a response quantity may be measured (such as deflection, strain etc.), but the specimen may not be loaded till failure. In such cases, it is valuable to derive a more general expression for the Bayes factor, by using prior and posterior PDF's of the predicted response.

2.1.2. Validation with response variable measurement

In Eq. (2), the probability of the data k for a given value of x, i.e., $P(k \mid x, n)$, is also the likelihood function of x, i.e., $L(x)$, where the failure probability x is the parameter of the binomial distribution. For a continuous distribution, the likelihood function is proportional to the

probability density of data y given the parameter x i.e., $L(x) \propto f(y|\, x)$ [18]. Consider x to be not the failure probability, but some general response quantity, with density function $f(x)$ and x_o the value predicted by the computational model under consideration. Then the probability of observing the data under the null hypothesis, $P(data|\, H_o\colon x = x_o)$ can be obtained from $L(x_o) = \varepsilon f(y|\, x_o)$ where ε is a positive constant [18]. Similarly, the probability of observing the data under the alternative hypothesis $P(data|\, H_1\colon x \neq x_o)$ can be obtained from $\int L(x)g(x)dx$ or $\int \varepsilon f(y\,|\,x)g(x)dx$, where $g(x)$ is the prior density of x under the alternative hypothesis. Since no information on $g(x)$ is likely to be available, one possibility is to assume $g(x) = f(x)$. Then, using Eq. (1) and Bayes theorem, the Bayes factor is computed as

$$B(x_o) = \frac{P\left(data\,\middle|\, H_0 : x = x_o\right)}{P\left(data\,\middle|\, H_1 : x \neq x_o\right)} = \frac{L(x_o)}{\int L(x) f(x)dx} = \frac{f(y\,|\,x_o)}{\int f(y\,|\,x) f(x)dx} = \left.\frac{f(x\,|\,y)}{f(x)}\right|_{x=x_0} \tag{5}$$

Thus, the Bayes factor simply becomes the ratio of posterior to prior PDFs of the predicted response when $g(x) = f(x)$. This result probabilistically quantifies the contribution to model validation of an experimental result that agrees with a given model prediction. If $g(x) \neq f(x)$, then the Bayes factor is computed using Eq. (5) with $g(x)$ instead of $f(x)$ in the denominator. Fig. 2 shows notional posterior and prior densities of model prediction x. Once again, $B > 1$ indicates data support for the model.

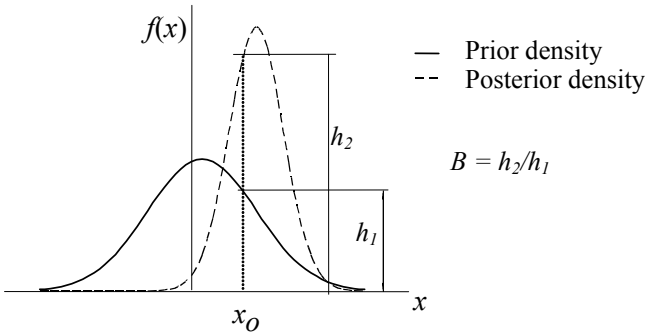


Fig. 2. Validation metric as a ratio of posterior and prior density values

If x_{true} is the true solution, x is the model output, and y is the experimental observation, then the following equations hold:

$$x_{true} = x + \varepsilon_{pred} \tag{6a}$$

$$x_{true} = y + \varepsilon_{exp} \tag{6b}$$

where ε_{pred} is the model prediction error and ε_{exp} is the measurement error. If we hypothesize that there is no prediction error, the observed value will simply be $y = x - \varepsilon_{exp}$. From this relation and a Gaussian experimental error assumption, we obtain $f(y|\, x) \sim N(x,\ \sigma^2_{\varepsilon_{\exp}})$. The likelihood function $L(x)$ in Eq. (5) can be created using $f(y|\, x)$. If there is only one observed value of y, then $L(x) = f(y|\, x)$. If multiple data are observed, the likelihood is constructed as a product of $f(y|\, x)$ values evaluated at each y. This can be used to test our hypothesis.

A Bayes factor less than unity denotes that ε_{pred} is significant and should not have been omitted and hence there is a need for estimating the total prediction error. It should be noted that the metric shown in Eq. (5) allows us to use non-Gaussian experimental errors also. Even when

$B > 1$, it is useful to quantify the prediction error, and to examine whether contributions from different errors cancel each other. Further the degree of confidence in the model prediction can be measured from the posterior probability of the null hypothesis being true i.e., $P(H_0|\ data)$ as $B/(B + 1)$ assuming that the prior probability $P(H_0)$ to be 0.5 in the absence of any prior knowledge. The following section presents methods for quantifying the errors and uncertainty in model prediction.

3. Error estimation

The total prediction error is a function of various error components that can be broadly divided into numerical solution errors and model form errors. Investigations on error combination are rarely available. This paper pursues a nonlinear combination method.

3.1 Numerical Error Components in Simulation

Several components of numerical errors in model prediction, such as data error, discretization error, stochastic analysis error (or UQ error), and measurement error are briefly discussed below.

3.1.1. Input data error (ε_d)

The measurement error in the input variables will be propagated to the prediction of the output. If the relationship between input and output is given by $u = f(x_1, x_2, ..x_m)$, then the error in the prediction of the output due to the measurement error in the input variables may be approximated using a first-order sensitivity analysis as

$$\varepsilon_d = \Delta u = \sum_{i=1}^{m} \left(\frac{\partial f}{\partial x_i} \right|_{x = \bar{x}} \delta x_i \tag{7}$$

in which δx_i is the measurement error in i^{th} input random variable x_i and $\left(\frac{\partial f}{\partial x_i} \right|_{x=\bar{x}}$ is the first order sensitivity coefficient of the model output u with respect to the i^{th} input random variable x_i. The measurement error in each input variable has been commonly quantified as a random variable with Gaussian distribution, with zero mean and a known or assumed variance, based on the instrument calibration.

3.1.2. Discretization error (ε_h)

Several methods to quantify the discretization error in finite element analysis are available in the literature. However, most of these methods do not quantify the actual error; instead, they quantify some surrogate measures to facilitate adaptive mesh refinement. The Richardson extrapolation (RE) method has been found to be suitable for model verification and validation, since it comes closest to quantifying the actual discretization error [19]. This method has been extended by the first author to stochastic finite element analysis [20]. It should be noted that RE requires that the model solution be asymptotically convergent and the domain is discretized uniformly (regular grid). The assumption of monotone truncation error convergence is not valid for very coarse models sometimes. In the Richardson extrapolation method, the error due to grid size (for a coarse mesh) is given by

$$\varepsilon_h = \frac{f_1 - f_2}{r^p - 1} \tag{8}$$

where the grid refinement ratio $r = h_2/ h_1$, and f_1 and f_2 are the solutions with coarse and fine meshes respectively. The order of convergence p can be obtained from the relation $p = \ln\left(\frac{f_3 - f_2}{f_2 - f_1}\right) / \ln(r)$ where f_3 is the solution with the finest grid size, and $r = h_2/ h_1 = h_3/ h_2$. Due to the input variable uncertainties, the finite element solutions f_1 and f_2 are stochastic. Therefore, based on Eq. (8), the discretization error (ε_h) is also a random variable.

3.1.3. Uncertainty quantification error (ε_{uq})

Errors in stochastic analysis are method-dependent, i.e. sampling error occurs (ε_{mc}) in Monte Carlo methods and truncation error (ε_{sm}) occurs in response surface methods (either conventional or polynomial chaos-based). For example, sampling error could be assumed to be a Gaussian random variable with zero mean and variance given by σ^2/N where N is the number of Monte Carlo runs and σ^2 is the original variance of the model output [21]. The truncation error (ε_{sm}) is simply the residual error in the response surface.

In this paper, due to the use of response surface techniques for uncertainty quantification, truncation error is used to represent ε_{uq}. A polynomial chaos-based response surface is used, which is found to have superior convergence characteristics than conventional response surface models [20]. The response surface is constructed by approximating both the input and output random variables through series expansions of independent standard random variables ξ_i. For example, a normal random variable can be expressed in terms of its parameters as $\mu + \sigma\xi$ where ξ is a standard normal variable. Similarly, a lognormal random variable with parameters λ and δ can be expressed as exp ($\lambda + \delta\xi$). The output response surface is expressed in terms of the input variables through a polynomial chaos expansion as

$$x = a_o + \sum_{i_1=1}^{n} a_{i_1} \Gamma_1(\xi_{i_1}) + \sum_{i_1=1}^{n} \sum_{i_2=1}^{i_1} a_{i_1 i_2} \Gamma_2(\xi_{i_1}, \xi_{i_2}) + \sum_{i_1=1}^{n} \sum_{i_2=1}^{i_1} \sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3} \Gamma_3(\xi_{i_1}, \xi_{i_2}, \xi_{i_3}) + ... \quad (9)$$

where x is the output and $\Gamma_p(\xi_{i_1},...,\xi_{i_p})$ are multi-dimensional Hermite polynomials of degree p. The design points for the data used to construct the response surface are chosen such that they are the roots of the Hermite polynomial of the order p+1 where p is the order of the response surface [2].

The series in Eq. (9) could be truncated to a finite number of terms. Thus the accuracy of the stochastic computational model depends on the order of the expansion. The truncation error ε_{sm} in the response surface of order p can be estimated by constructing additional higher order response surfaces (i.e., order p +1 or p + 2), and using the Richardson extrapolation method, similar to Eq. (8).

3.1.4. Output measurement error (ε_{exp})

The measurement error in the output variable is a separate error component, whereas the measurement error in the input variables is compounded through propagation in the prediction model. Output measurement error is quantified commonly as a random variable with Gaussian distribution, with zero mean and a known or assumed variance.

3.2. Model Form Error (ε_{model})

If multiple models are considered, Bayesian model averaging (BMA) [22, 23] may be used to reduce the model form uncertainty and model errors, but not quantify them explicitly. In some practical cases, only one model may be available, in which case BMA may not be useful. If a single model is employed, this paper makes use of the observed data to express the overall prediction error through a regression model consisting of the individual error components. The residual of such a regression analysis should include the model form error (after subtracting the experimental error effects). From Eqs. (6a) and (6b) and by denoting ε_{obs} as the difference between the data and prediction, i.e., $\varepsilon_{obs} = y - x$, we can obtain the following relation:

$$\begin{aligned} \varepsilon_{obs} &= y - x = \varepsilon_{pred} - \varepsilon_{exp} \\ &= \varepsilon_{num} + \varepsilon_{model} - \varepsilon_{exp} \\ &= f(\varepsilon_h, \varepsilon_{uq}, \varepsilon_d) + \varepsilon_{model} - \varepsilon_{exp} \end{aligned} \tag{10}$$

In Eq. (10), overall numerical error ε_{num} is a nonlinear function of the error components ε_h, ε_{uq} and ε_d. Therefore, it is constructed as a response surface with respect to ε_h, ε_{uq}, and ε_d, using a polynomial chaos expansion, similar to Section 3.1.3. The quantity $\varepsilon_{model} - \varepsilon_{exp}$ is simply the residual $\varepsilon_{residual}$ of such a response surface. Thus the distribution of model error ε_{model} is quantified by knowing the distributions of $\varepsilon_{residual}$ and ε_{exp}. However in most practical situations, the validation data is very limited. From a single validation experiment, one has the numerical values of prediction and observation, and estimates of the numerical errors in prediction, but not the experimental error. In other words, values $f(\varepsilon_h, \varepsilon_{uq}, \varepsilon_d)$, and ε_{obs} are available but the exact value for experimental error ε_{exp} cannot be estimated. Only the distribution of ε_{exp} is available or assumed, if at all. If we have a sufficient number of validation data, we can compute the difference ($\varepsilon_{obs} - f(\varepsilon_h, \varepsilon_{uq}, \varepsilon_d)$) and add a randomly generated term ε_{exp} to it each time to obtain an estimate of model form error ε_{model}. Since the sample size (number of observations made) is limited, an empirical distribution for ε_{model} cannot be constructed with confidence. However, one can compute the statistics like mean and standard deviation of model error from a set of validation experiments. Bootstrapping [24] (sampling with replacement) can be done on the given data set to generate a large number of statistics for model form error, thus obtaining the distributions for mean and standard deviation of model form error. Bootstrapping assumes that the data set in hand is representative of the intended population and no prior assumptions are made regarding distribution of the samples. Further the observations are assumed to be independent and sampling is purely random.

3.3 Sensitivity Analysis

The Bayesian validation metric given in Eqs. (3) and (5) depends explicitly on model output and uncertainties arising from validation experiments like lack of sufficient data points and random measurement errors etc. Also, the statistical and physical model parameters affect the model output and hence a model may be accepted or rejected based on our prior assumption in a Bayesian analysis. Thus, there exists an implicit relation between the Bayes factor and each of the above model parameters, and curves may be fitted to depict this relation. The sensitivity of the Bayes factor to these variables may be estimated from the slopes (first order sense) of such plots. For example, with reference to Eq. (3), the uncertainty due to a limited number of data

points in a pass/fail type of test may be quantified as $\sigma_{x_o}^2 = \frac{x_o(1-x_o)}{n}$. For particular known values of k and x_0, one can generate a plot of $B(x_0)$ versus $\sigma_{x_o}^2$ for different values of n, based on Eq. (3), as shown in Fig. 3.

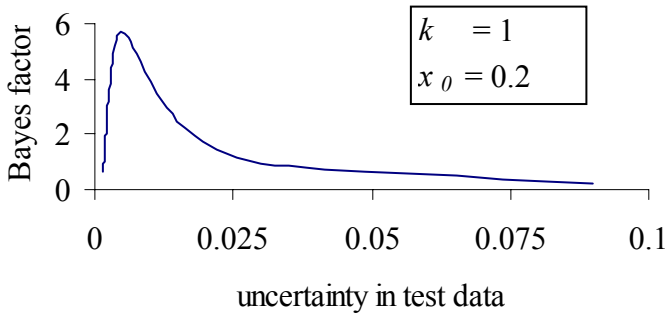


Fig. 3. Relation between $B(x_0)$ and $\sigma_{x_o}^2$

Similar relations between $B(x_0)$ and $\sigma_{\varepsilon_{\exp}}^2$, μ_x, σ_x etc in Eq. (5) can be derived to be used in a sensitivity analysis. A numerical example is provided to illustrate the proposed methodology.

4. Numerical example

The steady state heat transfer in a thin wire of length L, with thermal conductivity k, convective heat coefficient β is of interest. The temperature at midpoint of the wire needs to be predicted. We assume (acknowledging model form error) that this problem is essentially one dimensional and that the solution can be obtained from the boundary value problem

$$-k\frac{\partial T^2}{\partial x^2} + \beta T = Q(x) \qquad (11)$$

with known conditions $T(0)$ and $T(L)$

where $Q(x)$ is the heat source. Suppose k and β are assumed for the sake of illustration to be random variables, normally distributed with statistics N (5, 1) and N (0.5, 0.1) respectively. Also, the heat source $Q(x) = 25(2x\text{-}L)^2$ with $L = 4$. The wire is insulated at the ends, which are kept at zero temperature i.e., $T(0) = T(L) = 0$. It is required to predict $T(2.0)$.

The numerical solution $T(x)$ for Eq. (11) can be obtained using a finite-difference scheme with discretization step size h. Since k and β are random, the model prediction $T(2.0)$ is also random whose statistical distribution needs to be estimated. Since each computation of $T(2.0)$ using a finite-difference scheme could be computationally expensive, a response surface may be fitted to predict $T(2.0)$ as a function of input random variables k and β. A polynomial chaos-based response surface is used for this purpose, as mentioned in Section 3.1.3 earlier. Thus the random variables k and β are expressed as $(5+\xi_1)$ and $(0.5+0.1\xi_2)$ respectively, where ξ_1 and ξ_2 are standard normal variables. The design points for the data used to construct the response surface are chosen such that they are the roots of Hermite polynomial of the order p+1 where p is the order of the response surface. The corresponding values of k and β, with respect to these collocation points, are then used in the numerical model to compute the response $T(2.0)$. The

unknown coefficients of the stochastic response surface are then computed using standard regression techniques.

In this numerical example, a second order response surface in two variables is constructed for $T(2.0)$ and with step-size $h = 1$. The design points for k and β are selected at the collocation points obtained from the roots of 3rd order Hermite polynomials. The response surface is

$$T_{pred} = T(2.0) = 17.102 - 3.003\ \xi_1 - 0.698\ \xi_2 + 0.4964(\xi_1^2 - 1) + 0.0251(\xi_2^2 - 1) + 0.237\ \xi_1\xi_2 \quad (12)$$

where ξ_1 and ξ_2 are independent standard normal variables and $R^2 = 0.999$. The PDF of T_{pred} can be generated by simulating ξ_1 and ξ_2, and is found to have a lognormal distribution with mean 17.12 and variance 10.042. This is the prior density to be considered in the Bayesian model validation next.

4.1 Validation

Suppose for given values of k and β, the numerical model predicted a temperature of 18.5 degrees. A wire made of a material with properties k and β having the same measured values as input to the numerical model was tested three times repeatedly to measure the temperature at location $x = 2$.

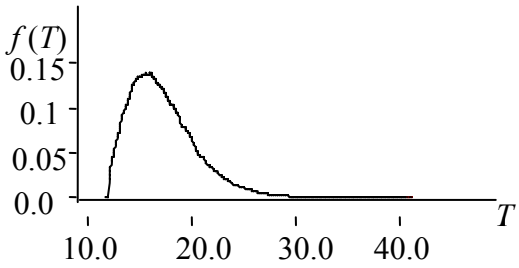


Fig. 4. PDF of $T(2.0)$

The measured temperature was different in each experiment i.e., 18.8, 18.2, 18.9 degrees. Assuming a Gaussian experimental error with zero mean, the true experimental value is assumed to be the mean of the three measurements, i.e., 18.633 degrees for the sake of illustration. Also, the experimental error is assumed to have a variance $\sigma^2_{\varepsilon\exp}$ estimated from the three measurements, again for the sake of illustration.

As described in Section 2.1.1, the likelihood function of the prediction is proportional to a normal density with mean $T(2.0)$ and variance $\sigma^2_{\varepsilon\exp} = 0.1433$. Also with the knowledge of $f(T(2.0))$, the validation metric is evaluated at $T = 18.5$ degrees. Using Eq. (5), the validation metric B is found to be 11.6 which is much greater than 1.0 indicating that the data matches very well with the prediction. However, one should be cautious in accepting this result since various errors like discretization error, input data error, truncation error and even model form error may be canceling each other to produce a result that is close to the measured value. Hence there is a need to estimate the various errors explicitly, as described below.

4.2 Error Estimation

The numerical model related to Eq. (11) was refined using $h = 0.5$ and $h = 0.25$ to estimate the convergence rate $p = 1.985 \approx 2$ as described in Section 3.1.2. The discretization error ε_h based on

the Richardson extrapolation method was obtained as a stochastic response surface in two variables as

$$\varepsilon_h = 5.9725 - 1.1\ \xi_1 - 0.1918\ \xi_2 + 0.1895(\xi_1^2 - 1) + 0.0057(\xi_2^2 - 1) + 0.0681\ \xi_1\xi_2 \qquad (13)$$

The discretization error was found to have a lognormal distribution with parameters $\lambda = 1.762$, and $\xi = 0.1923$ i.e., with mean 5.9725 degrees and a standard deviation of 1.15. Due to the use of the stochastic response surface, the uncertainty quantification error ε_{uq} is estimated by the truncation error $\varepsilon_{sm,}$ i.e., the residual error in the stochastic response surface for the model response $T(2.0)$, which was found to be a Gaussian variable with zero mean and a standard deviation of 0.2 degrees. This is much smaller than the FEM discretization error. The information on input data error ε_d was not available, and ε_d is assumed negligible in this example. Neglecting ε_{sm} and ε_d based on the above observations, the numerical error ε_{num} in Eq. (10) is approximated by ε_h. Thus Eq. (10) reduces to

$$\varepsilon_{obs} = \varepsilon_h + \varepsilon_{model} - \varepsilon_{exp} \qquad (14)$$

In Eq. (14), the distributions of ε_h and ε_{exp} are available from the above discussion. Specific values of ε_{obs} are available from each test. The number of tests, and thus the number of samples of ε_{obs}, is likely to be small in practical problems. In this example, nine values of T_{pred} (and hence ε_h) are calculated at nine collocation point values of ξ_1 and ξ_2 (using Eqs. 12 and 13 respectively), as shown in Table 1. Assume for the sake of illustration that nine corresponding values of T_{obs} (and hence ε_{obs}) are observed from nine tests, also shown in Table 1. A bootstrap resampling technique [24] (with replacement) can then be applied to generate ε_{model}, based on Eq. (14). Each time a value for $(\varepsilon_{obs} - \varepsilon_h)$ is resampled from the nine values shown in Table 1, a randomly generated value of ε_{exp} is added to it, and sample statistics of ε_{model} (mean and standard deviation) are calculated from nine such values in a resample. (Note that each resample contains the same number of data points as the original sample, i.e., nine in this case). The procedure is repeated and 10,000 values for $\mu_{\varepsilon model}$ and $\sigma_{\varepsilon model}$ are obtained, thus giving their distributions as shown in Table 2. This approach thus provides a measure of the uncertainties in the statistical parameters of model form error, since they are obtained by bootstrap resampling.

Table 1. Sample points for model form error

ε_h	T_{pred}	T_{obs}	ε_{obs}	$\varepsilon_{obs} - \varepsilon_h$
5.824	16.597	16.794	0.197	-5.627
8.126	12.902	12.997	0.095	-8.031
5.824	15.642	15.920	0.278	-5.546
4.174	23.222	23.310	0.088	-4.086
5.824	17.653	17.442	-0.211	-6.035
8.126	13.526	13.488	-0.038	-8.164
4.174	21.350	21.181	-0.169	-4.343
8.126	12.327	12.173	-0.154	-8.28
4.174	25.394	25.301	-0.093	-4.267

In this example, the mean model error $\mu_{\varepsilon model}$ was observed to follow a normal distribution (mean -6.03, see Table 2), and the standard deviation of model error $\sigma_{\varepsilon model}$ was observed to follow a Weibull distribution. The numerical error has a mean value of 5.9 degrees while the model form error has a mean value of -6.03 degrees.

Table 2. Distributions of model error statistics

Parameter	Mean	Variance	Type of distribution
$\mu_{\varepsilon model}$	-6.03	0.293	Normal
$\sigma_{\varepsilon model}$	1.6	0.066	Weibull

Looking at these numerical results, the two errors appear to have almost cancelled each other in Eq. (14), resulting in a small overall observed error ε_{obs} and hence a high Bayes factor in Section 4.1, indicating an optimistic model validation result. But the error quantification shows that the prediction model has large numerical and model form errors. Thus a model acceptance/ rejection criterion based on the overall ε_{obs} alone can lead to misleading conclusions, especially in other untested situations when the numerical and model form errors might become additive. This observation shows the importance of quantifying various error components, in addition to simply comparing prediction and observation.

Since the error components add up linearly in this example, the relative first order sensitivities of numerical, model and experimental errors to Bayes factor will simply be proportional to their respective standard deviations. The first order normalized sensitivities of the model parameters k and β to Bayes factor in Eq. (5) are found to be 0.978 and 0.208 respectively at their corresponding mean values. These values are computed by plotting the relation B vs. k and B vs. β separately, and normalizing the products of standard deviations
and slopes evaluated at their mean values.

5. Conclusion

V&V needs to quantify various errors under uncertainty and effectively compare them with imprecisely measured experimental data to assess the predictive capability of the model. A Bayesian approach was proposed for model validation in this paper, and developed for two situations. The first case performs model assessment using a pass/fail criterion and uses the Bayes factor as a metric. The second case includes the uncertainty in the experimental data explicitly and estimates the Bayes factor using prior and posterior distributions of the model output. The overall numerical error in prediction is expressed as a nonlinear response surface in terms of several errors such as discretization error, uncertainty quantification error and input data error, and compared with the observed error to estimate the statistics of the model form error. A bootstrapping technique is used to estimate the model form error from a limited number of experimental measurements. The simple numerical example resulted in a linear relation among various errors and the further work is needed to demonstrate the methodology for a more complex problem where the various errors are combined in a non-linear fashion. Also, the sensitivity analysis limited to a first-order evaluation showed that the validation metric is sensitive to the variance of each error component. Similarly model parameters that have an insignificant effect (small sensitivity) on Bayes factor can be omitted to reduce model complexity.

Acknowledgements

The research reported in this paper was supported by funds from Sandia National Laboratories, Albuquerque, NM (contract no. BG-7732), under the Sandia-NSF Life Cycle Engineering program (project monitors: Dr. Steve Wojtkiewicz and Dr. Thomas L Paez). The support is gratefully acknowledged.

References

1. R. L. Iman, and W. J. Conover. A distribution free approach to introducing rank correlation among input variables. *Communication on statistics- simulation and computation*, 11(3):311-334, 1982.
2. S. S. Isukapalli, and P. G. Georgopoulos. Computational Methods for the Efficient Sensitivity and Uncertainty Analysis of Models for Environmental and Biological Systems. *Technical Report,* State Univ. of New Jersey, 1999.
3. G. I. Schueller, C. G. Bucher, U. Bourgund, and W. Ouypornprasert. On efficient computational schemes to calculate failure probabilities. *Probabilistic engineering mechanics*, 4(1):10-18, 1989.
4. American Institute of Aeronautics and Astronautics, *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*, AIAA-G-077-1998, Reston, VA, 1998.
5. P. J. Roache. *Verification and Validation of in Computational Science and Engineering*, Hermosa, Albuquerque, USA, 1998.
6. M. Ainsworth, and J. T. Oden. A posteriori error estimation in finite element analysis. *Comput. Methods Appl. Mech. Engg*, 142:1-88, 1997.
7. O. C. Zienkiewicz, and R. L. Taylor. *The Finite Element Method. Volume 1-Basic Formulation and Linear Problems*, 4th edition, McGraw-Hill, 1989.
8. O. C. Zienkiewicz, and J. Z. Zhu. The super-convergent patch recovery and a posteriori estimates. Part 2: Error estimates and adaptivity. *Int. J. Numer. Methods Engg*, 33:1335-1382, 1992.
9. Defense Modeling and Simulation Office, *Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide*, Office of the Director of Defense Research and Engg, www.dmso.mil/docslib, Alexandria, VA, April, 1996.
10. T. G. Trucano. Aspects of ASCI Code Verification and Validation. Sandia Technical Report No. SAND2000-0390C, Sandia National Laboratories, Albuquerque, NM, 2000.
11. W. L. Oberkampf, and T. G. Trucano. Verification and Validation in Computational Fluid Dynamics. *Progress in Aerospace Sciences*, 38: 209-272, 2002.
12. R. G. Hills, and I. H. Leslie. Statistical Validation of Engineering and Scientific Models: Validation Experiments to Application. Report No. SAND2003-0706, Sandia National Laboratories, Albuquerque, NM, 2003.
13. A. Urbina, T. L. Paez, T. K. Hasselman, G. W. Wathugala, K. Yap. Assessment of Model Accuracy Relative to Stochastic System Behavior. *Proceedings of 44th AIAA Structures, Structural Dynamics, Materials Conference*, April 7-10, Norfolk, VA, 2003.
14. J. O. Berger, and L. R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc,* 91:109-122, 1996.
15. T. Leonard, and J. S. J. Hsu. *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge University Press, Cambridge, 1999.
16. H. Jeffreys. *Theory of Probability* (3rd. ed.), Oxford University Press, London, 1961.

17. R. Zhang, and S. Mahadevan. Bayesian Methodology for Reliability Model Acceptance. *Reliability Engineering and System Safety*, 80(1):95-103, 2003.
18. Y. Pawitan. *In All Likelihood: Statistical Modeling and Inference using Likelihood*, Oxford Science Publications, New York, 2001.
19. S. A. Richards. Completed Richardson extrapolation in space and time. *Comm. Numer. Methods Engg*, 13:573-58, 1997.
20. R. Rebba. *Computational Model Validation under Uncertainty.* Master's thesis, Vanderbilt University, Nashville, TN, 2002.
21. R. Y. Rubinstein. *Simulation and the Monte Carlo Method.* John-Wiley, New York, 1981.
22. D. Draper. Assessment and Propagation of Model Uncertainty. *J. of Royal Statistical Society Series B*, 57 (1):45-97, 1995.
23. J. A. Hoeting, D. Madigan, A. Raftery, and C. T. Volinsky. Bayesian Model Averaging: A Tutorial (with discussion). *Statistical Science*, 14(4):382 - 417, 1999.
24. B. Efron, B. Bootstrap method: Another look at the jackknife. *Ann. Statist*., 7:1-26, 1979.

Probabilistic Inversion for Chicken Processing Lines

Roger M. Cooke[a], *Maarten Nauta*[b], *Arie H. Havelaar*[b], *Ine van der Fels*[b]
[a]Department of Mathematics Delft University of Technology Delft, The Netherlands
[b]Microbiological Laboratory for Health Protection RIVM, Bilthoven, The Netherlands
r.m.cooke@its.tudelft.nl

Abstract: We introduce probabilistic inversion techniques as applied in a recent example of Campylobacter transmission. Such techniques are indicated when we wish to quantify a model whose parameters cannot be directly measured. In this a (partially specified) uncertainty distribution over measurable quantities predicted by the model can be "pulled-back" onto the parameter space of the model. If a probabilistic inversion is feasible, the solution is seldom unique and we require a method of selecting a preferred solution. If a problem is not feasible, we require a best fitting distribution. This study illustrates two such techniques, Iterative Proportional Fitting (IPF) (Kruithof 1937) and PARmeter Fitting for Uncertain Models (PARFUM) (Cooke 1994). In addition, we illustrate how expert judgement on predicted observable quantities in combination with probabilistic inversion may be used for model validation and/or model criticism.

Key words: Probabilistic inversion, IPF, PARFUM, Campylobacter, transport models, expert judgment, entropy, information.

1. INTRODUCTION

"Probabilistic inversion" denotes a family of techniques which have recently been introduced into the field of risk and environmental modelling. They enable quantification of non-measurable model parameters in terms of distributions over measurable quantities. This is particularly useful when expert judgement is used: experts are queried about outcomes of possible measurements, and their uncertainty distributions are pulled back onto the parameter space of a model which predicts the measured outcomes. Recent applications may be found in (Kraan and Cooke, 2000a,b), for a discussion see (Kraan 2002, Kurowicka and Cooke 2002). A complete mathematical discussion of the techniques employed here is found in (Du et al 2003).

The ‘pull-back’ distribution on model parameters may be pushed through the model to re-predict the quantities assessed by the decision maker. This provides an opportunity for model validation and/or criticism. If the re-predicted distributions agree with the original decision maker’s distributions, then the model provides a suitable vehicle for capturing the decision maker’s uncertainty. If these distributions do not agree, then the model is not suitable to represent the decision maker’s uncertainty. In this case the model must be re-evaluated and possibly revised. For more discussion and examples of this aspect see (Kraan and Cooke 2000b).

This paper gives an informal introduction to probabilistic inversion techniques, illustrated with a recent application to campylobacter transmission. This is a relatively

simple environmental transport model and illustrates nicely how probabilistic inversion applied to structured expert judgment can play a constructive role in model evaluation.

2. CAMPYLOBACTER TRANSMISSION

Campylobacter contamination of chicken meat may be responsible for up to 40% of the annual 100,000 cases of Campylobacter-associated gastroenteritis in The Netherlands, and a similar proportion of an estimated 30 deaths. A recent effort to rank various control options for Campylobacter contamination of chicken carcasses has led to the development of a mathematical model of a typical chicken processing line (Nauta et al in preparation). This model has been quantified in an expert judgment study involving 12 experts (van der Fels et al 2003). Key parameters in the model are transfer coefficients from the chickens' skin and intestines to the processing environment, and from the environment back to the chickens' skin. Experimental data on such transfer coefficients are not available, and experts are unable to quantify their uncertainty on the values of these coefficients. Hence, the model must be quantified by asking the experts about other quantities which, under specific circumstances, can be predicted by the model. These quantities typically involve aggregate phenomena with which experts are sufficiently familiar to render a judgment. The experts need not endorse, or even know the model. Their uncertainty distributions are combined to form a "decision maker's" distribution, as described in (van der Fels et al 2003).

A schematic representation of a typical broiler chicken processing line is given in Fig. 1.

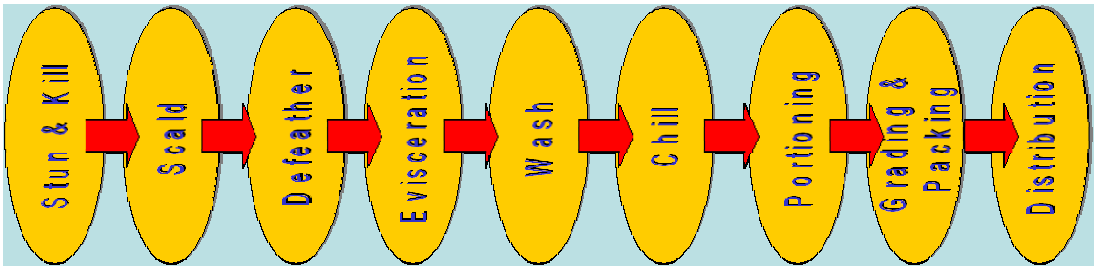


Figure 1. Broiler Chicken Processing line

For campylobacter transmission, the relevant phases are scalding, defeathering, evisceration, washing, and chilling. Two types of scalding processes are considered, namely low and intermediate temperature, as two types of chilling, namely air and spray chilling. Each phase is modelled as a physical transport process. A typical phase in the processing line is illustrated in Fig. 2.

N_{env} represents the number of campylobacter in the physical environment of the chicken in a processing phase (expressed in cfu, colony forming units). N_{ext} is the number on the exterior of the chicken, and C_{int} is the concentration in the intestines, containing the feces. The transfer coefficients are explained in Fig. 2 below, and depend on the processing phase S. Mass balance equations are formulated which say, eg, that the cfu's at the end of phase S on the exterior equals the number at the beginning, minus what is

transferred to the environment or inactivated/removed entirely, plus what moves onto the exterior during phase S:

$$\begin{cases} N_{ext,S}(i) = (1-a_{ext,S})(1-c_{ext,S})N_{ext,S-1}(i) + b_{env,S}N_{env,S}(i-1) + (1-a_{\text{int},S})w_{\text{int},S}(i)C_{\text{int}}(i) \\ N_{env,S}(i) = a_{ext,S}N_{ext,S-1}(i) + (1-b_{env,S})(1-c_{env,S})N_{env,S}(i-1) + a_{\text{int},S}w_{\text{int},S}(i)C_{\text{int}}(i) \end{cases}$$

In equilibrium we have $N_{env,S}(i) = N_{env,S}(i-1)$, so that:

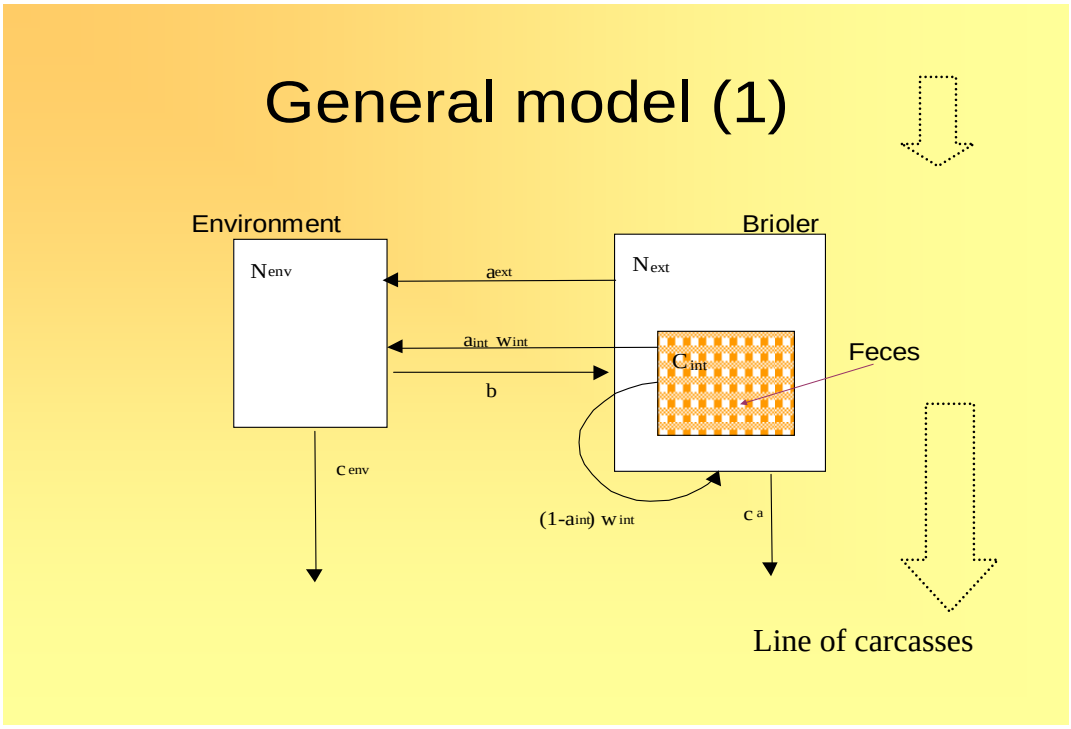


Figure 2. A typical phase in the chicken processing model

$$N_{env,S} = \frac{a_{ext}N_{ext,S-1} + a_{\text{int}}w_{\text{int}}C_{\text{int}}}{b_{env} + c_{env} - b_{env}c_{env}}$$

For more detail, see (Nauta et al in preparation). Ideally, we would like to have numerical values for the coefficients in Fig. 2. However, experimental data yielding these values are not available. Failing that, we would like to quantify the uncertainty in the transfer coefficients. Expert judgment could be applied for this purpose, if the experts had detailed knowledge of the interactions in each processing phase. Unfortunately, that is not the case. Experts are however able to quantify their uncertainty regarding the number of cfu's on a broiler in the situations described below taken from the elicitation protocol:

At the beginning of a new slaughtering day a thinned-flock is slaughtered in a "typical large broiler chicken slaughterhouse". Suppose each chicken of this flock to be infected with Campylobacter, both externally and internally. We suppose every chicken to be externally infected with 10^5 campylobacters per carcass and internally with 10^8 campylobacters per gram of caecal content *at the beginning of each slaughtering stage* (a hypothetical situation). So at the beginning of scalding, plucking etc., each chicken has this (identical) external and internal contamination rate.

Question A1:
All chickens of the particular flock are passing successively each slaughtering stage. How many campylobacters (per carcass) will be found *after* each of the mentioned stages of the slaughtering process, each time on the *first* chicken of the flock?

Experts respond to these questions, for each phase, by stating the 5, 50 and 95% quantiles of their uncertainty distributions. If distributions on the transfer coefficients in Fig. 2 are given, then a distribution, per processing phase, for the elicited variables can be computed from the mass balance equations by a Monte Carlo simulation. Thus, the elicited quantities may be expressed as (the processing phase S is suppressed in the notation):

$$
\begin{aligned}
&A1 = 10\text{^}5 \times (1-a_{ext,}) \times (1-c_a)+10\text{^}8 \times (1-a_{int}) \times w_{int} \times (1-ca);\\
&A2 = A1+b \times (a_{ext} \times 10\text{^}5 + (a_{int}) \times w_{int} \times 10\text{^}8)/(b + c_{env} - b \times c_{env});\\
&B1 = 10\text{^}4 \times (1 - a_{exr} - c_a + a_{exr} \times c_a);\\
&B2= B1 + b \times a_{exr} \times 10\text{^}4/(b + c_e - b*c_e);\\
&C= (1 - b - c_e + b \times c_e)\text{^}99 * b \times (a_{ext} * 10\text{^}5 + (a_{int}) \times w_{int} \times 10\text{^}8)/(b + c_{env} - b \times c_{env});\\
&W_{int} = W_{int}. \qquad (1)
\end{aligned}
$$

The variables A1, A2,… w_{int} are the variables assessed by the experts. Question A2 is similar to A1, but concerns the *last* chicken in the flock. Questions B1 and B2 are similar to A1 and A2, but refer to a flock in which the birds are externally contaminated, but not colonized (internally). Question C asks for the infection on the 100^{th} broiler of an uninfected flock which is processed after an internally and externally infected flock. W_{int} was queried directly. It is included here to indicate that its distribution must conform to the decision maker's quantile assessments. For the first 3 processing phases, we have 6 equations; for later phases the intestines are removed and the variable W_{int} is not defined. In total we have 39 such equations, counting the alternative processes for scalding and chilling[1]. The number of equations is equal to the number of transfer coefficients for the whole line.

Assuming distributions for coefficients on the right hand sides in (1) are known, we could sample from these distributions and build up distributions for the quantities on the left hand side. These quantities are assessed by the experts. We would like these distributions to comply with the quantiles given by the decision maker. The probabilistic inversion problem may now be expressed as follows: find a joint distribution over the

[1] The decision maker considered variable C for washing degenerate, i.e. zero with probability 1. Removing this would give 38 equations.

transfer coefficients, such that the quantiles of the quantities on left hand sides of the above equations agree with the decision maker's quantiles. If more than one such joint distribution exists, pick the least informative of these. If no such joint distribution exists, pick a "best fitting" distribution.

3. PROBABILISTIC INVERSION

Let X and Y be n- and m-dimensional random vectors, respectively, and let G be a function from $\Re^n$ to $\Re^m$. We call $x \in \Re^n$ an inverse of $y \in \Re^m$ under G if $G(x) = y$. Similarly we call X a probabilistic inverse of Y under G if $G(X) \sim Y$, where "~" means "has the same distribution as". If $\{Y \mid Y \in C\}$ is the set of random vectors satisfying constraints C, then we say that X is an element of the probabilistic inverse of $\{Y \mid Y \in C\}$ under G if $G(X) \in C$. Equivalently, and more conveniently, if the distribution of Y is partially specified, then we say that X is a probabilistic inverse of Y under G if $G(X)$ satisfies the partial specification of Y. In the current context, the transfer coefficients in Fig. 2 play the role of X, and the left hand sides of equations (1) play the role of Y. That is:

$Y = (A_{1,1},\ldots\ W_{int,1},\ldots\ A_{1,2}.\ldots..W_{int},5);$ (39 components in total) .

The joint distribution of these variables is partially specified by the decision maker, namely by given 5, 50 and 95% quantiles. The right hand sides of (1) constitute the function G.

If the function G could be inverted analytically, then it would be a simple matter to compute X as $G^{-1}(Y)$. Of course this is generally not possible, and we must devise other ways to find X. A number of approaches could be considered. A thorough discussion of this problem is found in (Du et al, 2003), and a shorter discussion in (Kurowicka and Cooke, 2002). By far most satisfactory to date are techniques based on *sample re-weighting*, and these have been applied to the chicken line model. We first choose an initial distribution for the transfer coefficients (X) such that, when we sample it a large number of times and compute Y via (1), some samples fall within each interquantile interval for each variable in Y, and all samples are physically possible. The choice of initial distribution is not further constrained, but it should cover the range of realistic values. We take N samples from X and compute N samples for Y, yielding N samples for (X,Y). When drawn from the initial distribution, each of the N samples has probability $1/N$. We now wish to re-weight these N samples such that, if we re-sample this distribution, drawing each sample (with replacement) with probability given by its weight, then the quantile constraints on Y are satisfied in the re-sampled distribution.

We describe two strategies for finding the weights, namely Iterative Proportional Fitting (IPF) and PARameter Fitting for Uncertain Models (PARFUM). These involve iteratively re-adjusting an initial set of weights so as to satisfy the constraints. For convenience, we describe this for one processing phase with 6 elicitation variables (Y is restricted to 6 components).

Since each sample contains a value for (A_1, A_2, B_1, B_2, C, W_{int}), and each component falls in one of 4 interquantile intervals, we may represent this sample as a 6-vector of components, each component taking values in {1, 2, 3, 4}. There are $4^6 = 4096$ possible vectors of this type, and we may think of each such vector as an *interquantile cell* containing a number of samples. Not all cells will be physically possible. It is easy to see from (1) that $A_1 \leq A_2$, and $B_1 \leq B_2$. Thus, if the 50% quantile for A_1 is above the 5% quantile for A_2, then it is impossible that A_1 could be above its median while A_2 is below its 5% quantile. Fortunately it is not necessary to figure out which combinations of interquantile intervals are feasible; sampling *X* and computing *Y* via (1) does that automatically. It is well to realize, however, that a large number of mathematically possible interquantile cells may actually be unfeasible under the function *G*. In a typical example for a processing phase, we would draw 65,000 samples and find that 150 – 300 of the 4096 interquantile cells were occupied. The weight assigned to each interquantile cell is simply the total weight of the samples falling in that cell. In our iterative algorithms, two samples falling in the same cell will be treated in the same way; therefore we can restrict our problem to that of finding weights for the non-empty interquantile cells. When these weights are found, we just distribute the cell weight uniformly over the samples in the cell to get the sample weights.

Rather than describe the IPF and PARFUM algorithms formally, it is appropriate here simply to illustrate them on a simple example and report the relevant mathematical facts. Details can be retrieved from the cited literature. For purposes of illustration, we consider only 2 elicitation variables, each with 4 interquantile intervals corresponding to the 5, 50 and 95% quantiles. The interquantile cells can be represented as a 4 × 4 matrix; where, for example, a sample is said to fall in cell (3,2) if it is between the 50 and 95% quantiles for variable 1 and between the 5 and 50% quantiles for variable 2.

We start with an initial distribution over *X* and generate an initial distribution over the interquantile cells, which we represent in Table 2. Note that 6 cells are empty. The marginals are shown in boldface.

0.1966	0.0006	0	0	**0.1972**
0.0407	0.1642	0.005	0	**0.2099**
0	0.0094	0.1196	0.0155	**0.1445**
0	0	0.0008	0.4476	**0.4484**
0.2373	**0.1742**	**0.1254**	**0.4631**	

Table 2. Initial distribution over interquantile cells.

The problem is now to adjust the non-empty cells in table 2 such that the margins equal 0.05, 0.45, 0.45, 0.05; which are the probabilities associated with the decision maker's interquantile intervals.

The IPF algorithm was introduced by (Kruithof 1937) and rediscovered by (Deming and Stefan 1942). Its convergence properties were studied by many, including (Fienberg, 1970, Csiszar, 1975). Simply stated, we first multiply each row by constant, so that the

column sums agree with the target. For the first row this constant is *target / row sum =0.05 / (0.1966+0.0006).* Then we multiply each column by the constant *(target / column sum)* to make the column sums agree with the target, then again the rows are multiplied by a constant, etc. Figure 4 illustrates the procedure. The target margins are shown by the starting distribution.

Csiszar (1975) showed that this algorithm converges if and only if there is a distribution with exactly the same zero's as the initial distribution which satisfies the target margins. In that case IPF converges to the distribution which has minimum information relative to the starting distribution, in the set of distributions with the target margins. This distribution may have zeros in cells where the starting distribution is non-zero. The result holds for arbitrary finite dimensions, and arbitrary finite numbers of cells per dimension. The target marginal distributions need not be the same on each dimension. It is evident that the criterion for convergence becomes more difficult to satisfy as the number of zero's increases. When the criterion is not met the probabilistic inversion problem is infeasible and IPF does not converge. In otherwords, there is *no* distribution over the non-zero cells in the starting matrix which has the target margins. In the case of two dimensions, it is known that IPF oscillates between 2 distributions, in case of non convergence (Csiszar and Tusnady 1983). Nothing is known about the behavior of IPF in higher dimensions when the condition for convergence is not satisfied.

The PARFUM algorithm (Cooke 1994, Du et al 2003) differs from IPF in the following way. Instead of first fitting the row sums, then the column sums, then again the row sums, etc; PARFUM successively averages the row and column sum fits. It is schematized in Fig. 4.

Unlike IPF, PARFUM *always* converges. If the problem is feasible, then it converges to a distribution P which minimizes the following functional:

$$F(P) = I(P^{row\ fit} \mid P) + I(P^{column\ fit} \mid P);$$

relative to the starting distribution. Here, $I(Q \mid P)$ denotes the relative information of Q with respect to P. If P is a solution, that is, if P's row- and column fits agree with the target, then $P^{row\ fit} = P = P^{column\ fit}$, so that $F(P) = 0$. Most importantly, if the problem is feasible, then PARFUM converges to a P with $F(P) = 0$, that is, it converges to a solution (see Du et al. 2003). If the problem is feasible then IPF is generally preferred. If infeasible, then IPF tends to distribute the lack of fit quite unevenly and tends to concentrate weight on a small number of samples. In such cases PARFUM often gives better results. Of course, if IPF does not cycle, we have no way of knowing on a finite number of iterations whether it is converging. Appeal to common sense is appropriate. These algorithms have several advantages relative to other methods. First, they are 'dumb' in the sense that they do not require intelligent steering. Second they avoid computationally expensive matrix manipulations, but simply loop repeatedly through the interquantile cells. Finally, since rows and columns are altered one at a time, the whole sample need not be stored in memory, and there is effectively no limit on the size of problems which can be tackled. There are disadvantages as well. Most significantly,

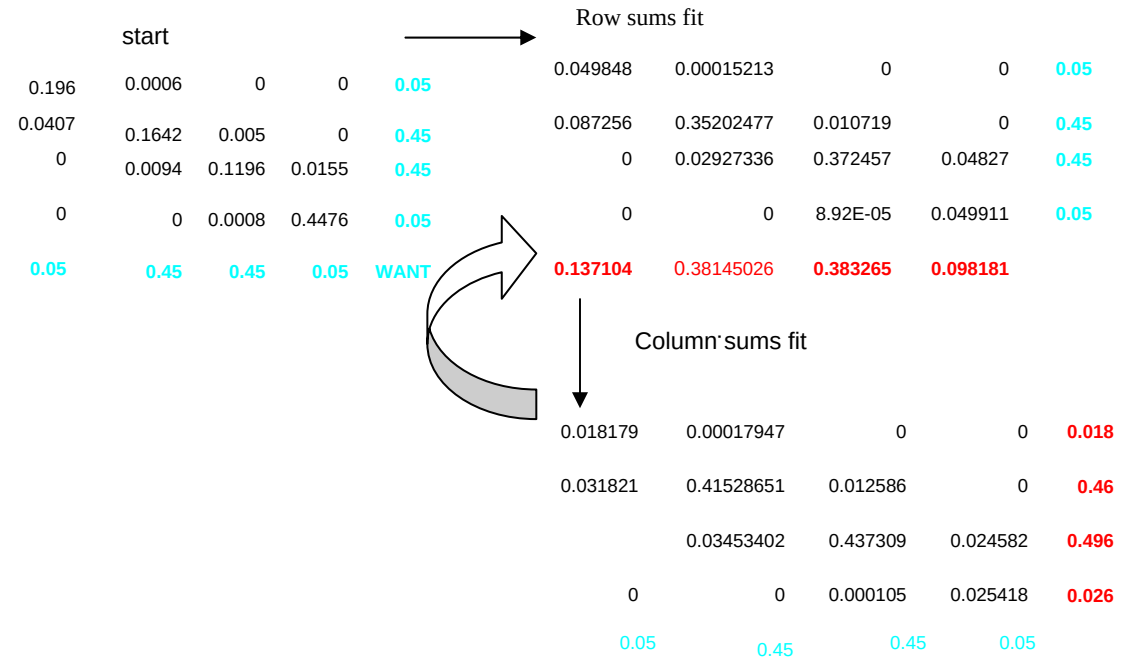


Figure 3. Iterative proportional fitting

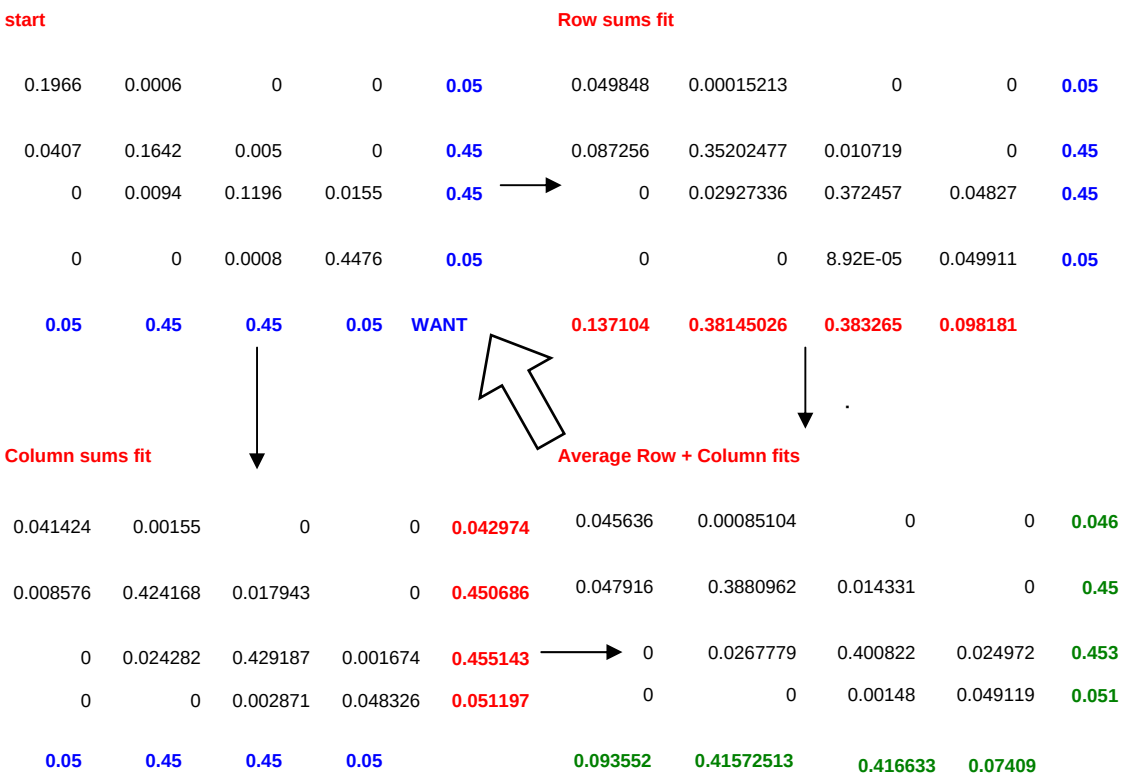


Figure 4. The PARFUM algorithm

it is impossible in practice to know if IPF is converging without verifying the condition for convergence, and this is just as hard as finding the solution. PARFUM has a distinct advantage in this regard. In case of infeasibility, neither algorithm yields information on how the original sample might be extended to yield better solutions.

4. RESULTS

The results with the model described above yielded a very poor fit between the re-predicted and decision maker distributions for some variables. Table 3 shows the results for defeathering. Especially bad fits are circled.

		DEFEATHERING	
elicitation variable	Quantile	PARFUM	IPF
A1	*5%*	0.053	0.014
	50%	0.424	0.175
	95%	0.871	0.719
A2	*5%*	0.030	0.033
	50%	0.256	0.151
	95%	0.543	0.654

Table 3. Re-predicted results of probabilistic inversion defeathering, showing the proportion of samples falling below the corresponding quantile.

Inspection of the experts' rationales revealed that the experts distinguished two transfer mechanisms from the exterior to the environment. Campylobactor in the pores of the skin would be difficult to remove, but on the feathers or skin surface they would come off more easily. It therefore makes a difference whether the birds have been contaminated during transport only (giving rise to only contamination of the exterior) or at the farm (resulting in intestinal colonization and contamination on the exterior. These two different situations had been the starting point of questions A and B. The processing model was therefore altered to include this second transport pathway. The coefficient a_{ext} is replaced by two coefficients, a_{xa} and a_{xb}. The equations for the elicited quantities now become:

$$A1 = 10\text{^}5 \times (1-a_{xa,}) \times (1-c_a)+10\text{^}8 \times (1-a_{int}) \times w_{int} \times (1-c_a);$$
$$A2 = A1+b \times (a_{xa} \times 10\text{^}5 + (a_{int}) \times w_{int} \times 10\text{^}8)/(b + c_{env} - b \times c_{env});$$
$$B1 = 10\text{^}4 \times (1-a_{xb}-c_a+a_{xb} \times c_a);$$
$$B2= B1 + b \times a_{x\ b} \times 10\text{^}4/(b + c_e – b \times c_e);$$
$$C= (1 – b - c_e + b \times c_e)\text{^}99 \times b \times (a_{xa} \times 10\text{^}5 + (a_{int}) \times w_{int} \times 10\text{^}8)/(b + c_{env} - b \times c_{env});$$
$$W_{int} = W_{int}. \qquad (2)$$

With the second model, the probabilistic inversion yielded better fits; partial results are shown in Table 4.

	Quantile	Scalding low IPF	Scalding low PARFUM	Scalding Intermediate IPF	Scalding Intermediate PARFUM	Defeathering IPF	Defeathering PARFUM
variable A_1	0.05	**0.05**	0.05	**0.04**	0.04	0.02	**0.05**
	0.5	**0.50**	0.53	**0.42**	0.45	0.78	**0.49**
	0.95	**0.95**	0.95	**0.81**	0.86	0.97	**0.94**
variable A_2	0.05	**0.05**	0.04	**0.07**	0.07	0.00	**0.04**
	0.5	**0.50**	0.41	**0.70**	0.65	0.14	**0.38**
	0.95	**0.95**	0.93	**0.95**	0.97	0.91	**0.72**

Table 4. Re-predictions with Model(2).

The bold values indicate the solution chosen. The PARFUM solution was chosen in 3 of the seven cases. There is still lack of fit, in particular for defeathering. However, overall, the model revision has produced a better fit. It is a truism that no model is fully adequate to reality. Information regarding the degree and locus of misfit is extremely valuable. The methods discussed here provide such information. The alternative is to search for compliant experts who will assess model parameters directly (and often anonymously); this cannot lead to model improvement.

5. CONCLUSIONS

Iterative sample re-weighting methods are available to solve probabilistic inversion problems, as illustrated in the model of chicken processing lines. IPF and PARFUM are easy to implement and have a solid theoretical foundation. They provide useful tools for the practicing risk modeller. The present study illustrates a fruitful interaction between the modellers and the experts made possible by querying experts on observable quantities and applying probabilistic inversion.

REFERENCES

1. Cooke, R.M. (1994) "Parameter fitting for uncertain models: modelling uncertainty in small models," *Reliability Engineering and System Safety* 44: 89--102.
2. Cooke, R.M. Hora, S. Young. M. Pasler-Ssauer, J. Miller, L. Kraan,B Lui, C. McKay, M. Helton, J. Jones, A. Harper, F. and Goossens L. (1994). "Joint USNRC\CEC consequence uncertainty study: Summary of objectives, approach, application, and results for the dispersion and deposition uncertainty assessment," Technical Report VOL. III, NUREG/CR-6244, EUR 15755 EN, SAND94-1453.
3. Csiszar, I.(1975) "I-divergence geometry of probability distributions and minimization problems." *Ann. of Probability* 3: 146-158.
4. Csiszar, I and Tusnady, G. (1983). "Information Geometry and alternating minimization procedures," *Statistics & Decisions, Supplement Issue,* No. 1: 205-237, Munich.

5. Deming, W.E. and Stephan, F.F. (1940). “On a last squares adjustment of sampled frequency table when the expected marginal totals are known,” *Ann. Math. Statist.* 11: 427-444.
6. Du, C. Kurowicka, D. and Cooke, R.M. (2003) “Iterative Algorithms for Probabilistic Inversion,” Delft Institute for Applied Mathematics, T.U. Delft.
7. Fienberg, S.E. (1970). “An iterative procedure for estimation in contingency tables” *Ann. of Math. Statist.* 41: 907-917.
8. Kraan. B.C.P and Cooke R.M (1997). “Post processing techniques for the joint cec/usnrc uncertainty analysis of accident consequence codes,” *Statis. Comput. Simul*. 57: 243-259.
9. Kraan. B.C.P and Cooke R.M. (2000a). “Processing Expert Judgements in Accident Consequence Modelling," *Radiation Protection Dosimetry* 90(3): 311-315.
10. Kraan. B.C.P and Cooke R.M. (2000b). “Uncertainty in Compartmental Models for Hazardous Materials - A Case Study,” *Journal of Hazardous Materials* 71: 253-268.
11. raan, B.C.P. (2002) *Probabilistic inversion in uncertainty analysis and related topics* PhD dissertation, department of Mathematics, Delft University of Technology
12. Kruithof, J. (1937). “Telefoonverkeersrekening,” De Ingenieur, 52(8): E15-E25.
13. Kurowicka, D. and Cooke, R.M. (2002). “Techniques for Generic Probabilistic Inversion,” *Probabilistic Safety Assessment and Management*, E.J. Bonano, A.L. Camp, M.J. Majors, R.A. Thompson (eds), Elsevier, 1543-1550.
14. Van der Fels-Klerx, H.J., Cooke, R.M., Nauta, M.J., Goossens, L.H.J., Havelaar, A.H. (2003). “A structured expert judgment study for a model of campylobacter contamination during broiler chicken processing.”
15. Nauta, M.J., Van der Fels-Klerx, H.J. and Havelaar, A.H. “A poultry processing model for quantitative microbiological risk assessment.” *Submitted.*

The Good, The Bad, and The Ugly of Predictive Science

François M. Hemez[1]

Engineering Sciences and Applications
Los Alamos National Laboratory
Los Alamos, New Mexico 87545, U.S.A.

Yakov Ben-Haim[2]

Faculty of Mechanical Engineering
Technion — Israel Institute of Technology
Haifa 32000, Israel

Abstract: In computational physics and engineering, numerical models are developed to predict the behavior of a system whose response cannot be measured experimentally. A key aspect of science-based predictive modeling is the assessment of prediction credibility. Credibility, which is demonstrated through the activities of Verification and Validation, quantifies the extent to which simulation results can be analyzed with confidence to represent the phenomenon of interest with accuracy consistent with the intended use of the model. This paper argues that assessing the credibility of a mathematical or numerical model must combine three components: 1) Improving the fidelity to test data; 2) Studying the robustness of prediction-based decisions to variability, uncertainty, and lack-of-knowledge; and 3) Establishing the expected prediction accuracy of the models in situations where test measurements are not available. A recently published Theorem that demonstrates the irrevocable trade-offs between *"The Good, The Bad, and The Ugly,"* or robustness-to-uncertainty, fidelity-to-data, and confidence-in-prediction, is summarized. The main implication is that high-fidelity models cannot, at the same time, be made robust to uncertainty and lack-of-knowledge. Similarly, equally robust models do not provide consistent predictions, hence reducing confidence-in-prediction. The conclusion of the theoretical investigation is that, in assessing the predictive accuracy of numerical models, one should never focus on a single aspect. Instead, the trade-offs between fidelity-to-data, robustness-to-uncertainty, and confidence-in-prediction should be explored. The discussion is illustrated with an engineering application that consists in modeling and predicting the propagation of an impact through a layer of hyper-foam material. A novel definition of sensitivity coefficients is suggested from the slopes of robustness-to-uncertainty curves. Such definition makes it possible to define the sensitivity of a performance metric to arbitrary uncertainty, whether it is represented with probability laws or any other information theory. *This publication has been approved for unlimited, public release on November 18, 2003 (LA-UR-03-8492, Unclassified).*

Keywords: Confidence, decision-making, fidelity-to-data, prediction, robustness, uncertainty.

1. INTRODUCTION

In computational physics and engineering, numerical models are developed to predict the behavior of a system whose response cannot be measured experimentally. A key aspect of science-based predictive modeling is to assess the ***credibility*** of predictions. Credibility, which is demonstrated through the activities of Verification and Validation (V&V), quantifies the extent to which simulation results can be analyzed with ***confidence*** to represent the phenomenon of interest with a degree of accuracy consistent with the intended use of the model [1].

The paper argues that assessing the credibility of a mathematical or numerical model must combine three components: 1) Improve the fidelity, R, to test data; 2) Study the robustness, a^*, of predictions to uncertainty and lack-of-knowledge; and 3) Establish the "prediction looseness," λ_Y, of the model. Prediction looseness here refers to the range of predictions expected from a model or family of models. Its

[1] Technical Staff Member, ESA-WR; E-mail: hemez@lanl.gov.

[2] Professor, Yitzhak Moda'i Chair in Technology and Economics; E-mail: yakov@technion.il.

importance stems from the fact that, to predict with confidence, there should be little difference (or small looseness λ_Y) between the predictions of equally credible models.

The discussion presented in this manuscript is kept at a conceptual level for the most part. Section 2 defines the notations used in the paper. Section 3 briefly discusses the concepts of fidelity-to-data and robustness-to-uncertainty. A Theorem, that unveils the antagonistic relationships between fidelity-to-data, robustness-to-uncertainty, and confidence-in-prediction, is summarized in Section 4. Finally, Section 5 suggests a novel definition of sensitivity coefficients, obtained from the slopes of robustness-to-uncertainty curves. Such definition makes it possible to define the sensitivity of a performance metric to arbitrary uncertainty, whether it is represented with probability laws or any other information theory.

2. BACKGROUND: FIDELITY, ROBUSTNESS, AND CONFIDENCE

Even though the conventional activities of model V&V are generally restricted to improving the fidelity-to-data through the correlation of test and simulation results, and the calibration of model parameters [2-3], the other two components are equally important. The main reason is that optimal models—in the sense of models that minimize the prediction errors with respect to the available test data—possess exactly zero ***robustness*** to uncertainty and lack-of-knowledge [4-5]. This means that small variations in the setting of model parameters, or small errors in the knowledge of the functional form of the models, can lead to an actual fidelity that is significantly poorer than the one demonstrated through calibration.

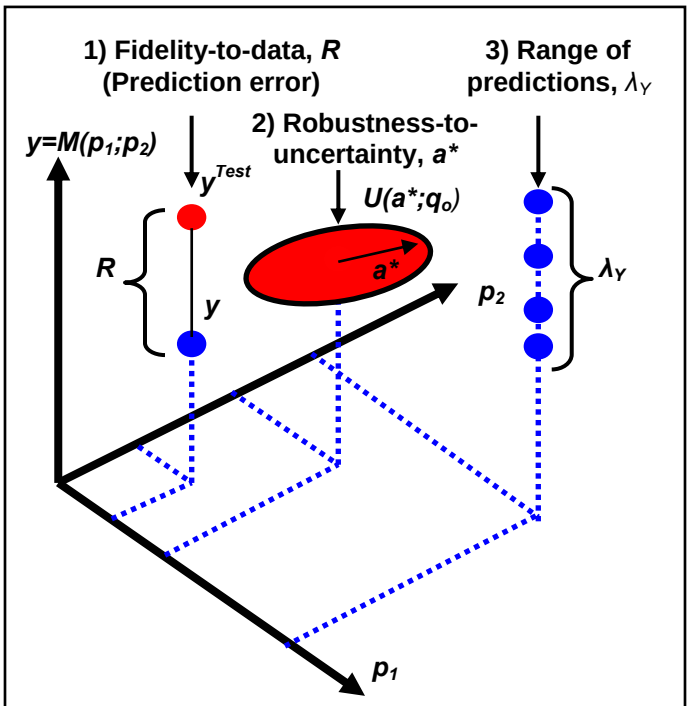


Figure 1. Illustration of the concepts of fidelity-to-data, robustness-to-uncertainty, and prediction accuracy (range, "looseness," or confidence).

Clearly, fidelity-to-data matters because no analyst will trust a numerical simulation that does not reproduce the measurements of past experiments or the information contained in historical databases. Robustness-to-uncertainty is equally critical to minimize the vulnerability of decisions to uncertainty and lack-of-knowledge. It may be argued, however, that the most important aspect of credibility is the assessment of confidence-in-prediction, which is generally not addressed in the literature. Assessing the confidence in prediction here refers to an assessment of prediction error away from settings where physical experiments have been performed, which must include a rigorous quantification of the sources of variability, uncertainty, and lack-of-knowledge, and their effects on model-based prediction.

The concepts of fidelity-to-data, robustness-to-uncertainty, and prediction confidence are illustrated in Figure 1. It is emphasized that, because this is work-in-progress to a great extent, the concept of prediction accuracy denoted in Figure 1 by the symbol λ_Y is somewhat broad. It is analogous to a range of predictions, or "looseness." Clearly, predicting a range of values relates to the notion of confidence-in-predictions. It is believed that future research will narrow down this definition, but a standard accepted throughout the scientific community is not, to the best of the author's knowledge, currently available.

2.1 Definitions and Notations

Throughout the manuscript, the numerical simulation is represented conceptually as a "black-box" input-output relationship between inputs p and q and outputs y. The notation is:

- The quantity y represents the observable outputs. They can be scalar quantities—which is the case assumed here for simplicity—or vector quantities. These model outputs are usually features extracted from a large-order or large-dimensional response.
- The quantity p denotes control parameters of the numerical simulation. These inputs include the control parameters that characterize the experimental configuration. Generally, there will be more than a single input parameter. Inputs p represent settings such as, for example, the angle of attack and flow velocity of an aero-elastic simulation that predicts a coefficient of lift $y=C_L$.
- The quantity q represents parameters that specify the structure and coefficients of the family of models developed to represent the physical phenomenon of interest. The inputs q can include discrete and continuous parameters that control the functional form of the model. Various models can be functions of different subsets of parameters q.

In the general case, the model is represented as:

$$y = M(p;q_o) \tag{1}$$

where q_o denotes nominal settings for the parameters q. In the following, the subscript $(\,)_o$ represents the nominal condition of a quantity.

A domain denoted by D_P represents the design space over which predictions must be obtained. This implies that the prediction accuracy must be established for all settings p in the design domain D_P. In the case, for example, of a two-dimensional operational space where $p = (p_1;p_2)$, like the one pictured in Figure 1, the prediction accuracy of the model $y = M(p;q_o)$ must be studied for all combinations $(p_1;p_2)$ that belongs to $D_P = [p_1^{(min)};p_1^{(max)}] \times [p_2^{(min)};p_2^{(max)}]$.

The quantity $M(p;q)$ is used to denote alternative possible physical models. This notation is introduced to recognize that some of the model parameters q may be subjected to parametric variability. Others may be uncertain, or represent an epistemic lack-of-knowledge about the functional form of the model. For example, the behavior of a particular material under a fast transient load may not be known with certainty. Having to choose between, say, a linear elastic model, a model of perfect plasticity, or a visco-elastic model with hardening represents an epistemic uncertainty denoted by q. In the absence of epistemic uncertainty, no alternative to $M(p;q_o)$ would be feasible. As the horizon of modeling uncertainty increases, more and more alternative models become candidates. The family of predictive models can therefore be represented in a generic sense by the equation:

$$U(a; q_o) = \left\{ M(p; q) \;\middle|\; \|q - q_o\| \le a \right\}, \quad \text{for} \quad a \ge 0 \tag{2}$$

where the symbol a and norm $\|\ \|$ are left undefined for now. It suffices to say that a is a positive scalar quantity that represents the ***horizon-of-uncertainty***. The meaning of definition (2) is that the family of models $U(a;q_o)$ becomes increasingly inclusive as the parameters q are allowed to differ from their nominal settings q_o. Note that these definitions are purposely broad to encompass a wide range of models and uncertainties.

Measurements are denoted by the symbol y^{Test}. Measurements are made at specific experimental configurations controlled by the parameters p. The notation used throughout this paper is that replicate measurements made to estimate the environmental variability are collected in the same vector or matrix quantity y^{Test}. Measurements made, on the other hand, for different configurations $p_1 \ldots p_m$ will be indexed as $y^{Test\,(1)} \ldots y^{Test\,(m)}$.

2.2 Fidelity, Robustness, and Confidence

Fidelity-to-data represents the distance R—assessed with the appropriate metrics, possibly a statistical test if probabilistic information is involved—between physical measurements y^{Test} and simulation predictions y at a setting $(p;q_o)$:

$$R = \left\| y^{Test} - y \right\| \tag{3}$$

Fidelity-to-data is pictured in Figure 1 as the vertical distance between a measurement y^{Test} and a prediction y for the physical experiment and numerical simulation performed at $(p_1;p_2)$.

Robustness-to-uncertainty refers to the range of settings q that provide no more than a given level of prediction error R_{Max}. The concept of robustness is illustrated in Figure 1 by showing a subset $U(a^*;q_o)$ of the design domain $[p_1^{(min)};p_1^{(max)}]\ x\ [p_2^{(min)};p_2^{(max)}]$. The significance of the concept of robustness-to-uncertainty is that all predictions made for settings q chosen inside the domain $U(a^*;q_o)$ are guaranteed not to exceed the error level R_{Max}. The a-parameter represents the "size" of the domain $U(a;q_o)$. The definitions of the sizing parameter and corresponding domain are arbitrary at this point because the purpose of this discussion is to introduce concepts. The only constraint to satisfy is that increasing values of the sizing parameter a must define nested domains $U(a;q_o)$. Reference [5] defines the families of domains as convex sub-spaces. This choice allows the analyst to accommodate a wide variety of uncertainty and lack-of-knowledge models.[3] Clearly, a large robustness-to-uncertainty (a^*) is more desirable than a small one (a') because the former subspace will encompass all events defined in the latter one, or $U(a';q_o) \subset U(a^*;q_o)$. A large robustness indicates that potentially large uncertainty does not deteriorate the prediction error by more than R_{Max}. Generally, a trade-off must be decided upon between the robustness-to-uncertainty (a^*) and prediction error (R_{Max}), or fidelity-to-data. Studying such trade-off is the basic concept of the information-gap theory for decision-making under severe uncertainty [4-5].

Finally, the symbol λ_Y in Figure 1 refers to the range of predictions made by a family of potentially different models. The importance of λ_Y stems from the fact that, to have confidence in predictions, there should be as much consistency as possible between the predictions provided by equally credible sources of information. Confidence is generally increased when different sources of evidence all reach the same conclusion. The concept of confidence-in-prediction is illustrated in Figure 1 by showing a range λ_Y of predictions obtained when different models are exercised to make predictions at a setting $(p_1;p_2)$ where no test data are available. The ultimate goal of model V&V is to establish predictive confidence by estimating the range of predictions λ_Y (or, equivalently, the lack-of-consistency) provided by equally credible sources of information.

Note that the terminology "model" is here defined in a broad sense. In any realistic application, sources of evidence include expert judgment, back-of-the-envelope calculations, measurements, and predictions obtained from phenomenological models or high-fidelity simulations. These can all be considered as models because they define a relationship between the inputs p and the outputs y. All available sources of information must be taken into account to assess the credibility of numerical simulations. It is equally important to understand, quantify, and eventually combine the uncertainty associated with each source of information. The integration of uncertainty and definition of total uncertainty metrics are areas of active research at Los Alamos National Laboratory, although not addressed in this publication [6].

[3] A first example is a probabilistic model of variability where the values of coefficients in the covariance matrix are controlled by the parameter a. A second example is a possibility structure defined to represent a lack-of-knowledge, where the size of intervals is proportional to the parameter a. A third example is a family of fuzzy membership functions defined to represent expert judgment and linguistic ambiguity, where the membership functions are parameterized by the uncertainty parameter a.

3. THE TRADE-OFF OF FIDELITY AND ROBUSTNESS

A family of models $U(a;q_o)$ such as equation (2) defines a model of ***information-gap***. In the theory of information-gap for decision-making, the difference between what is currently known and what needs to be known to make a decision is modeled. Models of ignorance are hence associated to gaps in knowledge [4-5]. This is a significant departure from other representations of uncertainty, such as the probability theory, that attempt to model the randomness itself. Doing so requires strong assumptions that might not be justifiable in the case of lack-of-knowledge.[4]

3.1 Fidelity-optimal Strategy for Model Selection

Let $M(p;q)$ be any physical model in the information-gap domain $U(a;q_o)$. The fidelity-to-data, R, of this particular model can be estimated by calculating a test-analysis correlation metric (3). Clearly, many models can potentially be included in the family $U(a;q_o)$, some with higher fidelity-to-data (or smaller R values) than others. For a given horizon-of-uncertainty a, an analyst might have the choice between several models to make predictions, and a natural strategy would be to use the model that exhibits the best fidelity to the existing data, R^*. In mathematical terms, this approach consists of solving an optimization problem defined as:

$$R^* = \min_{M \in U(a;\, q_o)} R \tag{4}$$

Solving the optimization (4) provides the set of optimal parameters q^* (or optimal model) for which the predictions $M(p;q^*)$ lead to the best fidelity-to-data R^*. This is precisely the problem that parameter calibration solves.

The inappropriateness of this strategy for choosing a model or making decisions comes from the fact that the horizon-of-uncertainty, a, is generally unknown. An example in mechanical engineering is the definition of a friction coefficient between two materials. A value of the friction coefficient might be available from the literature, but the extent of the variability is typically unknown. What it even more difficult to assess is the suitability of the Coulomb friction model—for which a friction coefficient is sought—to represent the mechanics of friction. These un-doubtfully involve stick-and-slip and complex micro-mechanics that the Coulomb friction can only approximate. The extent to which this model is in error compared to the "true-but-unknown" behavior is generally unknown. Therefore, any uncertainty model that aims at representing the lack-of-knowledge associated with the friction model would have to be associated to an unknown horizon-of-uncertainty.

3.2 Robustness-optimal Strategy for Model Selection

The main point of this discussion is that a natural trade-off arises between fidelity-to-data and robustness-to-uncertainty. Instead of fixing the horizon-of-uncertainty—which is practiced all the time, for example, in probabilistic analysis when standard deviation or total entropy are initialized and kept constant—and optimizing the fidelity-to-data, the robustness-to-uncertainty can be maximized for a given ***aspiration*** of fidelity-to-data. The fidelity aspiration is denoted as R_{Max}, and it represents a value of prediction error not to be exceeded. This means that a model is rejected if its fidelity-to-data is poorer than the aspiration, or $R>R_{Max}$. It could also happen that models are found that outperform our original fidelity aspiration, which would indeed be good news. In mathematical terms, this approach consists of solving the following embedded optimization problems:

[4] In probability theory, for example, the frequency of occurrence of random events needs to be assessed. Enough measurements and observations might not be available to derive a probability density function with confidence. In extreme cases, only ranges of values can be obtained. Similar difficulties are encountered with physics-based models, historical databases, and expert opinion. Defining a specific model of uncertainty—such as probabilities, possibilities, or a fuzzy structure—might require assumptions that the available evidence simply does not support.

$$a^* = \max_{a \geq 0} \left\{ \min_{M \in U(a;\, q_o)} R \leq R_{Max} \right\} \tag{5}$$

where a^* denotes the robustness-to-uncertainty, or largest amount of uncertainty that can be tolerated in our knowledge of the model and its parameters, while guaranteeing a fidelity-to-data at least equal to R_{Max}. As pointed out earlier, it could happen that the robust-optimal model features a better fidelity-to-data, or $R<R_{Max}$, a situation referred to in Reference [5] as opportunity from uncertainty.

3.3 Trade-off Between Fidelity and Robustness

Just like the fidelity-optimal strategy for model selection defines an ordering preference where the model $M(p;q^*)$ is preferred to the model $M(p;q)$ if $R^*<R$, the robustness-optimal strategy defines an ordering preference where the model $M(p;q^*)$ is preferred to the model $M(p;q)$ if it is more robust to the uncertainty, that is, $a^*>a$. As mentioned previously, a large robustness is more desirable than a small robustness because it indicates that potentially large sources of uncertainty do not deteriorate the prediction error by more than R_{Max}.

We are not advocating that fidelity-optimality, as a decision strategy for building and validating models, be systematically replaced with robustness-optimality. Instead, investigating the trade-off between the aspiration of fidelity-to-data R_{Max} and robustness-to-uncertainty a^* should be the basis for building and validating models. One significant advantage gained in doing so is that information-gap models can encompass a wide range of uncertainty: probabilistic or non-probabilistic, from parametric uncertainty to linguistic ambiguity and modeling lack-of-knowledge, etc. One practical limitation is the amount of calculation involved in the saddle-point optimization problem (5).

4. THE TRADE-OFFS OF FIDELITY, ROBUSTNESS, AND LOOSENESS

In this section we explore the "looseness" of model prediction: the range of predicted values deriving from models which all satisfy a specified fidelity requirement. The notion of prediction looseness (or range of predictions as it is also referred to below) is important because it relates to the confidence that one has in the predictions of equally credible models. We prove a Theorem whose meaning is that a change in the model that enhances fidelity-robustness to modeling error also increases the looseness of the model prediction. In other words, fidelity-robustness and prediction-looseness are antagonistic attributes of any modeling effort.

4.1 More Definitions

Let a^* be the robustness-to-uncertainty of model $M(p_k;q)$ at the experimental configuration p_k, as defined in equation (5). Let $U^*=U(a^*;q_o)$ denote the set of models whose fidelities are no worse than the aspiration R_{Max}, for the k^{th} experiment defined by parameters p_k. Note that both a^* an U^* depend upon the model specification, q. We have no reason to reject any model $M(p_k;q)$ in U^* if fidelity-to-data is used as the measure of merit. This is because all models $M(p_k;q)$ included in the family U^* satisfy, by definition, the aspiration of fidelity-to-data, $R<R_{Max}$. The "best" model is therefore non-unique, which is a well-established result in inverse problem mathematics. As discussed previously, some of these models may be more robust to the uncertainty than others, up to the upper limit $a=a^*$. An alternative model selection strategy is to identify models associated with the largest robustness a.[5]

If a^* is large, then U^* contains a wide range of models. The predictive looseness of the family of models, $M(p_k;q)$, that belong to U^* is simply defined as the range of predictions in U^*:

[5] Even with robustness-optimal model selection, uniqueness of the solution is not guaranteed. Because the family of models $U(a;q_o)$ is constructed as a family of increasingly-including—or nested—convex subsets, models that would possess the same fidelity-to-data, R, ***and*** the same robustness-to-uncertainty, a, can still occur on the convex hull of the domain $U(a;q_o)$.

$$\lambda_Y = \max_{M \in U(a^*;q_o)} M(p_k;q) - \min_{M \in U(a^*;q_o)} M(p_k;q) \tag{6}$$

The definition of a range (6) may not be the most appropriate to represent the concept of confidence-in-prediction provided by a family of models U^*, but it is clearly related. This choice is motivated by the fact that no standard is currently available in the scientific community to express confidence. It also offers simplicity, both conceptually and computationally.

Large robustness, a^*, and small range of predictions, λ_Y, are both desirable. We will say that robustness and range are ***sympathetic*** if a change in input variables or model form parameters q improves them both; otherwise they are ***antagonistic***:

$$\text{Sympathetic}: \left(\frac{\partial a^*}{\partial q}\right)\left(\frac{\partial \lambda_Y}{\partial q}\right) \le 0, \quad \text{Antagonistic}: \left(\frac{\partial a^*}{\partial q}\right)\left(\frac{\partial \lambda_Y}{\partial q}\right) \ge 0 \tag{7}$$

The Theorem enounced in Section 4.2 shows that, under fairly weak conditions, robustness and range are always antagonistic. The axioms of nesting and translation are needed to support the main result. Nesting expresses that, as the horizon-of-uncertainty increases, the family of models includes all previously included models, plus new ones. Translation expresses that two families of models that share the same horizon-of-uncertainty, a, only differ in their center points. The axioms are not further discussed because they can easily be realized with arbitrary models of uncertainty.

4.2 Theoretical Result of the Robustness-range Trade-off

The main Theorem is given as follows. Let $U(a;q_o)$ be an information-gap family of models that obeys the axioms of nesting and translation, and let $a=a(p_k;q_o;R_{Max})$–versus–R_{Max} be its robustness function. Consider two initial models, q_o^* and q_o'. If $a(p_k;q_o^*;R_{Max}) \ge a(p_k;q_o';R_{Max})$, then $\lambda_Y(q_o^*) \ge \lambda_Y(q_o')$. That is, robustness and range (or prediction looseness) are antagonistic.

A proof of the Theorem can be found in Reference [7]. The proof relies on the information-gap description of ignorance. No restrictive assumption is made regarding the type of models, sources or types of uncertainty, and their mathematical representations. This makes the theory applicable to a wide range of situations.

4.3 Exploring the Trade-offs of Predictive Accuracy

Three quantities are central to the information-gap analysis of modeling and forecasting: fidelity of the model to the data, R; robustness-to-uncertainty, a^*; and the range of predictions, λ_Y, from models of comparable fidelity. Combining the results of References [5] and [7] yields several trade-offs briefly discussed in the remainder.

First, the relationship between robustness (a) and looseness (λ_Y) can be written as:

$$\frac{\partial \lambda_Y}{\partial a} \ge 0 \tag{8}$$

which means that a revision of the model, with the purpose of enhancing robustness to modeling error, also increases the looseness of predictions. In other words, robustness and prediction looseness are antagonistic attributes of any model. Extending this result to the following three inequalities is trivial:

$$\frac{\partial a}{\partial R} \ge 0, \quad \frac{\partial \lambda_Y}{\partial a} \ge 0, \quad \frac{\partial \lambda_Y}{\partial R} \ge 0 \tag{9}$$

The inequalities (9) express the trade-offs between fidelity, robustness, and looseness:

- ***Robustness decreases as fidelity improves.*** The robustness-to-uncertainty gets larger if the prediction error increases. Numerical simulations made to better reproduce the available test data become more vulnerable to errors in modeling assumptions, errors in the functional form of the model, and uncertainty and variability in the model parameters.

- ***Looseness increases as robustness improves.*** The prediction looseness gets larger if the robustness-to-uncertainty increases. Numerical simulations that are more immune to uncertainty and modeling errors provide a wider, hence less consistent, range of predictions.
- ***Looseness decreases as fidelity improves.*** The range of predictions gets larger if the prediction error gets larger. Numerical simulations made to better reproduce the available test data provide more consistent predictions. Although intuitive, this result is not necessarily a good thing when the models are employed to analyze configurations of the system that are very different from those tested.

These trade-offs imply that it is not possible to have, simultaneously, high fidelity, large robustness, and small prediction looseness. High fidelity (small R) implies that the model is true to the measurements, which adds warrant to the model. Large robustness (large a) strengthens belief in the validity of the model or family of models. Small looseness (small λ_Y) implies that all the models that are equivalent in terms of fidelity, also agree in their predictions of the system behavior. The conflict between robustness, fidelity and prediction looseness is reminiscent of Hume's critique of empirical induction. Our analysis shows that past measurements, accompanied by incomplete understanding of the measured process, cannot unequivocally establish true predictions of the behavior of the system.

5. AN APPLICATION TO PREDICTABILITY IN ENGINEERING

The theoretical results discussed in Sections 3 and 4 are illustrated with an engineering application. The numerical simulation of the propagation of an impact wave through a layer of non-linear, crushable foam material is compared to physical measurements [8]. Details about the experimental set-up, finite element modeling, and sources of uncertainty can be obtained from Reference [9]. In the following, the main source of uncertainty analyzed in this work is briefly described, and analysis results are discussed to illustrate the trade-offs between fidelity-to-data, robustness-to-uncertainty, and prediction looseness. Final comments are made about the concept of sensitivity in the case where calculating partial derivatives or variance coefficients is not mathematically feasible.

5.1 Sources of Uncertainty

The main source of uncertainty considered in this analysis arises from a lack-of-knowledge about the material behavior of the crushable foam. It is illustrated in Figure 2 that shows data obtained from four physical experiments. The center curve (second curve from the left, in blue) represents the nominal model. The other three curves are also acceptable realizations of a material behavior. This uncertainty matters greatly because selecting a constitutive law that describes how the material behaves is a critical step of building the numerical simulation.

Because Figure 2 represents a rather severe lack-of-knowledge about the material, the predictions made from numerical simulations must account for this uncertainty. A pre-requisite is to obtain a mathematical representation of this uncertainty. Clearly, deriving a probability law based solely on the evidence captured by Figure 2 would be nothing short of crystal-ball reading. For the same reason, we are not confident postulating a possibility structure, basic Dempster-Shafer probability assignments, or fuzzy membership functions, to name only a few. It is recognized that more testing could be performed, and formal expert elicitation techniques are available to help capture knowledge. The merits of acquiring more knowledge, in one form or another, can never be over-stated. Nevertheless, Figure 2 illustrates a practical reality where decisions must often be made in the context of severe uncertainty because of constraints such as timetables, budgets, staffing, and lack of testing.

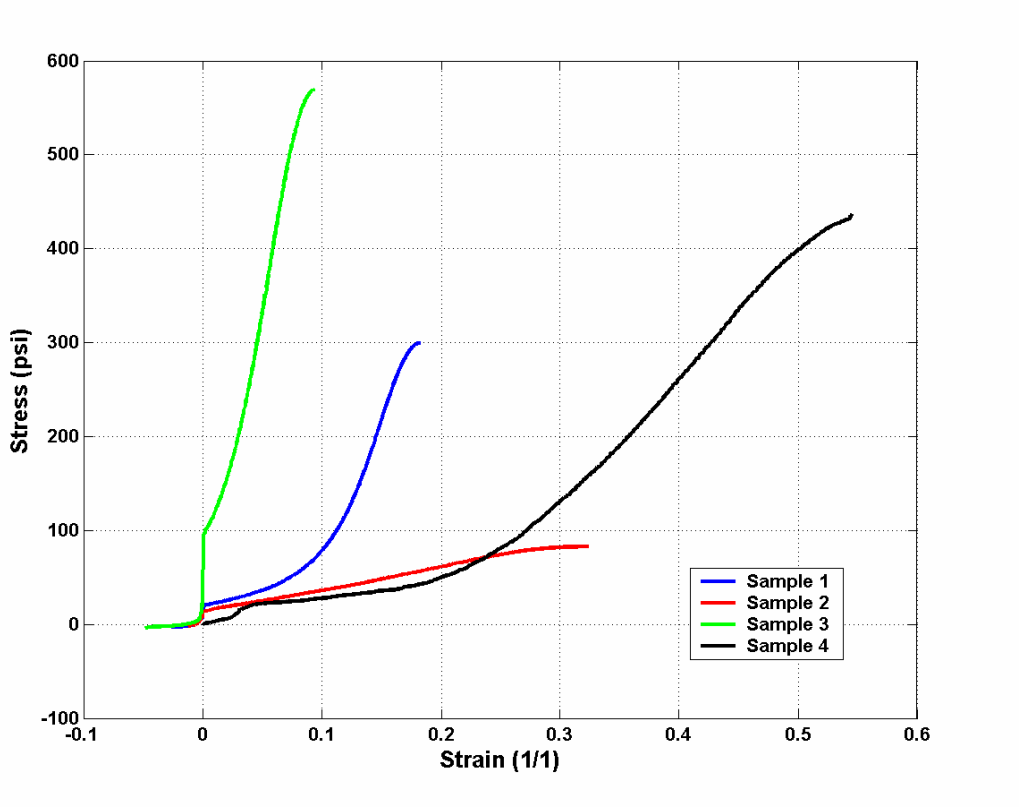


Figure 2. Four test samples of material behavior for the crushable foam.

In this study, the uncertainty pictured in Figure 2 is represented mathematically by a family of convex sets $U(a;q_o)$. The nominal material model q_o represents the baseline curve shown in blue color (second from the left). The horizon-of-uncertainty a measures the distance between any curve q and the nominal curve q_o using the Root Mean Square (RMS) metric. The convex domain $U(a;q_o)$ defined in equation (2) therefore becomes the set of predictions made by all possible numerical simulations $M(p;q)$ for which the material model q does not "deviate" from the nominal material model q_o by more than a RMS distance equal to a.

It is emphasized that such description of lack-of-knowledge can accommodate a wide variety of practical situations. Here, the symbol q represents variations of material coefficients for linear, bi-linear, or cubic material models. These variations can be made as arbitrary as desired, to the extent where the material models, q, do not violate the basic laws of physics. This procedure lets the analyst use his/her prior knowledge to restrict the variations of q around q_o to the ones that make sense. It equally is critical to realize that, in our analysis, the extent of the horizon-of-uncertainty, a, is unknown. The family of models $U(a;q_o)$ becomes increasingly inclusive of material models as the uncertainty represented by the parameter a increases. This is a significant departure from the representation of uncertainty provided by probability theory, and a key aspect of the analysis results presented next.

5.2 Consistency of Predictions

Figure 3 illustrates the ranges of predictions (maximum minus minimum values) obtained at different horizon-of-uncertainty levels. The horizontal axis represents ranges of predictions, in our case, ranges of peak acceleration of the impact wave as it propagates through the crushable foam material. The vertical axis represents the horizon-of-uncertainty. An expected trend is observed: As the horizon-of-uncertainty increases, the range of predictions made also grows. It expresses that the models included in a given family $U(a;q_o)$ make increasingly less consistent predictions. Discontinuities could be encountered, but

each curve is monotonically increasing because the domains $U(a;q_o)$ are nested within one another. Figure 3 illustrates the antagonistic nature of robustness-to-uncertainty and confidence-in-prediction.

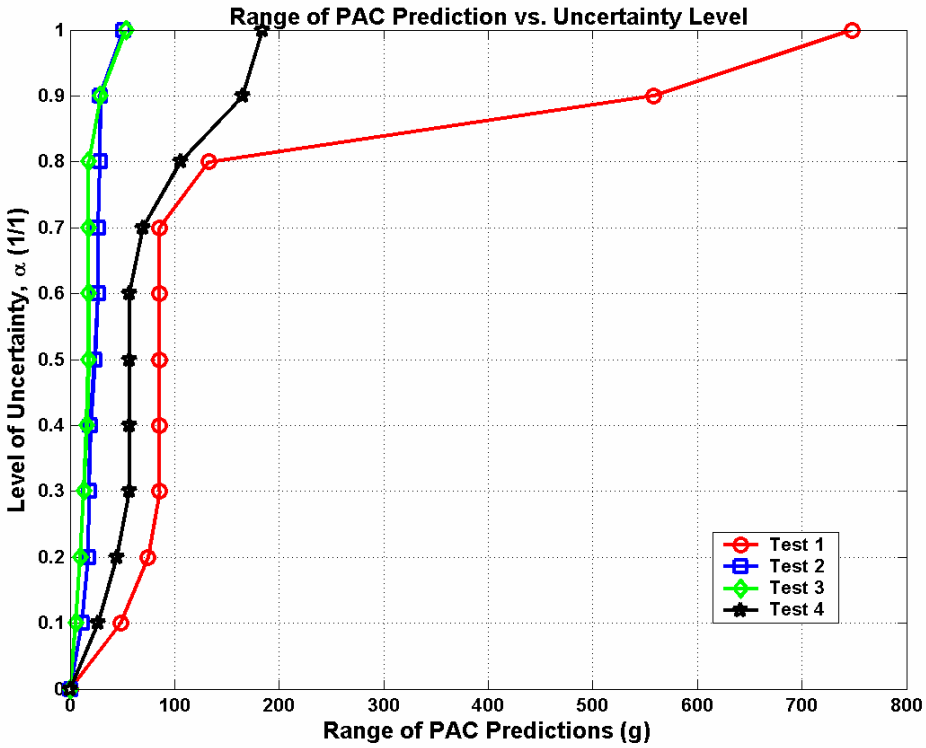


Figure 3. Ranges of predictions obtained at increasing levels of horizon-of-uncertainty.

The four curves shown in Figure 3 present results for four different settings of the system. Test 1 (in red color with circle symbols) simulates the propagation of a low impact through a thin layer of material. Test 2 (in blue color with square symbols) simulates the propagation of a low impact through a thick layer of material. Test 3 (in green color with diamond symbols) simulates the propagation of a high impact through a thin layer of material. Lastly, Test 4 (in black color with star symbols) simulates the propagation of a high impact through a thick layer of material. It can be observed that the family of models $U(a;q_o)$ provides more consistent predictions at any given level of uncertainty, a, for the second and third configurations of the system. If anything, this would tend to increase the credibility of the numerical simulation for predicting testing conditions similar to those of Tests 2 and 3. It does not, however, address prediction accuracy, which is analyzed next.

5.3 Accuracy of Predictions

The results of fidelity-to-data and robustness-to-uncertainty are examined. Figure 4 shows the robustness and opportunity curves obtained with the four configurations of the system (Tests 1 through 4), as a function of increasing levels of horizon-of-uncertainty. The horizontal axis represents test-analysis correlation errors, in our case, the error are expressed as percent between measured peak acceleration values y^{Test} and predicted values y. The vertical axis represents the horizon-of-uncertainty.

For each configuration of low/high impact and thin/thick foam, two curves are shown. The robustness curves, in solid lines, are the worst possible test-analysis correlation errors that can be obtained from a family of models at a given horizon-of-uncertainty. As expected, fidelity-to-data worsens with uncertainty, hence, illustrating the antagonistic nature between truthfulness to data and robustness-to-uncertainty.

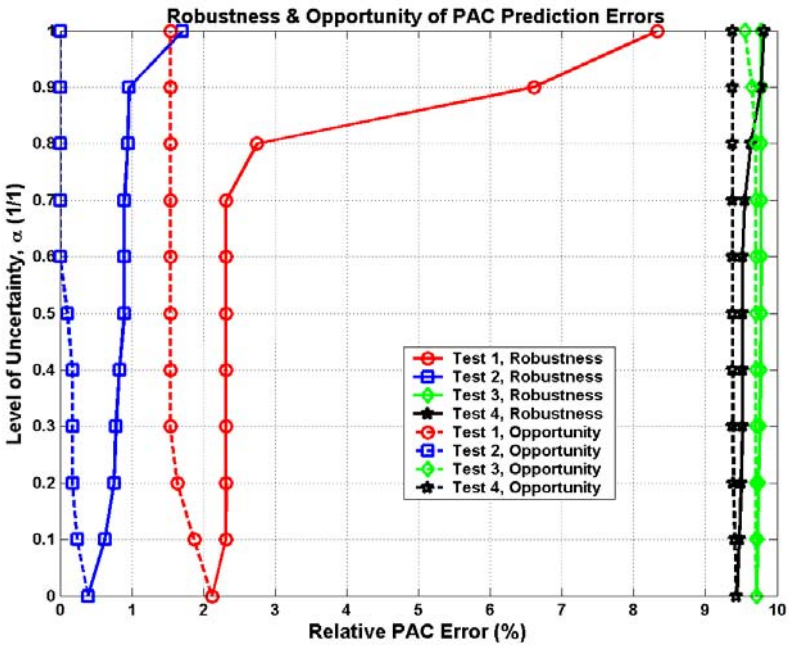


Figure 4. Test-analysis errors obtained at increasing levels of horizon-of-uncertainty.

The opportunity curves, shown in dashed lines, are the best possible test-analysis correlation errors that can be obtained from a family of models at a given horizon-of-uncertainty. Opportunity curves illustrate that uncertainty can sometimes be taken advantage of. For Tests 1 and 2, Figure 4 shows that tolerating a small level of uncertainty can yield material models that are more truthful to the test data.[6] As mentioned previously, the robustness and opportunity curves can be discontinuous, but their monotonically increasing (for robustness) or decreasing (for opportunity) natures are guaranteed by the fact that increasing horizon-of-uncertainty levels, a, generate nested domains $U(a;q_o)$.

When the robustness and opportunity curves are considered together, Figure 4 shows the ranges of predictive accuracy that can be obtained for each configuration of the system. For example, the prediction error for Test 1 is guaranteed within [1.5%; 2.7%] at the horizon-of-uncertainty level of $a=0.8$. No matter which material model is selected from the family $U(a;q_o)$, its predictive accuracy of Test 1 will be no worst than 2.7% but no better than 1.5%. A slight increase in uncertainty from $a=0.8$ to $a=0.9$ results in a potential deterioration of predictive accuracy from the interval [1.5%; 2.7%] to [1.5%; 6.6%]. Such analysis that examines the trade-off between fidelity-to-data and robustness-to-uncertainty provides useful guidelines to estimate a tolerable level of uncertainty.

Similarly, combining Figures 3 and 4 shows that trade-offs must be made between fidelity-to-data, robustness-to-uncertainty, and looseness-in-prediction. Tolerating increased levels of uncertainty in the knowledge of the material behavior offers the potential benefit of slightly improving the fidelity-to-data while, at the same time, reducing the consistency of predictions made from the family of models.

5.4 Sensitivity Analysis

The last question examined is sensitivity analysis (SA). SA provides valuable information to assess the effect of a parametric change or an uncertainty on a performance metric. Local methods generally rely on small perturbation theory, while global methods rely on statistical analysis. Examples are finite

[6] Figure 4 shows that the prediction error for Test 1 is equal to 2% with the nominal material model q_o (when $a=0$, no uncertainty). If the material model is allowed to "deviate" from the nominal model up to the horizon of-uncertainty of $a=0.2$, models can be found that yield no more than 1.5% test-analysis error.

differences in the former case, and variance-based effect analysis in the latter case. The question asked is how to handle a non-probabilistic, non-parametric uncertainty such as the material lack-of-knowledge described in Figure 2.

The performance metric of interest in this work is the fidelity-to-data, R. Lack-of-knowledge is represented by an information-gap horizon-of-uncertainty parameter, a. Sensitivity coefficients are therefore defined as:

$$S_a = \frac{\partial R}{\partial a} \tag{10}$$

We already know from the theoretical investigation that an increase in robustness worsens the fidelity-to-data, which means that $S_a > 0$. Nevertheless, slopes S_a of the robustness-versus-fidelity curves are of great interest because they indicate "how fast" robustness can be lost. A visual illustration is provided in Figure 4, where it can be observed that the numerical simulation is least sensitive to the lack-of-knowledge for predicting the third and forth configurations of the system (Tests 3 and 4). A qualitative SA is summarized in Table 1. The sensitivity coefficients are estimated from Figure 4 by averaging the slopes of the four robustness curves over the entire range of fidelity-to-data values R.

Table 1. Qualitative sensitivity analysis.

	Thin layer of material	**Thick layer of material**
Low impact condition	**Test 1: High (S_a~6)**	**Test 3: Low (S_a<1)**
High impact condition	**Test 2: Medium (S_a~1)**	**Test 4: Low (S_a<1)**

SA results are useful to understand the vulnerability of the simulation to the uncertainty. Clearly, our analysis shows that the model is most vulnerable to uncertainty for Test 1. This means that a significant deterioration in predictive accuracy could result from a small increase in lack-of-knowledge. Caution should be exercised when the numerical simulation is employed to analyze configurations similar to the low impact/thin foam pad settings of Test 1. It is of interest to mention that the measured data provided by Test 1 were the only data used in this study to calibrate the material models. The fact that the average sensitivity coefficient S_a for Test 1 is the highest should not come as a surprise: It is, yet, another manifestation of the theoretical result that ***calibrated models have the least robustness-to-uncertainty*** [5].

Although the sensitivity coefficients S_a do not convey any physical meaning, their main advantage is that they can be estimated based on the robustness-versus-performance functions (such as shown in Figure 4), for any type of uncertainty or lack-of-knowledge. It is emphasized that the information-gap representation of ignorance adopted throughout this work is not exclusive of other representations of uncertainty. Reference [8], for example, presents results where a probabilistic description of uncertainty is handled within the information-gap framework.

6. CONCLUSION

This work studies the relationship between several aspects of prediction accuracy. The main conclusion is that, in assessing the predictive accuracy of numerical models, one should never focus on a single aspect only. Instead, the trade-offs between fidelity-to-data, robustness-to-uncertainty, and confidence-in-prediction should be explored. One consequence that cannot be emphasized enough is that the calibration of numerical models—which focuses solely on the fidelity-to-data aspect—is not a sound strategy for selecting models capable of making accurate predictions. Calibration leaves models vulnerable to modeling uncertainty. It is further established that models selected for their robustness to uncertainty will tend to make inconsistent predictions. This finding seems discouraging because one would like to make accurate predictions while being robust to the sources of uncertainty and lack-of-knowledge. It is, however, a fundamental limitation of predictive science that scientists and engineers

should not loose sight of. The trade-off simply expresses that obtaining accurate predictions is conditioned by the assumptions upon which the models are built.

REFERENCES

[1] Doebling, S.W., "Structural Dynamics Model Validation: Pushing the Envelope," *International Conference on Structural Dynamics Modeling: Test, Analysis, Correlation and Validation*, Madeira Island, Portugal, June 3-5, 2002.

[2] Mottershead, J.E., Friswell, M.I., "Model Updating in Structural Dynamics: A Survey," *Journal of Sound and Vibration*, Vol. 162, No. 2, 1993, pp. 347-375.

[3] Hemez, F.M., Doebling, S.W., "Inversion of Structural Dynamics Simulations: State-of-the-art and Orientations of the Research," *25th International Conference on Noise and Vibration Engineering (ISMA)*, Leuven, Belgium, September 13-15, 2000, pp. 403-413.

[4] Ben-Haim, Y., "Robust Rationality and Decisions Under Severe Uncertainty," *Journal of the Franklin Institute*, Vol. 337, 2000, pp. 171-199.

[5] Ben-Haim, Y., **Information-Gap Decision Theory: Decisions Under Severe Uncertainty**, Academic Press, 2001.

[6] Anderson, M.C., Booker, J., Hemez, F.M., Joslyn, C., Reardon, B., Ross, T., "Quantifying Total Uncertainty in a Validation Assessment Using Different Mathematical Theories (U)," *14th Biennial Nuclear Explosives Design Physics Conference*, Los Alamos National Laboratory, Los Alamos, New Mexico, October 20-24, 2003. *Unclassified*.

[7] Ben-Haim, Y., Hemez, F.M., "Robustness-to-uncertainty, Fidelity-to-data, and Prediction-looseness of Models," *22nd SEM International Modal Analysis Conference*, Dearborn, Michigan, January 26-29, 2004.

[8] Hemez, F.M., Ben-Haim, Y., Cogan, S., "Information-gap Robustness for the Test-analysis Correlation of a Non-linear Transient Simulation," *9th AIAA/ISSMO Symposium on Multi-disciplinary Analysis and Optimization*, AIAA-2002-5420, Atlanta, Georgia, September 4-6, 2002.

[9] Hemez, F.M., Wilson, A.C., Doebling, S.W., "Design of Computer Experiments for Improving an Impact Test Simulation," *19th SEM International Modal Analysis Conference,* Kissimmee, Florida, February 5-8, 2001, pp. 977-985.

Our Calibrated Model has No Predictive Value: An Example from the Petroleum Industry

J.N. Carter[a], *P.J. Ballester*[a], *Z. Tavassoli*[a] *and P.R. King*[a]

[a] Department of Earth Sciences and Engineering,
Imperial College, London, SW7 2AZ, United Kingdom.
Email: j.n.carter@imperial.ac.uk.

Abstract: It is often assumed that once a model has been calibrated to measurements then it will have some level of predictive capability, although this may be limited. If the model does not have predictive capability then the assumption is that the model needs to be improved in some way.

Using an example from the petroleum industry, we show that cases can exit where calibrated models have no predictive capability. This occurs even when there is no modelling error present. It is also shown that the introduction of a small modelling error can make it impossible to obtain any models with useful predictive capability.

We have been unable to find ways of identifying which calibrated models will have some predictive capacity and those which will not.

Keywords: Prediction, Calibration, Uncertainty, Petroleum

1. INTRODUCTION

In many studies involving numeric models of complex real world situations, for example petroleum reservoirs and climate modelling, it is implicitly assumed that if the model has been carefully calibrated to reproduce previously observed behaviour, then the model will have some predictive capacity. It is recognised that predictability may only be achievable for a finite period of time, and that any prediction will be uncertain to some extent.

Two types of error are considered in most calibration exercises: measurement error and model error. Measurement errors are fixed at the time the measurement was made, they generally have well defined statistics and can be handled appropriately. Model errors are due to approximations, such as a loss of spatial, or temporal, resolution, and the non-inclusion of all of the relevant physics. The assumption that is normally made is that if the model errors are sufficiently unimportant, so that when the model has been calibrated to measurement data, then we have some level of acceptable predictability. If the model does not have predictability, then the model errors are assumed to be too large and we need to use a "better" model. Where "better" probably means improved resolution, spatial or temporal, and/or the inclusion of more physics.

In this paper we present the results of a study, for a petroleum reservoir, where a well calibrated model has no predictive value. Even though the calibration and truth models had identical physics and identical spatial and temporal resolution. A second study shows

Correspondence to J.N. Carter

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

that where there are slight differences in the physics between the calibration and truth models, then the problems encountered are even worse.

In the next section the experimental set-up is described, this is followed by the results for cases with/without modelling errors. Finally we draw some conclusions from our observations.

2. EXPERIMENTAL SET-UP

In this section we describe our three parameter reservoir model and our methodology for calibrating the model against the available measurements.

2.1. Model Description

Our model is a cross-section of a simple layered reservoir, with a single vertical fault midway between an injector producer pair, as shown in figure 1. The model that we calibrate has three parameters: the vertical displacement (throw) of the fault; the permeability of the poor quality sand; and the permeability of the good quality sand. The geological layers are assumed to be homogeneous (ie they have constant physical properties). The "truth" case, which is used to generate the measurements for the calibration, is a variant of the calibration model, but with fixed parameter values. In the case of no model error, then the "truth" case is a member of the set of all possible calibration models. The size and type of model error is chosen by how a specific calibration model is perturbed to obtain the truth case. In the work presented in this paper, the model error is obtained by introducing small variations into the spatial properties of the geological layers. The permeability and porosity in each grid block are randomly perturbed. The maximum variations that are allowed is $\pm 1\%$ of the unperturbed mean values. These perturbations are much lower than would be expected for a real world rock that had been classified as homogeneous. A more extensive description of the model can be found a paper that deals with estimating model errors[4].

2.2. Calibration Methodology

Our procedure to produce a calibrated model is as follows:

1. Choose "truth" values for the three model parameters;
2. Select the level of measurement and model error to be used;
3. From the truth case produce the measurements required for the calibration process (three years of monthly data);
4. Calibrate the model against the measurements;
5. Predict the behaviour for years 4-10.

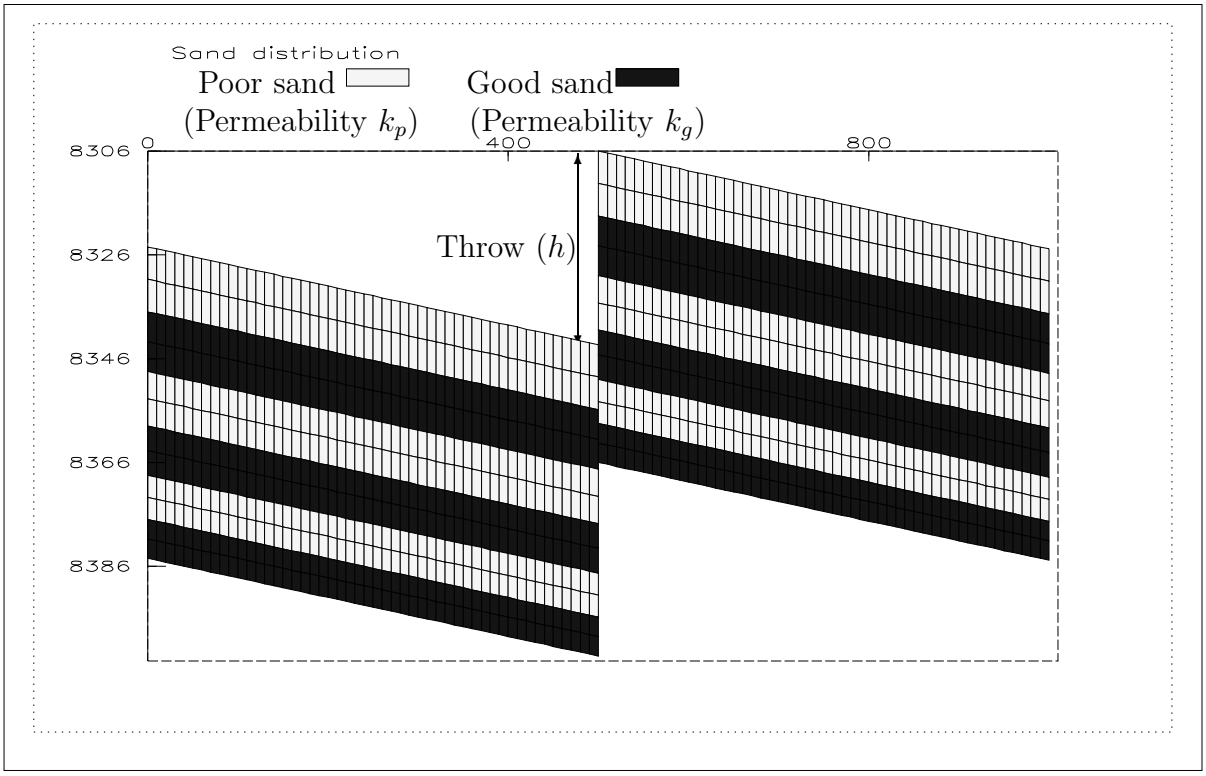


Figure 1. Reservoir model showing the fault throw and the geological, and simulation, layers.

We have considered the truth case: $h = 10.4$, $k_p = 1.31$ and $k_g = 131.7$ with and without model error. No measurement error was added, but we assumed Gaussian noise with a 1% standard deviation when calculating the likelihood that a proposed calibration matches the truth.

In order to quantify the degree of the model calibration against measurements, we define first an objective function for the calibration period, Δ_m, as follows

$$\triangle_m = \frac{1}{36} \sum_{j=1}^{36} \sum_{k=1}^{3} \frac{|sim(j,k) - obj(j,k)|}{2\sigma_{jk}} \tag{1}$$

where $sim(j,k)$ is the simulated response for production series k of the model at time j, $obj(j,k)$ is the corresponding true value and σ_{jk}, an estimation of what would be the associated measurement error. We consider three production series: Oil Production Rate, Water Production Rate (or Water Cut) and Water Injection Rate.

Likewise, the objective function for the prediction period, Δ_f, is

$$\triangle_f = \frac{1}{7} \sum_{j=37}^{43} \sum_{k=1}^{3} \frac{|sim(j,k) - obj(j,k)|}{2\sigma_{jk}} \tag{2}$$

The ranges that the model parameters were allowed to take are: $h \in (0, 60)$, $k_g \in (100, 200)$ and $k_p \in (0, 50)$.

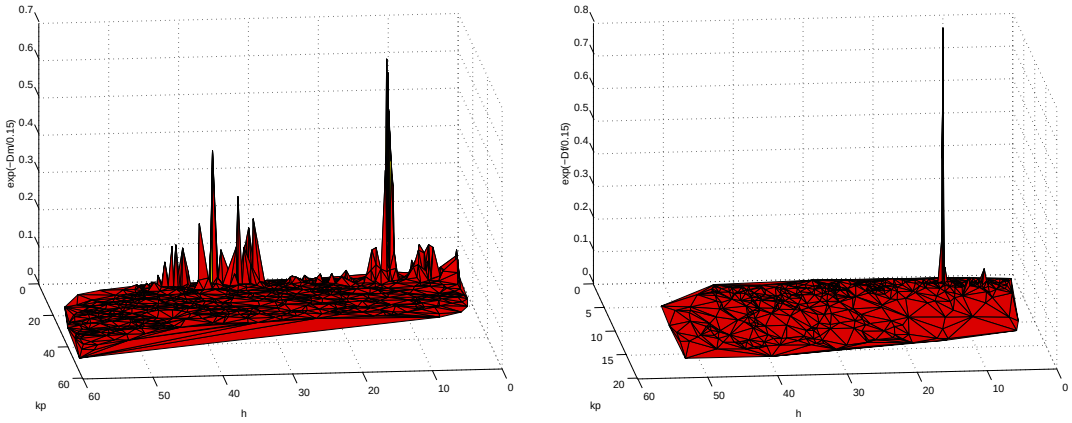


Figure 2. Calibrations of the model (with no modelling error) to a) history period, b) prediction period.

2.3. Genetic Algorithm

Our chosen search method is a Steady-state Real-parameter Genetic Algorithm. This is used because we need to search for multiple good optima within a parameter space that seems to contain very many local optima. It is a development of a previously published study [1] and has been developed to solve the type of problem described in this paper.

In brief the details are: a steady-state population of 50 individuals is used, parents are selected randomly (without reference to their fitness), crossover is performed using vSBX[1, 3], and culling is carried out using a form of tournament selection involving 10 individuals, a total of 7000 individuals are generated.

3. RESULTS

In this section we present the results of two studies: the first is with no modelling error present; the second has a low level of modelling error.

3.1. Calibration with No Modelling Error

Figure 2a shows the result of calibrating the model against the data for the first 36 months. The truth model has exactly the same physics and structure as the calibration models, and the truth model is a member of the set of possible calibration models.

The very large spike, with $h \approx 10$, corresponds to the truth case. We can also see notable local optima with $0 < h < 8$, $30 < h < 38$ and $40 < h < 45$. The global optimum has a small basin of attraction around it and has proved difficult to identify in previous work[2], the easiest optimum to find has been the one with $30 < h < 38$. The rather noisy structure of the objective surface is largely an artifact of the of the way that k_g is sampled. Any point with an acceptable objective value is plotted no matter what value of k_g was used. This means that it is possible for two points to have identical values for h and k_p but different values of the objective function. Hence a vertical line would be plotted. Figure 3 shows a contour plot, centred on $h = 5.0$ and $k_p = 1.65$, of the objective

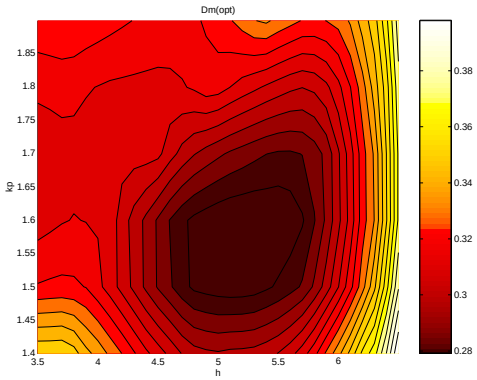


Figure 3. Surface plot for Δ_m, where k_g has been optimised so as to minimise Δ_m, $h \in (3.5, 6.5)$ and $k_p \in (1.40, 1.90)$

function Δ_m. The figure was generated by conducting a grid search on a fine grid. At each point on the grid, k_g was optimised, this results in a much smoother representation of the objective.

Figure 2b shows the result of calibrating the model to the prediction period. The only substantial point found corresponds to the truth model. All of the other local optima that can be seen in figure 2a are unable to match the observations during the prediction period. We conclude that for this model you can only obtain a good prediction from the truth case, and that good matches from the history matching phase have no predictive value.

3.2. Calibration with Modelling Error

The result of matching the calibration model to data generated by a truth case that includes modelling errors is shown in figure 4a. Superficially the figure is similar to figure 2a. The important difference is that the global optima now occurs for $h \approx 32$. This is within the largest basin of attraction and is usually found by most search algorithms. The optima associated with the "true" parameter values is of much lower quality.

If we now look at the calibration to the prediction period, figure 4b, we see that the global optimum for the history matching period has no predictive value. None of the models that have some predictive value correspond to the truth case (the spike at $h \approx 10$ has the wrong values for k_p and k_g). The objective values obtained are low compared to those in figure 2b.

4. CONCLUSIONS

In this paper we have examined, for a particular case, our ability to calibrate a model and then to make accurate predictions. This has been carried out for cases with and without modelling errors, but no measurement error.

From these studies we make the following observations:

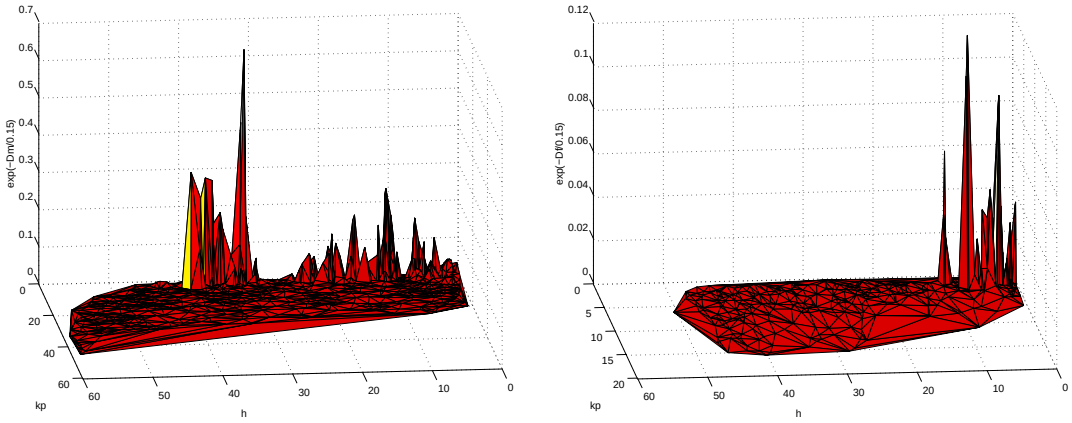


Figure 4. Calibrations of the model (with modelling error) to a) history period, b) prediction period.

- The basin of attraction a round a global optimum may be sufficiently small that search algorithms may not find them. The basins of attraction associated with other local optima may be much larger and hence easier to find.
- When there is no modelling error present, some of the non-global optima may be of quite good quality. However only the global optimum is able to make an accurate prediction.
- When small amounts of modelling error are present, then the global optimum is no longer associated with the truth. The local optimum that has parameter values of the truth case is not of significant quality and could easily be disregarded.
- None of the models tested in the presence of modelling errors have valuable predictive power. In particular the global optima from the history matching period was unable to provide an accurate prediction.

In summary: in the absence of model errors, and with very low measurement errors, it is possible to obtain calibrated models that do not have any predictive capability; such models may be significantly easier to identify than the correct model; we are unable to differentiate between calibrated models with or without predictive capabilities; the introduction of even small model errors may make it impossible to obtain a calibrated model with predictive value.

In this analysis there is nothing that seems to be unique to this model. In particular there is the issue of data availability, adding more measurements does not appear to offer a guaranty of avoiding this dilemma. If the observations made with this model are not unique to the model, and we have no reason to believe that the model is unique, then this presents a potentially serious obstacle to the use of models of this type for prediction.

Our concern is that if we cannot successfully calibrate and make predictions with a model as simple as this, where does this leave us when are models are more complex, have substantive modelling errors, and we have poor quality measurement data.

REFERENCES

1. P.J. Ballester and J.N. Carter,Real-parameter Genetic Algorithms for Finding Multiple Optimal Solutions in Multi-modal *Optimisation, Genetic and Evolutionary Computation Conference* 2003:706, Lecture Notes in Computer Science 2723, Pub Springer-Verlag, Heidelberg,2003.
2. M.D. Bush and J.N. Carter, Applications of a Modified Genetic Algorithm to Parameter Estimation in the Petroleum Industry, *Intelligent Engineering Systems Through Artificial Neural Networks* 6:397, 1996.
3. J.N. Carter, Introduction to using Genetic Algorithms, in Soft Computing and Intelligent Data Analysis in Oil Exploration, Eds M. Nikravesh, F. Aminzadeh and L.A. Zadeh, *Developments in Petroleum Science* 51, Elsevier Science, 2003.
4. J.N. Carter, Using Bayesian Statistics to Capture the Effects of Modelling Errors in Inverse Problems, accepted for publication in *Mathematical Geology*, expected to be published in 2004.

Towards Guidance in Assessing and Communicating Uncertainties

Peter H.M. Janssen[a], Arthur C. Petersen[a], Jeroen P. van der Sluijs[b], James S. Risbey[c] and Jerome R. Ravetz[d]

[a] Netherlands Environmental Assessment Agency (RIVM/MNP), P.O. Box 1, 3720 BA, Bilthoven, The Netherlands (Peter.Janssen@rivm.nl);
[b] Copernicus Institute for Sustainable Development and Innovation, Dept. of Science Technology and Society, Utrecht University, The Netherlands;
[c] School of Mathematical Sciences, Monash University, Clayton, Australia;
[d] Research Methods Consultancy (RMC), London, UK.

Abstract: In the daily practice of science for policy, as experienced by governmental agencies which inform policymakers on the state and outlook of the environment, there is a pressing need for guidance in assessing and communicating uncertainties. This need extends beyond the quantitative assessment of uncertainties in model results per se, and focuses instead on the entire process of environmental assessment, running from problem framing towards reporting the results of the study. Using the Netherlands Environmental Assessment Agency (RIVM/MNP) as a case, the development, structure and content of such a guidance system is highlighted. Conditions for a successful implementation of the guidance system are discussed, and some prospects for future work are outlined.

Keywords: uncertainty assessment, communication, guidance, typology, tool catalogue, science-policy interaction

1. INTRODUCTION

At the onset of 1999, the Netherlands National Institute for Public Health and the Environment (RIVM) was faced with a credibility crisis due to public criticism in a Dutch quality newspaper by an employee of the RIVM. He criticized the institute for suggesting an unjustified level of certainty in reporting their environmental studies, by not duly accounting for uncertainty and relying too much on the virtual reality of poorly validated models. His criticism attracted much media attention in the Netherlands and triggered extensive public and political debate on the credibility, reliability and quality of environmental statistics and model-based environmental foresight, as well as on the role and position of science in policy-making (pp. 285-287 in [1];[2]).

This event does not stand alone, (see, e.g., the recent upheaval around 'The Skeptical Environmentalist' by Bjørn Lomborg; compare [3]), but can be seen as typical for the role and position of science for policy and society in a world which is becoming increasingly interlinked and complex. Now decisions are urgent, stakes are high and diverse, values are in dispute, uncertainty and ignorance involved are high, and trust is fragile (Funtowicz and Ravetz, [4,5,6]). All these problems are common for sustainability, risk and safety issues. The changing relationship between science, policy and society calls for processes and arrangements where issues such as transparency and novel forms of quality control (e.g. extended peer review), public participation, multiple perspectives, reflexivity, transdisciplinarity and accountability are at the forefront in establishing knowledge that is more socially robust (Gibbons and Nowotny et al. [7,8,9,10]).

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

At RIVM, the above-mentioned credibility crisis was the impetus for developing a system of guidance for assisting its employees in their daily practice of performing research to advise policy-makers and the public on the state and outlook of the environment, placing special focus on the assessment and communication of uncertainties. In this paper the development of this guidance system will be described, and major parts of it will be highlighted. We will end with discussing conditions for its successful implementation and outlining prospects of future work in this area.

2. ON THE DEVELOPMENT OF THE GUIDANCE SYSTEM

After the media affair in 1999 there was a national and international review of the RIVM's environmental assessment activities. These reviews led to the start-up of a multidisciplinary project on uncertainty assessment and to the development of guidance system for uncertainty assessment and communication in an environmental assessment setting. It was judged that the scope of the guidance system should extend beyond the mere quantitative assessment of uncertainties in model results per se, and should focus instead on the entire process of environmental assessment. It should involve issues such as problem framing, stakeholder participation, indicator selection, appraisal of the knowledge base, mapping and assessment of relevant uncertainties, and reporting of the uncertainty information. Choices and judgments for all these aspects are potentially of key importance for ascertaining that the most relevant uncertainties are identified and for communicating them. Therefore the guidance system should explicitly stimulate reflection on these issues, since this is expected to lead to more conscious choices and a better way of dealing with uncertainties. It should provide a prioritized list of uncertainty types and sources that need particular attention for the case at hand, in view of its societal context and the function of the assessment. Moreover the system is intended to give advance warnings of which bottlenecks can occur with respect to dealing with these uncertainties and what additional effort should then be made in the field of uncertainty assessment. It should offer advice on the selection of quantitative and qualitative methods and tools to adequately estimate uncertainties in the given context and to communicate them to scientific researchers, the clients (usually ministries), other actors in the policy process, and the broader public.

Commissioned by RIVM, the development of the guidance system started in September 2001, under the leadership of Dr. Jeroen van der Sluijs in close cooperation with RIVM and with a number of international uncertainty management specialists. In October 2001, an expert workshop was held to obtain input and feedback from the uncertainty management experts on a first sketch of the guidance system. This led to a draft version, in the form of a detailed questionnaire, which was subsequently presented to employees of the RIVM in a user workshop in November 2001. Though considered generally as a very thorough basis for uncertainty assessment, the detailed guidance document was judged by many of the users as being too comprehensive to be easily applicable in all cases. They preferred a shorter, pragmatic, easy-to-use version which could be applied at varying depths/levels, and which would offer specific hints and suggested actions on dealing with uncertainty. Therefore it was decided that in 2002 there would be developed a concise *mini-checklist* covering the major points in mapping and communicating uncertainties, as well as an associated *quickscan* version, which includes hints and preferred actions.

All this resulted in a suite of components (see figure 1), called the RIVM/MNP[1] Guidance for Uncertainty Assessment and Communication, denoted by Guidance for short in the sequel. The Guidance can be consulted in the various stages of the process by various users at a frequency and level which suits their individual needs best. For instance, *at the beginning* of a project, the guidance can play an important role in designing and elaborating the way uncertainty will be dealt with during the project; *during* a project, it can be of assistance in performing the uncertainty assessment and communicating the results; *after* a project, it can be of use in reviewing and evaluating the project. The group of intended users of the Guidance covers a large fraction of the employees of the Netherlands Environmental Assessment Agency[1] (RIVM/MNP) (e.g. project leaders, project-team members, researchers or policy advisers), as well as others (e.g. stakeholders involved in an extended peer-review of the project). Project leaders will typically use those components of the Guidance which are at a high level of aggregation (the *Mini-Checklist* and the *Quickscan*), while project-team members, researchers and policy advisers will more often also take up parts of the more detailed Guidance.

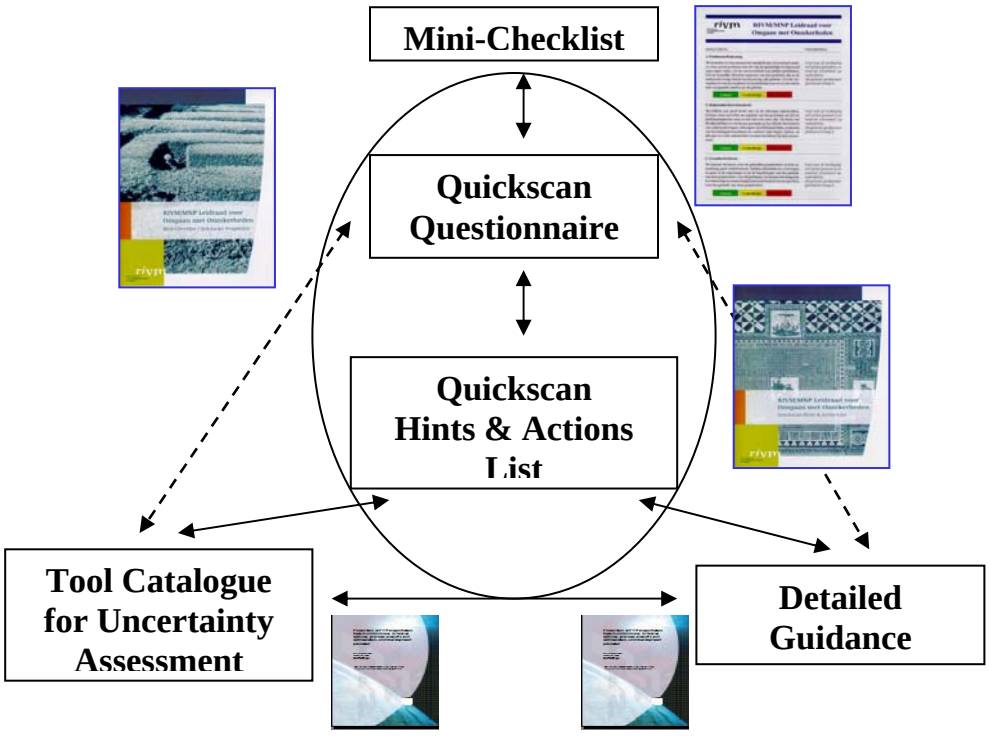


Figure 1. Structure of the RIVM/MNP Guidance for Uncertainty Assessment and Communication

The interrelationships between the components that constitute the Guidance is as follows (figure 1): The **Mini-Checklist** ([11])) is a back-of-the-envelope kind of tool which can also serve as a portal to the other components of the Guidance. By concisely presenting the potentially important issues in the various stages of the environmental assessment process - running from problem framing to reporting the results - it functions as a reminder list and instrument for reflection on the (desired) way of dealing with uncertainties and value-loadings in

[1] The Dutch name of the Netherlands Environmental Assessment Agency is 'Milieu- en Natuur Planbureau', abbreviated as MNP. It forms a part of the Netherlands National Institute for Public Health and the Environment (RIVM).

providing policy advice. It renders a brief account of the way uncertainty has been dealt with, and points to the **Quickscan Questionnaire** ([11]) if further elaboration is desired. This latter document (optionally) refers the user to the **Quickscan Hints & Actions List** ([12]) which has an advisory function and describes possible implications of the answers given to the questions in the Quickscan Questionnaire.

Some of the hints and actions point directly to parts of the **Detailed Guidance** ([13]) and the **Tool Catalogue for Uncertainty Assessment** ([14]). These documents can be considered as basic components underlying the complete Guidance. Although the Quickscan documents are presented as autonomous components which can be used in a separate mode, they are intrinsically related to these basic components. The 'Detailed Guidance' has been set up as an elaborate questionnaire for a deeper analysis of various aspects of dealing with uncertainty. It also contains a glossary of terms related to uncertainty assessment and communication. The associated 'Tool Catalogue for Uncertainty Assessment' offers information on different quantitative and qualitative methods and tools that can be utilized to assess uncertainties. The above documents are freely available on the Internet (see e.g. www.nusap.net).

The ultimate decision which components of the Guidance to use largely depends on the importance, the nature and the level of the uncertainties in the assessment concerned and on the resources available. The mini-checklist and the quickscan components are the parts of the Guidance that will be used most frequently. If one has the time and the mandate, then the detailed Guidance will be used to supplement and deepen the analysis. The mini-checklist and quickscan are discussed in the sequel.

3. STRUCTURE OF MINI-CHECKLIST AND QUICKSCAN DOCUMENTS

The mini-checklist concisely covers six central uncertainty-related themes in the environmental assessment process, including problem framing, stakeholder participation, selection of indicators, appraisal of the knowledge base, mapping and assessment of relevant uncertainties, reporting of the uncertainty information. It asks the user to reflect explicitly on how these issues are dealt with in the study at hand. The quickscan documents, consisting of a questionnaire and an associated hints and action list, elaborate this in more detail. In the sequel we will highlight point by point the six central themes addressed in these documents.

3.1 Problem Framing

In this stage the problem and its context and history are outlined, by identifying major issues, past work, the level of contention and the (expected) role of the assessment in the policy or decision making process. The user is explicitly asked to consider various views/perspectives on the problem, and to pay attention to the problem's interconnectedness with other problems. He/she is asked to be specific on what knowledge is needed with regard to the problem, and into which research questions this is translated. Possibly relevant aspects which are not dealt with in these research questions have to be indicated. Moreover it should be outlined what role the study is expected to play in the policy process, and what the relation is with previous studies on the subject (policy context and problem history).

3.2 Involvement of Stakeholders

This step concerns the identification of the main parties (stakeholders/actors) and their views and roles with respect to the problem, as well as the aspects of the problem about which they disagree. On the basis of all this, it has to be decided *if*, *how* (e.g. in formulating research questions, contributing information/data, evaluating findings/results), and *when* (at the beginning, during, after) one should involve *which* stakeholders in the study.

Important in this respect is a characterization of the problem at hand in terms of a number of features: level of dissensus on policy goals regarding the problem, type of knowledge needed, and the decision stakes and uncertainties involved. This characterization can lead to useful suggestions for stakeholder involvement, and can serve as an initial step in deciding on an appropriate level and form of participation for the current study. See Box 1 below.

Box 1: Problem characteristics and suggestions for stakeholder involvement

Problem characteristics	**Suggested stakeholder involvement**
There is dissensus about policy goals with respect to the problem and/or about the direction in which solutions need to be found	Involve the stakeholders or their views in defining and framing the problem and selecting the indicators; be explicit about the limited scope of the study and its results.
Decision stakes are high	Be transparent and open: let stakeholders 'take a look behind the scenes' (in all stages of the study); aim for a broad composition of the advisory panel; involve stakeholders in the review of the study.
There is dissensus about the (type of) knowledge required to solve the problem	Discuss and – if possible –use knowledge produced or put forward by stakeholders (including other research institutes); motivate the chosen approach (especially the choices about involving certain scientific disciplines) and state the potential limitations; signal and discuss the controversies with respect to the knowledge base, and account for deviating theories and approaches to the problem; provide for external review.
Major uncertainties exist regarding the behavior of the (natural and social) system(s) under study	If feasible, use knowledge and information produced or put forward by stakeholders, including knowledge and information derived from non-scientific sources, in order to be able to come up with a study of the required quality. When communicating intermediary and final results, be specific on the lack of knowledge and clearly state the consequences for the quality and the scope of the conclusions. Provide for external review or even counter-expertise. Deliver a clear mapping of the uncertainties.

Source: [11,12]

3.3 Selection of Indicators

In environmental assessments, the relevant features of the problem under study are typically expressed in terms of indicators or target variables. Selection of indicators is therefore an important step in shaping a study, and it is important to substantiate the final choices, discussing their shortcomings and associated controversies as well. This involves judging how well the selected indicators address key aspects of the problem as it has been framed, and how much support there is among scientists and within society (including decision-makers/politicians) for the use of these indicators for the problem at hand. Moreover, there should be an examination of how to deal with a potential lack of support, giving attention to differences in views and interest, and specifying what the consequences of these differences will be for the meaning and value of the study. Consider giving the stakeholders a role in defining or revising indicators.

3.4 Appraisal of Knowledge Base

This stage is concerned with answering the question of the adequacy of the available knowledge base for the assessment. It involves questions like: What quality criteria are relevant for answering the research questions? What knowledge and methods are needed to obtain answers of the required quality? What are the most important bottlenecks in the way of achieving this, in the light of existing controversies and weaknesses in the knowledge base? What will be the effect on the quality of the results, and which actions should be taken to clear these bottlenecks? In this way the user gets useful information for (re)shaping the study, in consultation with the client, and for adequately focusing the assessment and its reporting.

3.5 Mapping and Assessment of the Relevant Uncertainties

In this step the user is asked to identify the uncertainties most relevant to the problem, and to estimate what effort will be required to map these uncertainties adequately, providing information on their extent, nature (being it epistemic or stochastic) and location. Moreover the possible consequences of the uncertainties for the conclusions of the study have to be indicated, and an indication should be given on how to assess the most important uncertainties and their consequences, within the limitations of the available resources (time, money, people, expertise, etc.). The actual planning and performing of the uncertainty assessment completes this step.

Box 2: Aspects of uncertainty requiring additional attention

Problem characteristics	**Uncertainty aspects requiring special attention**
Various assumptions are critical	Be explicit about points of departure, assumptions and framing of the study; evaluate the critical choices made and discuss their consequences for the robustness of the policy-relevant conclusions.
The estimate of an indicator is close to a (legal) norm or (policy) target for that indicator	Pay additional attention to uncertainties for the indicator(s) concerned: • indicate the nature of the uncertainties, e.g., uncertainty due to limited knowledge or due to intrinsic variability (in nature, human behavior or social systems); • give attention to how these uncertainties can be translated in terms of accomplishing/not accomplishing policy goals, or exceeding/not exceeding norms, and to the potential size and seriousness of effects and risks; • investigate the possibilities to reduce (policy-relevant) uncertainty, and discuss these
A small change in an indicator estimate may have a significant influence on estimated costs, impacts or risks	
There is dissensus about policy goals	Pay additional attention to the role of value-laden uncertainties and stakeholder views and interests. Discuss the implications of uncertainties for the socio-political context/arena.
Decision stakes are high	Pay additional attention to the influence of views and values on the selection of indicators and on the conclusions. Discuss the implications of uncertainties for the socio-political context/arena.
There is dissensus about the (type of) knowledge required to solve the problem	Pay additional attention to the issues where the points of view differ most with respect to the (type of) knowledge required, and discuss the effects on the conclusions.
Major uncertainties exist regarding the behavior of the (natural and social) system(s) under study	Pay additional attention to the consequences of this uncertainty for the conclusions. Be explicit about ignorance and controversies, and about what these mean for the conclusions.
The assessment method used has typical uncertainties associated with it, which require additional attention	Determine which specific uncertainties are associated with the chosen assessment method (measurements, models, scenarios, expert judgment).

Source: [11,12]

To support the user in these tasks various hints and tools are provided. For example, Box 2 lists a number of triggers which point at policy-relevant uncertainties requiring additional attention. For identifying the most important uncertainties, the uncertainty matrix presented in table 1 can be used; see Box 3 for more background information. A tool catalogue is provided ([14]) to assist the user in choosing appropriate methods for dealing with the identified uncertainties. In this document comprehensive information is given on various quantitative and qualitative uncertainty assessment techniques (global sensitivity analysis, NUSAP, expert elicitation, scenario analysis, model quality assessment etc.). The presented information concerns a brief description of the specific technique and its goals, strengths and limitations, required resources, as well as guidelines for its use and warnings for typical pitfalls. It is supplemented by references to handbooks, software, example case studies, websites, experts etc. The tool catalogue is a 'living document', which will be made available on the web in the future, and to which descriptions of additional tools can be added.

Location ↓		**Level of uncertainty** *(from determinism, through probability and possibility, to ignorance)*			**Nature of uncertainty**		**Qualification of knowledge base**			**Value-ladenness of choices**		
		Statistical uncertainty	**Scenario uncertainty**	**Recognized ignorance**	**Epistemic**	**Variability**	**–**	**0**	**+**	**–**	**0**	**+**
Context												
Expert judgment												
Model	**structure**											
	implementation											
	parameters											
	inputs											
Data												
Outputs												

Table 1: Uncertainty Matrix (for explanation, see Box 3)

3.6 Reporting of Uncertainty Information

Reporting of uncertainty information preferably takes place during the whole environmental assessment process, not only at the final delivery of results. In this communication it is necessary to be aware of (i) the context of the reporting (why, to whom, on behalf of whom, when, where) and (ii) the robustness of the main messages for uncertainties in the knowledge base and for any deviations from the employed assumptions and choices. Reporting on the policy-relevant uncertainties and their possible consequences for policy making, politics and society should take place in a clear manner, tailored to the intended audience(s). In written reporting, the results should be presented in a balanced and consistent way, providing a traceable account and adequate backing of the presented material. For example, this can be achieved by offering the information in a step-wise fashion. Taking account of the fact that readers often scan a text selectively it is recommended to present important uncertainty information explicitly at strategic points, e.g. in the introduction, conclusions, summary, text-boxes. See also Box 4.

Box 3: The Uncertainty Matrix (see also Table 1)

A central element in the 'mapping and assessment' phase is the use of an extended version of the uncertainty matrix, originally proposed in Walker et al. [15]. This matrix is a heuristic device for classifying and reporting the various dimensions of uncertainty, and to improve communication among analysts as well as between them and policymakers and stakeholders. It is based on the uncertainty typology presented in [15], which classifies uncertainties according to three dimensions: their *location* (where they occur), their *level* (where uncertainty manifests itself on the gradual spectrum between deterministic knowledge and total ignorance) and their *nature* (whether uncertainty primarily stems from knowledge imperfection (epistemic uncertainty) or is a direct consequence from inherent variability/stochasticity). By explicitly adding two additional columns in the matrix denoted '*qualification of knowledge base'* and *'value-ladenness of choices*', we have extended the original uncertainty typology with two dimensions, see Table 1. The category 'qualification of knowledge' refers to the level of underpinning and backing of the information (e.g. data, theories, models, methods, argumentation etc.) involved in the assessment of the problem; it points at the methodological acceptability and the rigour and strength of the employed methods, knowledge and information, and thus it characterizes to a certain extent their (un)reliability. If desired, a so-called pedigree-analysis can be employed to assess the level of underpinning in a semi-quantitative way on basis of a number of quality criteria ([14,16]). The second additional category ('value-ladenness of choices') refers to the inevitable presence of values and biases in the various choices and assumptions involved. This concerns choices and assumptions regarding the way the scientific questions are framed, data are selected, interpreted and rejected, methodologies and models are devised and used, explanations and conclusions are formulated etc.
Both added dimensions characterize important features which directly relate to uncertainty: If underpinning is weak, this indicates that the statement of concern is surrounded by much (knowledge-related) uncertainty. If value-ladenness is high for relevant parts of the assessment, then it is imperative to analyze whether or not the results of the study are highly influenced by the choices and assumptions involved, and whether this could lead to a certain arbitrariness, ambiguity or uncertainty of the policy-relevant conclusions. This could then be a reason to explicitly deal with different views and perspectives in the assessment and to discuss the scope and robustness of the conclusions in an explicit manner.
Source: [12]

4. DISCUSSION

The foregoing illustrates that the guidance system provides structure to the task of uncertainty management and can be employed in a flexible way. The Guidance stimulates reflection and deliberation on how uncertainties are (to be) handled and communicated effectively and helps to avoid pitfalls. Tools for uncertainty assessment are made more easily available, and can be selected in a more tailored manner, on basis of problem characteristics. Although the guidance system was initially developed in the context of environmental assessments, it can be applied in other application areas as well, with some minor adaptations as appropriate.

For its success, some conditions are essential. Firstly, the commitment of the higher management is crucial, since time and effort spent on dealing with uncertainties must be considered as relevant. By its primary focus on the policy-relevance of uncertainties, the Guidance puts this issue explicitly to the fore. Secondly, it would certainly help if the use of the Guidance was included as standard activity in the prevailing project management. At RIVM/MNP this has been established by including the Guidance formally in the quality assurance system. Thirdly, its further application and institutionalization will require motivating and training the potential users of the Guidance, showing them the benefits of its use. We are currently developing training sessions for employees of the RIVM/MNP in the use of the Guidance. To support the introduction in daily practice, a web-based version of the Guidance is under construction, providing quick and easy access to its various parts. We hope to stimulate this introduction process further by providing appropriate uncertainty assessment tools, and by gradually building an expertise network and a 'good-practice examples' data-base in using the Guidance.

In fact, we currently are only at the start of implementing the presented systematic ideas of uncertainty management in an institutional setting. The above-mentioned activities must therefore be considered as first steps. In due time - after its fuller implementation - the use of the Guidance will be evaluated, leading possibly to further adaptations. For the time being we can already identify two major issues which will deserve future attention: The first one refers to the further deployment and development of the tool catalogue with methods for: (i) propagating and analyzing qualitative and semi-quantitative uncertainty information (e.g. concerning value-loadings, assumptions, pedigree scores), especially in expert-reasoning and model-based calculation 'chains'; and for (ii) synthesizing qualitative and (semi)-quantitative uncertainty information ([16]). The second issue concerns the analysis of various contexts of science, policy and society interactions in order to find suitable arrangements and forms for knowledge production and uncertainty communication; thus enhancing the effective use of science for policy or society.

Notwithstanding that there is still a long way to go, one can consider the Guidance - with its specific focus on problem context and socio-political embedding, accountability, transparency and reflexivity, participation and extended peer review - as a useful contribution towards new social practice of science in a postmodern era, as exemplified by e.g. the post-normal science and mode 2 science paradigms ([4-10]).

Box 4: Reporting of Uncertainty Information

Communicating and reporting about uncertainty entails a number of issues that should be taken into consideration. The RIVM/MNP Guidance discerns (1) context of communication of uncertainty; (2) target audiences; (3) language; (4) methods; (5) format and (6) content.

With regard to *context* authors have to ask themselves why the uncertainty is being reported (e.g. political purpose, scientific purposes, required by legislation, requested by stakeholders), and at which stage, and what setting (e.g. report, meeting, press article, internet, scientific journal). This will influence the scope of the reporting.

The *target audiences* may stretch out over the stakeholders for the problem of concern. Although the target audience might not correspond to the whole set of stakeholders, it is surely a subset of those. The type of audience will determine amongst other things the 'language' of the communication/report as well as the main messages of interest. Since the audience can be quite diverse or disparate, clear and transparent communication of the results is required, but misinterpretations can not always be avoided.

The *language* used in the communication and reporting of uncertainty is one of the most important issues. Careful design of communication and reporting should be done in order to avoid information divide, misunderstandings, and misinterpretations. The communication of uncertainty should be understandable by the audience. There should be clear guidelines to facilitate clear and consistent use of terms provided. Values should be made explicit in the reporting process. Potential ambiguity in the wording of the report or in use of metaphors should better be avoided.

The *method* used to manage uncertainty (quantitative sensitivity and uncertainty analysis, quality assurance (e.g. NUSAP, pedigree analysis) etc.) and hence, the type of information generated, is a crucial aspect of communicating and reporting uncertainty and should be described. Uncertainty methods can operate in the foreground when applied explicitly to produce information on uncertainty (e.g. written material, graphs), or in the background as when run 'behind' a model and results are embedded in the output (e.g. model outputs, scenarios).

A variety of different reporting *formats* and media can be used (numbers, words, narratives, graphs, pictures, multimedia, internet). No one format is more valid than others. The choice of format depends on communication settings, type of audience, and uncertainty management methods.

With regard to *content* one could think of making explicit the major assumptions on which the main messages are based, discussing the robustness of the major conclusions in the light of these assumptions as well as of uncertainties in the underlying knowledge base. Moreover important areas of ignorance and controversies should be stated explicitly, giving background on how these issues have been dealt with, and what this means for the main conclusions. If considered relevant for the given context, clear information could be given on the nature and causes of policy-relevant uncertainties and on the potential consequences for policy, politics and society (e.g. in terms of effects and risks). Indicate – if considered policy relevant – what can and can not be done about these uncertainties, and which uncertainty aspects deserve additional attention in the future.

Source: [11,12,13]

ACKNOWLEDGEMENTS

We like to acknowledge the help of many colleagues from RIVM/MNP as well as many uncertainty experts who have aided in the development of the Guidance by written contributions or by discussions, comments and reviews.

REFERENCES

1. M.B.A. van Asselt. *Perspectives on Uncertainty and Risk : The PRIMA Approach to Decision Support.* Kluwer Academic Publishers ,Dordrecht, The Netherlands, 2000.
2. J. P. van der Sluijs. A way out of the credibility crisis of models used in integrated environmental assessment. *Futures* 34:133–146, 2002.
3. R.A. Pielke, Jr. Science policy: Policy, politics and perspective. *Nature* 416:367-368, 2002.
4. S.O. Funtowicz and J. R. Ravetz. *Uncertainty and quality in science for policy.* Kluwer Academic, Dordrecht, The Netherlands, 1990.
5. S.O. Funtowicz and J. R. Ravetz. Science for the Post-Normal Age. *Futures* 25:735-755, 1993.
6. S.O. Funtowicz and J. R. Ravetz. Post-Normal Science - an insight now maturing, *Futures* 31: 641-646, 1999.
7. M. Gibbons, C. Limoges, H. Nowotny, S. Schwartzman, P. Scott and M. Trow. *The New Production of Knowledge.The Dynamics of Science and Research in Contemporary Societies.* Sage, London, 1994.
8. M. Gibbons, Science's new social contract with society. *Nature* 402:C81-C84, 1999.
9. H. Nowotny, P. Scott, and M. Gibbons. *Re-Thinking Science. Knowledge and the Public in an Age of Uncertainty*. Polity Press, Cambridge, 2001.
10. H. Nowotny, P. Scott, and M. Gibbons. Mode 2 Revisited: The New Production of Knowledge, *Minerva*, forthcoming, 2003.
11. A.C. Petersen, P.H.M. Janssen, J.P. van der Sluijs, J.S. Risbey, J.R. Ravetz *RIVM/MNP Guidance for Uncertainty Assessment and Communication: Mini-Checklist & Quickscan Questionairre*. RIVM/MNP; ISBN 90-6960-105-1, 2003.
12. P.H.M. Janssen, A.C. Petersen, J.P. van der Sluijs, J.S. Risbey, J.R. Ravetz . *RIVM/MNP Guidance for Uncertainty Assessment and Communication: Quickscan Hints & Actions List*. RIVM/MNP, ISBN 90-6960-105-2, 2003.
13. J.P. van der Sluijs, J.S. Risbey, P. Kloprogge, J.R. Ravetz, S.O. Funtowicz, S.Corral Quintana, Â Guimarães Pereira, B. De Marchi, A.C. Petersen, P.H.M. Janssen, R. Hoppe, and S.W.F. Huijs. *RIVM/MNP Guidance for Uncertainty Assessment and Communication: Detailed Guidance.* Utrecht University, 2003.
14. J.P. van der Sluijs, P.H.M. Janssen, A.C. Petersen, P. Kloprogge, J.S. Risbey, W. Tuinstra, M.B.A. van Asselt, J.R. Ravetz, *RIVM/MNP Guidance for Uncertainty Assessment and Communication: Tool Catalogue for Uncertainty Assessment.* Utrecht University, (forthcoming), 2004.
15. W.E Walker,.P., Harremoës, J. Rotmans, J.P. van der Sluijs, M.B.A. van Asselt., P. Janssen, M.P. Krayer von Krauss: Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integrated Assessment*, Vol.4, No.1, pp. 5-17, 2003.
16. J.P. van der Sluijs, P Kloprogge, J. Risbey, J. Ravetz. Towards a synthesis of qualitative and quantitative uncertainty assessment: applications of the Numeral, Unit, Spread, Assessment, Pedigree (NUSAP) system. Paper for the *International Workshop on Uncertainty, Sensitivity, and Parameter Estimation for Multimedia Environmental Modeling*, August 19-21, 2003, Rockville, Maryland.

Probability Bounds Analysis Is a Global Sensitivity Analysis

S. Ferson, J. Hajagos and W.T. Tucker

Applied Biomathematics, 100 North Country Road, Setauket, New York 11733, USA
E-mail: scott@ramas.com, janos@ramas.com, troy@ramas.com

Abstract: Probability bounds analysis provides analysts a convenient means to characterize the neighborhood of possible results that would be obtained in probabilistic calculations of plausible alternative inputs. For this reason, it constitutes a method of global sensitivity analysis that does not require any notion of decomposing or partitioning total uncertainty. We show the relationship between probability bounds analysis and the methods of interval analysis and probabilistic sensitivity analysis from which it is jointly derived, and indicate how the method can be used to assess the quality of probabilistic models such as those developed in Monte Carlo simulations for risk analyses. We also illustrate how a meta-level sensitivity analysis can be conducted within a probability bounds analysis by pinching inputs to precise distributions or real values.

Keywords: probability bounds analysis, interval analysis, second-order probability, sensitivity analysis, convolution, robust Bayes, meta-level sensitivity analysis

1. INTRODUCTION

Sensitivity analysis is the general term for quantitative study of how the inputs to a model influence the results of the model. Sensitivity analysis has many manifestations in probabilistic risk analyses and there are many disparate approaches based on various measures of influence and response. Sensitivity analyses are conducted for fundamentally two reasons: to understand the reliability of the conclusions and inferences drawn from an assessment, and to focus future empirical studies so that effort might be expended to improve estimates of inputs that would lead to the most improvement in the estimates of the outputs. Because of the obvious and fundamental importance of sensitivity analyses in calculations, there has been a confluence of ideas to this issue from disparate analytical disciplines.

Leamer [1] defined global sensitivity analysis as a systematic study in which "a neighborhood of alternative assumptions is selected and the corresponding interval of inferences is identified". There are two disparate ways to effect such a study. One natural way is to bound the neighborhood with interval ranges. Another natural way is to ascribe a probability distribution to the elements in the neighborhood. Consider, for example, the context of a deterministic calculation. When the model involves uncertainty about the real-valued quantities used in the calculation, the definition of global sensitivity analysis is equivalent to that of interval analysis [2,3,4,5]. Probability theory, implemented perhaps by Monte Carlo simulation, can also be viewed as a global sensitivity analysis of a deterministic calculation in that it yields a distribution describing the probability of alternative possible values about a point estimate [6,7,8,9]. In the figure below these two possible paths are shown as right and left downward arrows respectively.

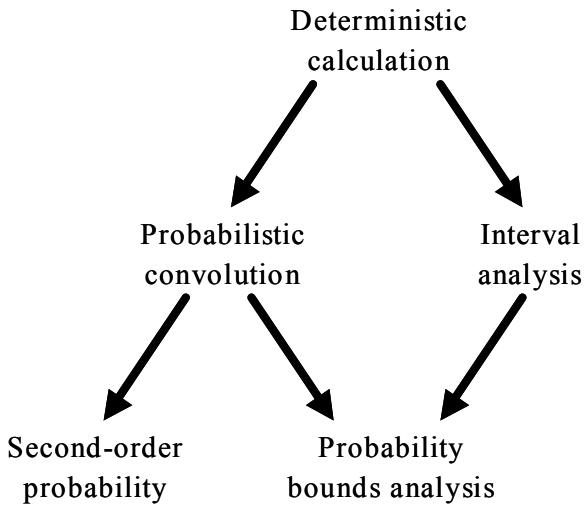


Figure 1. Relationships among different calculation strategies. Arrows represent generalizations.

Of course, the calculations on which it might be desirable to conduct sensitivity analyses are not all deterministic. In fact, many of them are already probabilistic, as is the case in most modern risk analyses and safety assessments. One could construct a probabilistic sensitivity analysis of a probabilistic calculation. The resulting analysis would be a second-order probabilistic assessment. However, such studies are often difficult to conduct because of the large number of calculations that are required. It is also sometimes difficult to visualize the results in a way that is easily comprehensible. Alternatively, one could apply bounding arguments to the probabilistic calculation and arrive at interval versions of probability distributions. We call such calculations "probability bounds analysis" (PBA) [10,11,12]. This approach represents the uncertainty about a probability distribution by the set of cumulative distribution functions lying entirely within a pair of bounding distributions called a "probability box" or a "p-box". Probability bounds analysis is a global sensitivity analysis of a probabilistic calculation because it defines neighborhoods of probability distributions (i.e., the p-boxes) that represent the uncertainty about imperfectly known input distributions and projects this uncertainty through the model to identify a neighborhood of answers (also characterized by a p-box) in a way that guarantees the resulting bounds will entirely enclose the cumulative distribution function of the output. A probability distribution is to a p-box the same way a real scalar number is to an interval. The bounding distributions of the p-box enclose all possible distributions in the same way that the endpoints of the interval circumscribe the possible real values.

Probability bounds analysis is related to other forms of uncertainty analysis. It is a marriage of probability theory and interval analysis that generalizes and is faithful to both traditions. As depicted in Figure 1, PBA can arise either by bounding probability distributions (the left path down to PBA) or by forming probability distributions of intervals (the right path). The advantage of this marriage is that variability (aleatory uncertainty) and incertitude (epistemic uncertainty) are treated separately and propagated differently so each maintains its own character. PBA is a comprehensive global sensitivity analysis that is an

alternative to complicated second-order or nested Monte Carlo methods. PBA is very similar in spirit to Bayesian sensitivity analysis (which is also known as robust Bayes [13]), although the former concerns arithmetic and convolutions, and the latter addresses the issues of updating and aggregation. Unlike Bayesian sensitivity analysis, probability bounds analysis is always easy to employ because it does not depend on the use of conjugate pairs to make calculations simple. PBA is a practical approach to computing with imprecise probabilities [14]. Like a Bayesian sensitivity analysis, imprecise probabilities are represented by a class of distribution functions. PBA is simpler because it defines the class solely by reference to two bounding distributions. (It therefore cannot fully represent a situation in which there are intermediate distributions lying within the bounds that are excluded from the class. In the context of risk and safety assessments, however, this is rarely a significant drawback.)

2. PBA CIRCUMSCRIBES POSSIBLE DISTRIBUTIONS GIVEN UNCERTAINTY

PBA can produce rigorous bounds around the output distribution from an assessment. These bounds enclose all the possible distributions that could actually arise given what is known and what is not known about the model and its inputs. Because it is based on the idea of bounding rather than approximation, it provides an estimate of its own reliability [15,16, cf. 17]. Probability bounds analysis can comprehensively account for possible deviations in assessment results arising from uncertainty about

- distribution parameters,
- distribution shape or family,
- intervariable dependence, and even
- model structure.

Moreover, it can handle all of these kinds of uncertainties in a single calculation that gives a simple and rigorous characterization of how different the result could be given all of the professed uncertainty. The requisite computations used in PBA are actually quite simple and have been implemented in straightforward algorithms [18,19,15,16,11,20]. The computations are generally much faster than even simple Monte Carlo convolution and vastly faster than a numerically intensive sensitivity analysis with traditional methods [21,22,23,24,6,7,25,8].

Probability bounds analysis is useful whenever the uncertainty about the marginal distributions can be characterized by interval bounds about their cumulative distribution functions. These bounds can be specified using empirical information available about each distribution. For instance, if the parameters of a normal distribution can be given within interval ranges, best-possible bounds on the distribution are easy to construct. If the shape of the underlying distribution is not known, but some statistics such as the mean, mode, variance, etc. can be specified (or given as intervals), rigorous bounds can generally be constructed that are guaranteed to enclose the true distribution subject to the given constraints. Often these bounds will be optimally narrow given the stated information. The resulting p-boxes are distribution-free in the sense that they make no assumptions whatever about the distribution family (whether it is normal, lognormal, Weibull, etc.). Such bounds on distributions can then be combined according to the calculations in the assessment. Currently, software is available to handle (i) arithmetic convolutions (addition, multiplication, minimum, etc.), (ii) magnitude comparisons (greater than, less than), (iii) logical operations (conjunction, disjunction, etc.), and (iv) transformations (logarithm, exponentiation, roots, etc.).

It is also possible to handle uncertainty about the dependencies among variables in a model. Recent algorithmic developments permit uncertainty about the dependencies among variables to be propagated through the calculations of a probabilistic assessment [26]. A pair-wise dependency may be modeled with any of the following six assumptions:

(i) independence,
(ii) comonotonicity (maximal correlation),
(iii) countermonotonicity (minimal correlation),
(iv) linear relationship and correlation within a specified interval,
(v) linear relationship with unknown correlation,
(vi) signed (positive or negative) but otherwise unknown dependency, and
(vii) unknown dependency (including any nonlinear relationship).

For the first three cases, a convolution between two probability distributions yields a well defined probability distribution. For the latter four cases, the results are given as bounds on a (cumulative) distribution function. For each binary operation, the bounds obtained are generally the best possible bounds, i.e., they could not be any narrower yet still contain all the possible distributions permissible under the assumption.

Unlike approaches based on conventional Monte Carlo simulation, the algorithms employed for these operations yield rigorous answers that lack sampling error. In fact, the results are exact at each point of discretization, of which there may be arbitrarily many. The results are guaranteed to enclose the true distributions. Although it is straightforward to ensure that bounds remain rigorous (sure to contain the true distributions) in sequential calculations, the best possible nature of the bounds may be lost in some complication calculations. Maintaining the optimality of the bounds is, in general, a computationally challenging task that can require other methods [14]. Nevertheless, the methods of probability bounds analysis developed over the last two decades provide risk and safety analysts a practical and convenient means to conduct comprehensive sensitivity analyses on their calculations.

3. META-LEVEL SENSITIVITY ANALYSES

As outlined above, probability bounds analysis is a kind of sensitivity analysis with considerable comprehensiveness. It is possible and sometimes of interest to perform a sensitivity analysis on the results of an assessment conducted with PBA. This would, of course, constitute a meta-level sensitivity analysis. This section explores the use of pinching studies that hypothetically assess the impact on result uncertainty of additional empirical knowledge.

One of the fundamental purposes of sensitivity studies is to learn where focusing future empirical efforts would be most productive. This purpose requires estimating the value of additional empirical information. Of course, the value of information not yet observed cannot be measured, but it can perhaps be predicted. One strategy to this end is to assess how much less uncertainty the calculations would have if extra knowledge about an input were available. This might be done by comparing the uncertainty before and after "pinching" an input, i.e., replacing it with a value without uncertainty. Of course, one does not generally know the correct value without uncertainty, so this replacement must be conjectural in nature. To pinch a parameter means to hypothetically reduce its uncertainty for the purpose of the thought-experiment. The experiment asks what would happen if there were less uncertainty about this

number. Quantifying this effect amounts to measuring the contribution by the input to the overall uncertainty in a calculation.

The estimate of the value of information for a parameter will depend on how much uncertainty is present in the parameter, and how it affects the uncertainty in this final result. The sensitivity could be computed with an expression like

$$100\left(1-\frac{\text{unc}(T)}{\text{unc}(B)}\right)\%$$

where B is the base value of the risk expression, T is the value of the risk expression computed with an input pinched, and unc( ) is a measure of the uncertainty of the answer. The result is an estimate of the value of additional empirical information about the input in terms of the percent reduction in uncertainty that might be achieved in the expression when the input parameter is replaced by a better estimate obtained from future empirical study. The pinching can be applied to each input quantity in turn and the results used to rank the inputs in terms of their sensitivities. (Note that these reductions will not generally add up to 100% for all the input variables.) In principle, one could also pinch multiple inputs simultaneously to study interactions.

There are multiple possible ways to define unc( ) to measure uncertainty. In the context of probability bounds analysis, one obvious measure is the area between the upper and lower bounds of the p-box. As the p-box approaches a precise probability distribution where all epistemic incertitude has evaporated and only the natural variability remains, this area approaches zero. An analyst might also elect to define unc( ) as some measure of dispersion or perhaps the heaviness of the tails [27] of the p-box. Using different measures will obviously allow the analyst to address different questions in a sensitivity analysis. If the measure of uncertainty is a scalar quantity (i.e., a real number), then the sensitivities that come from the analysis will also be scalars and can be ordered.

There are also multiple possible ways to pinch uncertainty. Pinching in different ways can result in strongly different estimates of the overall value of information. Several strategies are possible in estimating sensitivities from comparative PBA assessments:

(i) replace an input with a point value,
(ii) replace an input with a precise distribution function, or
(iii) replace an input with a zero-variance interval.

Replacing a p-box with a precise probability distribution would be pinching away the incertitude about the distribution. Replacing a p-box or a distribution function with a point value would be pinching away both the incertitude and the variability of the quantity. For inputs that are known to be variable (variance greater than zero), such a pinching is counterfactual, but it may nevertheless be informative. In particular, in may be especially useful in planning remediating strategies. In some situations, it may be reasonable to replace a p-box with a p-box shaped like an interval but prescribed to have a variance of zero. The effect of this would be to pinch away the variability but leave uncertainty. Such a replacement might be reasonable for p-boxes having a core (a region along the abscissa for which the upper bound of the p-box is one and lower bound is zero).

This approach of pinching inputs and recalculating the assessment is not unfamiliar to Monte Carlo analysts. Many routinely conduct sensitivity studies of the proportional contribution of variability in each variable to the overall variability in the calculated risk distribution. To determine the effects of variability in a Monte Carlo simulation using this method, each variable containing variability (i.e., expressed as a probability distribution) is reduced in turn to its mean or other appropriate point value, and the simulation is repeated. The measure of sensitivity is often the proportional effect of variability in each variable on the model, which is computed as the variance in the risk distribution from each of the simulations divided by the variance in the risk distribution from the base model result. Although the general idea of pinching is known to Monte Carlo analysts, the notions of pinching to a precise distribution and pinching to a zero-variance interval has no analog in Monte Carlo sensitivity analyses.

Figure 2 shows a numerical example of pinching to a precise distribution. The top panel of the figure depicts of addition of two p-boxes *a* and *b* (assuming independence). This is the "base case" against which the pinchings will be compared. The area between the upper and lower bounds for the sum $a+b$ is 2.12. The middle panel of the figure shows the first pinching. The p-box *a* is replaced with a precise probability distribution that lies entirely within the p-box. When a distribution replaces the p-box in the addition with *b* (which is still the same p-box), the result is the p-box shown at the far right on the middle panel. This p-box has an area of about 1.14. The percentage reduction in this area compared to that of the p-box for the sum shown on the top panel is 46.24%. This percent, which labels the sum on the middle panel, represents the sensitivity measure for pinching the variable *a* to a precise probability distribution. The bottom panel of Figure 2 shows the reduction of uncertainty (area) for the sum $a+b$ from pinching the p-box for *b* to a precise distribution. Compared to the base case in the top panel, the area is reduced by 47.17%. In this case, the potential reduction in uncertainty from additional information about *a* and *b* are roughly the same.

Figure 3 shows a similar set of sensitivity analyses based on pinching p-boxes to precise distribution functions. The calculation for the base case in this figure (shown in the top panel) was made without making any assumption about the dependence between the variables *a* and *b*. For this reason, even though the p-boxes for the variables *a* and *b* are just the same as were used in Figure 2, the area of the sum grows to about 3.05. The second panel of Figure 3 depicts pinching the p-box for the variable *a* to a precise distribution and its consequence for the resulting uncertainty about the sum. The third panel likewise shows the pinching for variable *b*. Both panels are annotated with the percent reduction in the area of the p-box for the sum compared to the base case in the top panel. The reduction in uncertainty from pinching the variable *a* in this situation is perhaps surprisingly small. The sensitivity to changing *b* is more than three times greater than that of *a*. The bottom panel shows the effect of pinching the dependence from the Fréchet case of assuming nothing about dependence to assuming independence. (The pinching could have specified any particular dependence.)

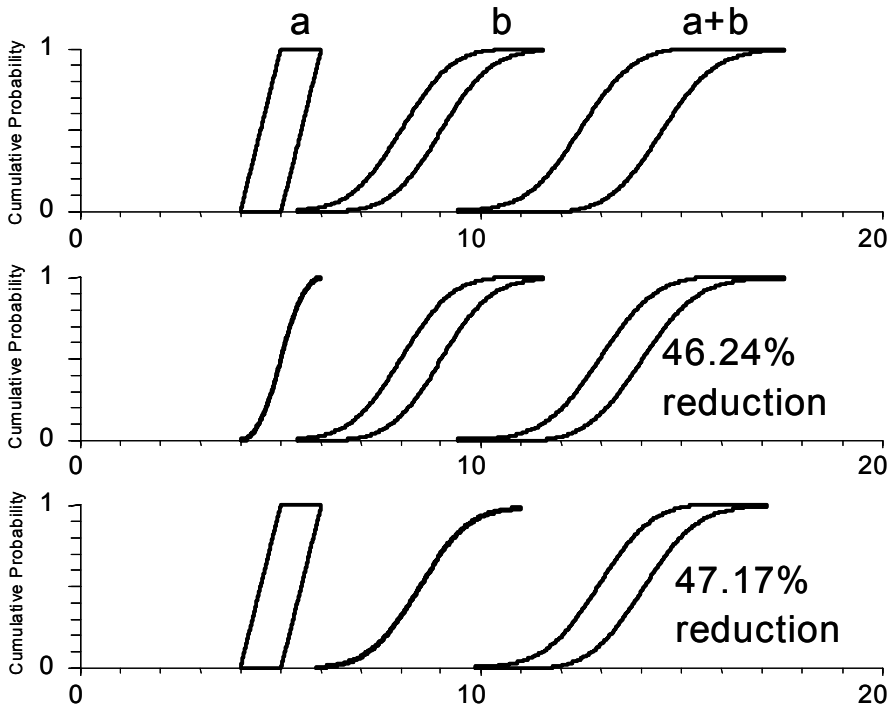


Figure 2. Meta-sensitivity analysis by pinching a p-box to a precise distribution.

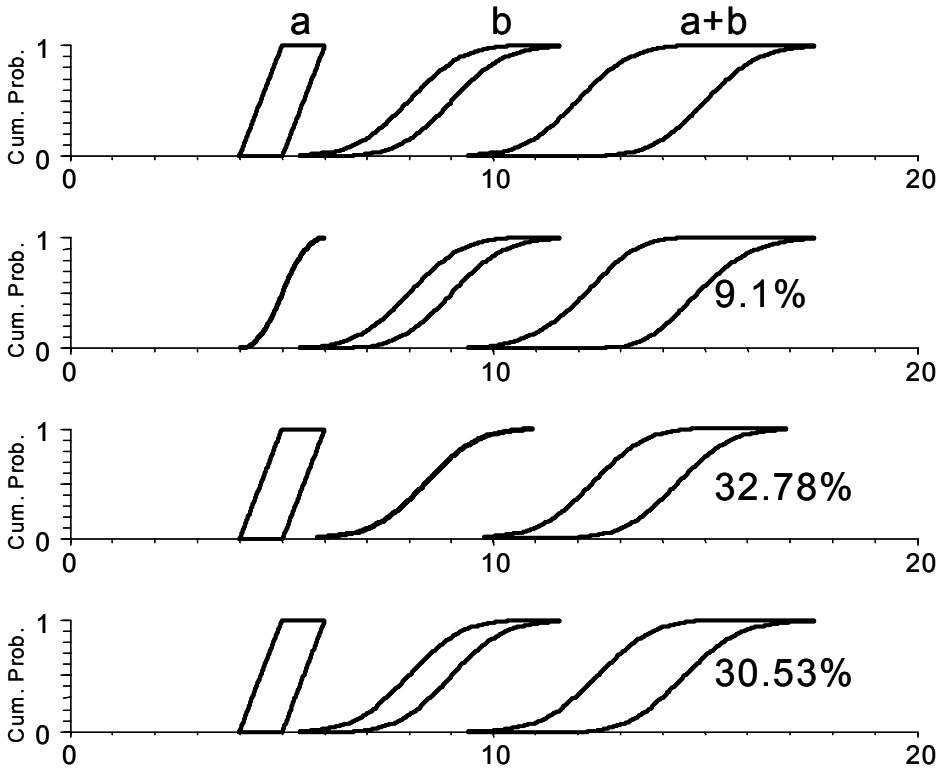


Figure 3. Meta-sensitivity analysis for the Fréchet case without dependence assumptions.

Figure 4 shows a third hypothetical sensitivity study. The base case in the top panel is identical to the base case shown in Figure 2, but in this study, the p-boxes are pinched to scalar values. The second and third panels of Figure 4 depict the additions resulting from pinching one of the addends to a point value. The observed percentage reduction in the area of each resulting sum compared to the base case is shown beside its p-box. What would the reductions in uncertainty have been if the base calculation had not assumed independence? The pinchings would have yielded exactly the same results, simply because dependence assumptions have no effect when either of the addends is a point. Thus, the lower two panels of Figure 4 would look exactly the same. However, if the base calculation had not assumed independence, then the base uncertainty about the sum $a+b$ would have been slightly greater (area = 3.05, compared to 2.12 under independence). That would make the rightmost p-box in the top panel of Figure 4 noticeably wider. Therefore the reductions in uncertainty by pinching to a point would have been somewhat greater than they were for the independent case. Instead of 50.4% and 52.36% reductions, pinching the variables a and b to points under no assumption about dependence would have respectively yielded 65.54% and 66.9% reductions in uncertainty as measured by the area within the resulting p-boxes.

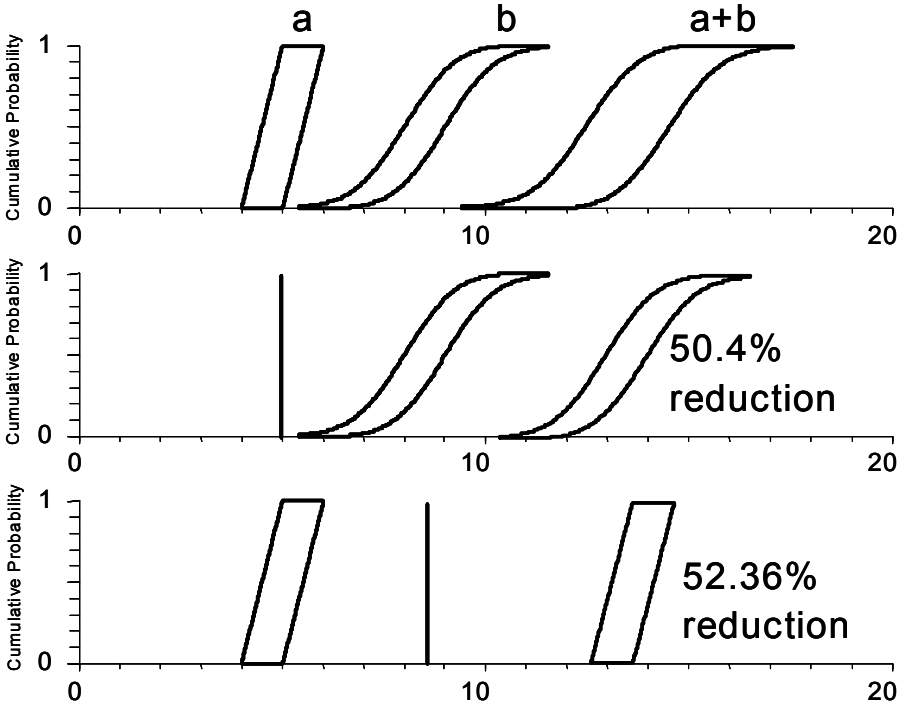


Figure 4. Meta-sensitivity analysis by pinching a p-box to a point value.

4. CONCLUSIONS

Many probabilistic assessment conducted using Monte Carlo simulations employ what-if sensitivity studies to explore the possible impact on the assessment results of varying the inputs. For instance, the effect of the truncation of some variable might be explored by re-running the model with various truncation settings, and observing the effect on the risk

estimate. The effect of particular parameter and probability distribution choices, and assumptions regarding dependencies between variables can also be examined in this way. Model uncertainty can be probed by running simulations using different models. However, such studies are often very difficult to conduct because of the large number of calculations that are required. While informative, this approach is rarely comprehensive because when there are multiple uncertainties at issue (as there usually are), the shear factorial problem of computing all of the possible combinations becomes prohibitive. Usually, in practice, only a relatively tiny number of such analyses can be performed. Probability bounds analysis can be used to automate such what-if sensitivity studies and vastly increase their comprehensiveness.

Sensitivity analysis can also be conducted at a meta-level by hypothetically replacing a p-box in a probability bounds analysis with a precise distribution or perhaps a scalar number to evaluate the potential reduction of uncertainty of the result under additional knowledge.

ACKNOWLEDGMENTS

We thank Cliff Joslyn of Los Alamos National Laboratory for helpful discussions. This work was supported in part by a Small Business Innovation Research grant (1R43ES10511) to Applied Biomathematics from the National Institutes of Environmental Health Sciences (NIEHS), a component of the National Institutes of Health (NIH), and by contract 19094 with Sandia National Laboratories (SNL), under the Epistemic Uncertainty project directed by William Oberkampf. Any opinions, findings, conclusions or recommendations expressed herein are solely those of the authors and do not necessarily reflect the views of the NIEHS, NIH or SNL.

REFERENCES

1. Leamer, E.E. 1990. Sensitivity analysis would help. *Modelling Economic Series*, edited by C.W.J. Granger, Clarendon Press, Oxford.
2. Dwyer, P. 1951. *Linear Computations*. John Wiley, New York.
3. Moore, R.E. 1966. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, New Jersey.
4. Alefeld, G. and J. Herzberger. 1983. *Introduction to Interval Computations.* Academic Press, New York.
5. Neumaier, A. 1990. Interval Methods for Systems of Equations. Cambridge University Press, Cambridge.
6. Iman, R.L. and J.C. Helton. 1985. A Comparison of Uncertainty and Sensitivity Analysis Techniques for Computer Models. NUREG/CR-3904. National Technical Information Service, Springfield, Virginia.
7. Morgan, M.G. and M. Henrion. 1990. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press, Cambridge, England.
8. Helton, J.C. and F.J. Davis 2000. *Sampling-based Methods for Uncertainty and Sensitivity Analysis.* Sandia National Laboratories, SAND99-2240, Albuquerque, New Mexico.
9. Helton, J.C. and F.J. Davis. 2002. *Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems*. SAND2001-0417, Sandia National Laboratories, Albuquerque, New Mexico.
10. Ferson, S. 1994. Naive Monte Carlo methods yield dangerous underestimates of tail probabilities. Proceedings of the High Consequence Operations Safety Symposium, Sandia National Laboratories, SAND94–2364, J.A. Cooper (ed.), pp. 507–514.
11. Ferson, S. 2002. *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers.* Lewis Publishers, Boca Raton, Florida.

12. Ferson, S., V. Kreinovich, L. Ginzburg, K. Sentz and D.S. Myers. 2003. *Constructing Probability Boxes and Dempster-Shafer Structures*. Sandia National Laboratories, SAND2002-4015, Albuquerque, New Mexico.
13. Insua, D.R. and F. Ruggeri (eds.) 2000. *Robust Bayesian Analysis*. Lecture Notes in Statistics, Volume 152. Springer-Verlag, New York.
14. Walley, P. 1991. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.
15. Berleant, D. 1993. Automatically verified reasoning with both intervals and probability density functions. *Interval Computations* (1993 no. 2): 48-70.
16. Berleant, D. 1996. Automatically verified arithmetic on probability distributions and intervals. Pp. 227-244 in *Applications of Interval Computations*, edited by B. Kearfott and V. Kreinovich, Kluwer Academic Publishers.
17. Adams. E. and U. Kulisch 1993. *Scientific Computing with Automatic Result Verification.* Academic Press, Boston.
18. Yager, R.R. 1986. Arithmetic and other operations on Dempster-Shafer structures. *International Journal of Man-machine Studies* 25: 357-366.
19. Williamson, R.C. and T. Downs. 1990. Probabilistic arithmetic I: Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* 4: 89-158.
20. Ferson, S. and J. Hajagos. 2004. Arithmetic with uncertain numbers: rigorous and (often) best-possible answers. *Reliability and Engineering and System Safety* [to appear].
21. Iman, R.L. and W.J. Conover. 1980. Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. *Communications in Statistics A9*:1749-1842.
22. Iman, R.L. and W.J. Conover. 1982. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics B11*:311-334.
23. Iman, R.L., J.C. Helton, J.E. Campbell. 1981. An approach to sensitivity analysis of computer models, Part 1. Introduction, input variable selection and preliminary variable assessment. *Journal of Quality Technology* 13:174-183.
24. Iman, R.L., J.C. Helton, J.E. Campbell. 1981. An approach to sensitivity analysis of computer models, Part 2. Ranking of input variables, response surface validation, distribution effect and technique synopsis. *Journal of Quality Technology* 13:232-240.
25. Saltelli, A., K. Chan, E.M. Scott (eds.) 2001. *Sensitivity Analysis*. John Wiley, New York.
26. Ferson, S., W.T. Tucker, D. Berleant, J. Zhang and L. Ginzburg. 2004 [tentative]. *Dependence in Dempster-Shafer Theory and Probability Bounds Analysis*. Sandia National Laboratories, SAND2004-xxxx, Albuquerque, New Mexico.
27. Hettmansperger, T. and M. Keenan. 1980. Tailweight, statistical inference, and families of distributions - a brief survey. *Statistical Distributions in Scientific Work*, Volume 1, G.P. Patil et al. (eds.), Kluwer, Boston.

Sampling-Based Methods for Uncertainty and Sensitivity Analysis

J.C. Helton

Department of Mathematics and Statistics,
Arizona State University, Tempe AZ 85287-1804
E-mail: jchelto@sandia.gov

Abstract: Sampling-based methods for uncertainty and sensitivity analysis are reviewed. The following topics are considered: (i) Definition of probability distributions to characterize epistemic uncertainty in analysis inputs, (ii) Generation of samples from uncertain analysis inputs, (iii) Propagation of sampled inputs through an analysis, (iv) Presentation of uncertainty analysis results, and (v) Determination of sensitivity analysis results.

Keywords: Epistemic uncertainty, Latin hypercube sampling, Monte Carlo, Sensitivity analysis, Uncertainty analysis

1. INTRODUCTION

Sampling-based (i.e., Monte Carlo) approaches to uncertainty and sensitivity analysis are both effective and widely used [1-4]. Analyses of this type involve the generation and exploration of a mapping from uncertain analysis inputs to uncertain analysis results. The underlying idea is that analysis results $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_{nY}(\mathbf{x})]$ are functions of uncertain analysis inputs $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$. In turn, uncertainty in $\mathbf{x}$ results in a corresponding uncertainty in $\mathbf{y}(\mathbf{x})$. This leads to two questions: (i) What is the uncertainty in $\mathbf{y}(\mathbf{x})$ given the uncertainty in $\mathbf{x}$?, and (ii) How important are the individual elements of $\mathbf{x}$ with respect to the uncertainty in $\mathbf{y}(\mathbf{x})$? The goal of uncertainty analysis is to answer the first question, and the goal of sensitivity analysis is to answer the second questions. In practice, the implementation of an uncertainty analysis and the implementation of a sensitivity analysis are very closely connected on both a conceptual and a computational level.

The following sections summarize the five basic components that underlie the implementation of a sampling-based uncertainty and sensitivity analysis: (i) Definition of distributions $D_1, D_2, \ldots, D_{nX}$ that characterize the uncertainty in the components $x_1, x_2, \ldots, x_{nX}$ of $\mathbf{x}$ (Sect. 2), (ii) Generation of a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{nS}$ fro the $\mathbf{x}$'s in consistency with the distributions $D_1, D_2, \ldots, D_{nX}$ (Sect. 3), (iii) Propagation of the sample through the analysis to produce a mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$, from analysis inputs to analysis results (Sect. 4), (iv) Presentation of uncertainty analysis results (i.e., approximations to the distributions of the elements of $\mathbf{y}$ constructed from the corresponding elements of $\mathbf{y}(\mathbf{x}_i)$, $i = 1, 2, \ldots, nS$) (Sect. 5), and (v) Determination of sensitivity analysis results (i.e., exploration of the mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$) (Sect. 6). Space limitations in this presentation preclude the presentation of detailed examples of the indicated analysis components; however, examples can be found in the published descriptions of an uncertainty and sensitivity analysis carried out for the Waste

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

Isolation Pilot Plant (e.g., [5-7]). The presentation then ends with a concluding summary (Sect. 7).

Only probabilistic characterizations of uncertainty are considered in this presentation. Alternative uncertainty representations (e.g., evidence theory, possibility theory, fuzzy set theory, interval analysis) are an active area of research [8, 9] but are outside the intended scope of this presentation.

2. CHARACTERIZATION UNCERTAINTY

Definition of the distributions D_1, D_2, …, D_{nX} that characterize the uncertainty in the components x_1, x_2, …, x_{nX} of $\mathbf{x}$ is the most important part of a sampling-based uncertainty and sensitivity analysis as these distributions determine both the uncertainty in y and the sensitivity of $\mathbf{y}$ to the elements of $\mathbf{x}$. The distributions D_1, D_2, …, D_{nX} are typically defined through an expert review process [10-13], and their development can constitute a major analysis cost. A possible analysis strategy is to perform an initial exploratory analysis with rather crude definitions for D_1, D_2, …, D_{nX} and use sensitivity analysis to identify the most important analysis inputs; then, resources can be concentrated on characterizing the uncertainty in these inputs and a second presentation or decision-aiding analysis can be carried out with these improved uncertainty characterizations.

The scope of an expert review process can vary widely depending on the purpose of the analysis, the size of the analysis, and the resources available to carry out the analysis. At one extreme is a relatively small study in which a single analyst both develops the uncertainty characterizations (e.g., on the basis of personal knowledge or a cursory literature review). At the other extreme, is a large analysis on which important societal decisions will be based and for which uncertainty characterizations are carried out for a large number of variables by teams of outside experts who support the analysts actually performing the analysis.

Given the breadth of analysis possibilities, it is beyond the scope of this presentation to provide an exhaustive review of how the distributions D_1, D_2, …, D_{nX} might be developed. However, as general guidance, it is best to avoid trying to define these distributions by specifying the defining parameters (e.g., mean and standard deviation) for a particular distribution. Rather, distributions can be defined by specifying selected quantiles (e.g., 0.0, 0.1, 0.25, …, 0.9, 1.0), which should keep the individual supplying the information in closer contact with the original sources of information or insight than is the case when a particular named distribution is specified. Distributions from multiple experts can be aggregated by averaging.

3. GENERATION OF SAMPLE

Several sampling strategies are available, including random sampling, importance sampling, and Latin hypercube sampling [14, 15]. Latin hypercube sampling is very popular for use with computationally demanding models because its efficient stratification properties allow for the extraction of a large amount of uncertainty and sensitivity information with a relatively small sample size.

Latin hypercube sampling operates in the following manner to generate a sample of size nS from the distributions $D_1, D_2, \ldots, D_{nX}$ associated with the elements of $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$. The range of each x_j is exhaustively divided into nS disjoint intervals of equal probability and one value x_{ij} is randomly selected from each interval. The nS values for x_1 are randomly paired without replacement with the nS value for x_2 to produce nS pairs. These pairs are then randomly combined without replacement with the nS values for x_3 to produce nS triples. This process is continued until a set of nS nX-triples $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{i,nX}]$, $i = 1, 2, \ldots, nS$, is obtained, with this set constituting the Latin hypercube sample. In addition, effective correlation control procedures are available for use with Latin hypercube sampling [16, 17]. The popularity of Latin hypercube sampling recently led to the original article being designated a *Technometrics* classic in experimental design [18].

Latin hypercube sampling is a good choice for a sampling procedure when computationally demanding models are being studied. When the model is not computationally demanding, many model evaluations can be performed and random sampling works as well as Latin hypercube sampling.

4. PROPAGATION OF SAMPLE THROUGH THE ANALYSIS

Propagation of the sample through the analysis to produce the mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i))]$, $i = 1, 2, \ldots, nS$, from analysis inputs to analysis results is often the most computationally demanding part of a sampling-based uncertainty and sensitivity analysis. The details of this propagation are analysis specific and can range from very simple for analyses that involve a single model to very complicated for large analyses that involve complex systems of linked models [7, 19].

When a single model is under consideration, this part of the analysis can involve little more than putting a DO loop around the model that (i) supplies the sampled input to the model, (ii) runs the model, and (iii) stores model results for later analysis. When more complex analyses with multiple models are involved, considerable sophistication may be required in this part of the analysis. Implementation of such analyses can involve (i) development of simplified models to approximate more complex models, (ii) clustering of results at model interfaces, (ii) reuse of model results through interpolation or linearity properties, and (iv) complex procedures for the storage and retrieval of analysis results.

5. PRESENTATION OF UNCERTAINTY ANALYSIS RESULTS

Presentation of uncertainty analysis results is generally straight forward and involves little more than displaying the results associated with the already calculated mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$. Presentation possibilities include means and standard deviations, density functions, cumulative distribution function (CDFs), complementary cumulative distribution functions (CCDFs), and box plots [2, 15]. Presentation formats such as CDFs, CCDFs and box plots are usually preferable to means and standard deviations because of the large amount of uncertainty information that is lost in the calculation of means and standard deviations.

6. DETERMINATION OF SENSITIVITY ANALYSIS RESULTS

Determination of sensitivity analysis results is usually more demanding that the presentation of uncertainty analysis results due to the need to actually explore the mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$, to assess the effects of individual components of $\mathbf{x}$ on the components of $\mathbf{y}$. A number of approaches to sensitivity analysis that can be used in conjunction with a sampling-based uncertainty analysis are listed and briefly summarized below. In this summary, (i) x_j is an element of $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$, (ii) y_k is an element of $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_{nY}(\mathbf{x})]$, (iii) $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{i,nX}]$, $i = 1, 2, \ldots, nS$, is a random or Latin hypercube sample from the possible values for $\mathbf{x}$ generated in consistency with the joint distribution assigned to the x_j, (iv) $\mathbf{y}_i = \mathbf{y}(\mathbf{x}_i)$ for $i = 1, 2, \ldots, nS$, and (v) x_{ij} and y_{ik} are arbitrary elements of $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively.

Scatterplots. Plots of points $[x_{ij}, y_{ik}]$ for $i = 1, 2, \ldots, nS$ can reveal nonlinear or other unexpected relationships. Natural starting point in complex analysis that can help development of sensitivity analysis strategy using one or more additional techniques. Additional information: Sect. 6.6.1, [20].

Cobweb Plots. Plots of points $[\mathbf{x}_i, y_{ik}] = [x_{i1}, x_{i2}, \ldots, x_{i,nX}, y_{ik}]$ for $i = 1, 2, \ldots, nS$. Provides two-dimensional representation for a $nX + 1$ dimensional quantity. Generalization of a scatterplot. Provides more information in a single plot frame than a scatterplot but is harder to read. Additional information: Sect. 11.7, [21]

Correlation. Provides measure of the strength of the linear relationship between x_j and y_k. Equal to standardized regression coefficient in linear regression relating y_k to x_j; also equal in absolute value to the square root of the R^2 value associated with the indicated regression. Often referred to as Pearson correlation coefficient. Additional information: Sect. 6.6.4, [20].

Regression Analysis. Provides algebraic representation of relationships between y_k and one or more x_j's. Usually performed in stepwise fashion with initial inclusion of most important x_j, then two most important x_j's, and so on until no more x_j's that significantly affect y_k can be identified. Variable importance indicated by order of selection in stepwise process, changes in R^2 values as additional variables are added to the regression model, and standardized regression coefficients for the x_j's in the final regression model. Additional information: Sects. 6.6.2, 6.6.3, 6.6.5, [20].

Partial Correlation. Provides measure of the strength of the linear relationship between y_k and x_j after the linear effects of all other elements of $\mathbf{x}$ have been removed. Additional information: Sect. 6.6.4, [20].

Rank Transformations. Replaces values for y_k and x_j with their corresponding ranks. Smallest valued assigned a rank of 1; next largest value assigned a rank of 2; tied values are assigned their average rank; and so on up to the largest value, which is assigned a rank of nS. Converts a nonlinear but monotonic relationship between y_k and x_j to a linear relationship. Produces rank (i.e., Spearman) correlations, rank regressions, standardized rank regression coefficients and partial rank correlation coefficients. Additional information: Sect. 6.6.6, [20]; [22].

Nonparametric Regression. Seeks more general models than those obtained by least squares regression. Attempts to find models that are local in the approximation to

the relationship between y_k and multiple x_j's. Better at capturing complex nonlinear relationships than traditional regression or rank regression. Can be applied in stepwise manner. Sequential changes in R^2 values with addition of successive variables to the model provides indication of variable importance. Very promising. Additional information: [23-25].

Tests for Patterns Based on Gridding. Grids can be placed on the scatterplot for y_k and x_j and then various statistical tests can be performed to determine if the distribution of points across the grid cells appears to be nonrandom. Appearance of a nonrandom pattern indicates that x_j has an effect on y_k. Possibilities include (i) tests for common means, common medians, and common distributions for values of y_k based on partitioning the range of x_j and (ii) tests for no influence based on partitioning the ranges of x_j and y_k. Additional information: Sects. 6.6.8, 6.6.9, [20]; [26].

Tests for Patterns Based on Distance Measures. Considers relationships within the scatterplot for y_k and x_j such as the distribution of distances between nearest neighbors. Provides way to identify nonrandom relationships between y_k and x_j. Avoids problem of defining appropriate gridding associated grid-based methods. Additional information: [27-30].

Trees. Searches for relationships between y_k and multiple x_j's by successively subdividing the sample elements $\mathbf{x}_i$ on the basis of observed effects of individual x_j's on y_k. Additional information: [31].

Two-Dimensional Kolmogorov-Smirnov Test. Provides way to test for nonrandom patterns in the scatterplot for y_k and x_j that does not require the imposition of a grid. Additional information: [32-34].

Squared Differences of Ranks. Seeks to identify presence of nonlinear relationship between y_k and x_j. Based on squared differences of consecutive ranks of y_k when the values of y_k have been ordered by the corresponding values of x_j. Additional information: [35].

Top-Down Concordance with Replicated Samples. Uses top-down coefficient of concordance and replicated (i.e., independently generated) samples. Sensitivity analysis with some appropriate technique performed for each sample. Top-down coefficient used to identify important variables by seeking variables with similar rankings across all replicates. Additional information: [36].

Variance Decomposition. The variance decomposition proposed by Sobol' and others is formally defined by high-dimensional integrals involving the x_j and $y_k(\mathbf{x})$. Provides decomposition of variance $V(y_k)$ of y_k in terms of the contributions V_j of individual x_j's to $V(y_k)$ and also the contributions of various interactions between the x_j to $V(y_k)$. In practice, indicated decomposition is obtained with sampling based methods. Two samples from $\mathbf{x}$ of size nS are required to estimate all V_j; $nX + 2$ samples of size nS are required to estimate all V_j and also the contributions of each of the x_j's and its interactions with other elements of $\mathbf{x}$ to $V(y_k)$. Conceptually very appealing but can be computationally demanding as more samples and probably larger samples required than with other sampling-based approaches to sensitivity analysis. Additional information: [37, 38]

7. SUMMARY

Sampling-based uncertainty and sensitivity analysis is widely used, and as a result, is a fairly mature area of study. However, there still remain a number of important challenges and areas for additional study. For example, there is a need for sensitivity analysis procedures that are more effective at revealing nonlinear relations than those currently in use. Among the approaches to sensitivity analysis listed in Sect. 6, nonparametric regression [23-25], the two-dimensional Kolmogorov-Smirnov test [32-34], tests for nonmonotone relations [35], tests for nonrandom patterns [26-30], and complete variance decomposition [37, 38] have not been as widely used as some of the other approaches and merit additional investigation and use. As another example, sampling-based procedures for uncertainty and sensitivity analysis usually use probability as the model, or representation, for uncertainty. However, when limited information is available with which to characterize uncertainty, probabilistic characterizations can give the appearance of more knowledge than is really present. Alternative representations for uncertainty such as evidence theory and possibility theory merit consideration for their potential to represent uncertainty in situations where little information is available [8, 9]. Finally, a significant challenge is the education of potential users of uncertainty and sensitivity analysis about (i) the importance of such analyses and their role in both large and small analyses, (ii) the need for appropriate separation of aleatory and epistemic uncertainty in the conceptual and computational implementation of analyses of complex systems [39-43], (ii) the need for a clear conceptual view of what an analysis is intended to represent and a computational design that is consistent with that view [44], (iv) the role that uncertainty and sensitivity analysis plays in model and analysis verification, and (v) the importance of avoiding deliberately conservative assumptions if meaningful uncertainty and sensitivity analysis results are to be obtained.

ACKNOWLEDGMENTS

Work performed for Sandia National Laboratories, which is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Security Administration under contract DE-AC04-94AL-85000. Editorial support provided by F. Puffer and J. Ripple of Tech Reps, a division of Ktech Corporation.

REFERENCES

[1] R. L. Iman, "Uncertainty and Sensitivity Analysis for Computer Modeling Applications," presented at *Reliability Technology - 1992, The Winter Annual Meeting of the American Society of Mechanical Engineers, Anaheim, California, November 8-13, 1992*, Vol. 28, pp. 153-168. New York, NY, 1992.

[2] J. C. Helton, "Uncertainty and Sensitivity Analysis Techniques for Use in Performance Assessment for Radioactive Waste Disposal," *Reliability Engineering and System Safety*, vol. 42, pp. 327-367, 1993.

[3] D. M. Hamby, "A Review of Techniques for Parameter Sensitivity Analysis of Environmental Models," *Environmental Monitoring and Assessment*, vol. 32, pp. 135-154, 1994.

[4] S. M. Blower and H. Dowlatabadi, "Sensitivity and Uncertainty Analysis of Complex Models of Disease Transmission: an HIV Model, as an Example," *International Statistical Review*, vol. 62, pp. 229-243, 1994.

[5] J. C. Helton, D. R. Anderson, H.-N. Jow, M. G. Marietta, and G. Basabilvazo, "Performance Assessment in Support of the 1996 Compliance Certification Application for the Waste Isolation Pilot Plant," *Risk Analysis*, vol. 19, pp. 959 - 986, 1999.

[6] J. C. Helton, "Uncertainty and Sensitivity Analysis in Performance Assessment for the Waste Isolation Pilot Plant," *Computer Physics Communications*, vol. 117, pp. 156-180, 1999.

[7] J. C. Helton and M. G. Marietta, "Special Issue: The 1996 Performance Assessment for the Waste Isolation Pilot Plant," *Reliability Engineering and System Safety*, vol. 69, pp. 1-451, 2000.

[8] G. J. Klir and M. J. Wierman, "Uncertainty-Based Information," Physica-Verlag, New York, NY 1999.

[9] J. C. Helton, J. D. Johnson, and W. L. Oberkampf, "An Exploration of Alternative Approaches to the Representation of Uncertainty in Model Predictions," *Reliability Engineering and System Safety (to appear)*.

[10] S. C. Hora and R. L. Iman, "Expert Opinion in Risk Analysis: The NUREG-1150 Methodology," *Nuclear Science and Engineering*, vol. 102, pp. 323-331, 1989.

[11] M. C. Thorne and M. M. R. Williams, "A Review of Expert Judgement Techniques with Reference to Nuclear Safety," *Progress in Nuclear Safety*, vol. 27, pp. 83-254, 1992.

[12] R. J. Budnitz, G. Apostolakis, D. M. Boore, L. S. Cluff, K. J. Coppersmith, C. A. Cornell, and P. A. Morris, "Use of Technical Expert Panels: Applications to Probabilistic Seismic Hazard Analysis," *Risk Analysis*, vol. 18, pp. 463-469, 1998.

[13] M. McKay and M. Meyer, "Critique of and Limitations on the use of Expert Judgements in Accident Consequence Uncertainty Analysis," *Radiation Protection Dosimetry*, vol. 90, pp. 325-330, 2000.

[14] M. D. McKay, R. J. Beckman, and W. J. Conover, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, vol. 21, pp. 239-245, 1979.

[15] J. C. Helton and F. J. Davis, "Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems," *Reliability Engineering and System Safety*, vol. 81, pp. 23-69, 2003.

[16] R. L. Iman and W. J. Conover, "A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables," *Communications in Statistics: Simulation and Computation*, vol. B11, pp. 311-334, 1982.

[17] R. L. Iman and J. M. Davenport, "Rank Correlation Plots for Use with Correlated Input Variables," *Communications in Statistics: Simulation and Computation*, vol. B11, pp. 335-360, 1982.

[18] M. D. Morris, "Three Technometrics Experimental Design Classics," *Technometrics*, vol. 42, pp. 26-27, 2000.

[19] R. J. Breeding, J. C. Helton, E. D. Gorham, and F. T. Harper, "Summary Description of the Methods Used in the Probabilistic Risk Assessments for NUREG-1150," *Nuclear Engineering and Design*, vol. 135, pp. 1-27, 1992.

[20] J. C. Helton and F. J. Davis, "Sampling-Based Methods," in *Sensitivity Analysis*, A. Saltelli, K. Chan, and E.M. Scott, Eds. New York, NY: Wiley. pp. 101-153, 2000.
[21] R. M. Cooke and J.M. van Noortwijk, "Graphical Methods," in *Sensitivity Analysis*, A. Saltelli, K. Chan, and E.M. Scott, Eds. Wiley: New York, NY, 2000, pp. 245-264.
[22] R. L. Iman and W. J. Conover, "The Use of the Rank Transform in Regression," *Technometrics*, vol. 21, pp. 499-509, 1979.
[23] T. J. Hastie and R. J. Tibshirano, *Generalized Additive Models*. London: Chapman & Hall, 1990.
[24] J. S. Simonoff, *Smoothing Methods in Statistics*. New York: Springer-Verlag, 1996.
[25] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*. Oxford: Clarendon, 1997.
[26] J. P. C. Kleijnen and J. C. Helton, "Statistical Analyses of Scatterplots to Identify Important Factors in Large-Scale Simulations, 1: Review and Comparison of Techniques," *Reliability Engineering and System Safety*, vol. 65, pp. 147-185, 1999.
[27] B. D. Ripley, "Tests of "Randomness" for Spatial Point Patterns," *Journal of the Royal Statistical Society*, vol. 41, pp. 368-374, 1979.
[28] P. J. Diggle and T. F. Cox, "Some Distance-Based Tests of Independence for Sparsely-Sampled Multivariate Spatial Point Patterns," *International Statistical Review*, vol. 51, pp. 11-23, 1983.
[29] G. Zeng and R. C. Dubes, "A Comparison of Tests for Randomness," *Pattern Recognition*, vol. 18, pp. 191-198, 1985.
[30] R. Assunçao, "Testing Spatial Randomness by Means of Angles," *Biometrics*, vol. 50, pp. 531-537, 1994.
[31] S. Mishra, N. E. Deeds, and B.S. RamoRao, "Application of Classification Trees in the Sensitivity Analysis of Probabilistic Model Results," *Reliability Engineering and System Safety*, vol. 79, pp. 123-129, 2003.
[32] J. A. Peacock, "Two-Dimensional Goodness-Of-Fit Testing in Astronomy," *Monthly Notices of the Royal Astronomical Society*, vol. 202, pp. 615-627, 1983.
[33] G. Fasano and A. Franceschini, "A Multidimensional Version of the Kolmogorov-Smirnov Test," *Monthly Notices of the Royal Astronomical Society*, vol. 225, pp. 155-170, 1987.
[34] J. E. Garvey, E. A. Marschall, and R. Wright, A., "From Star Charts to Stoneflies: Detecting Relationships in Continuous Bivariate Data," *Ecology*, vol. 79, pp. 442-447, 1998.
[35] S. C. Hora and J. C. Helton, "A Distribution-Free Test for the Relationship Between Model Input and Output when Using Latin Hypercube Sampling," *Reliability Engineering and System Safety*, vol. 79, pp. 333-339, 2003.
[36] J. C. Helton, F. J. Davis, and J. D. Johnson, "A Comparison of Uncertainty and Sensitivity Analysis Results Obtained with Random and Latin hypercube sampling," *Reliability Engineering and System Safety (submitted)*.

[37] A. Saltelli, S. Tarantola, and K. P.-S. Chan, "A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output," *Technometrics*, vol. 41, pp. 39-56, 1999.
[38] G. Li, C. Rosenthal, and H. Rabitz, "High-Dimensional Model Representations," *The Journal of Physical Chemistry*, vol. 105, pp. 7765-7777, 2001.
[39] G. Apostolakis, "The Concept of Probability in Safety Assessments of Technological Systems," *Science*, vol. 250, pp. 1359-1364, 1990.
[40] J. C. Helton, "Treatment of Uncertainty in Performance Assessments for Complex Systems," *Risk Analysis*, vol. 14, pp. 483-511, 1994.
[41] F. O. Hoffman and J. S. Hammonds, "Propagation of Uncertainty in Risk Assessments: The Need to Distinguish Between Uncertainty Due to Lack of Knowledge and Uncertainty Due to Variability," *Risk Analysis*, vol. 14, pp. 707-712, 1994.
[42] M. E. Paté-Cornell, "Uncertainties in Risk Analysis: Six Levels of Treatment," *Reliability Engineering and System Safety*, vol. 54, pp. 95-111, 1996.
[43] J. C. Helton, "Uncertainty and Sensitivity Analysis in the Presence of Stochastic and Subjective Uncertainty," *Journal of Statistical Computation and Simulation*, vol. 57, pp. 3-76, 1997.
[44] J. C. Helton, "Mathematical and Numerical Approaches in Performance Assessment for Radioactive Waste Disposal: Dealing with Uncertainty," in *Modelling Radioactivity in the Environment*, E. M. Scott, Ed. New York: Elsevier Science, 2003, pp. 353-390.

Estimation Procedures and Error Analysis for Inferring the Total Plutonium Produced by a Graphite-Moderated Reactor

Patrick G. Heasler[a], Tom Burr [c], Bruce Reid[a], Christopher Gesh[a], Charles Bayne[b]

[a]Pacific Northwest National Laboratory, Richland, WA, US;
[b]Oak Ridge National Laboratory, Oak Ridge, TN, US;
[c]Los Alamos National Laboratory, Los Alamos, NM, US
E-mail: patrick.heasler@pnl.gov, bayneck@ornl.gov, tburr@lanl.gov

ABSTRACT

GIRM (Graphite Isotope Ratio Method) is a technique that can determine total Plutonium (Pu) production in a graphite moderated reactor. In the GIRM methodology, total Pu production is estimated by measuring isotopic ratios of trace elements in extracted graphite samples from the target reactor.

Many safeguards problems require an estimate of total Pu production. For example, a declaration of total Pu might need to be verified through GIRM. In some cases, reactor information (such as core dimensions, coolant details, and operating history) are so well documented that reactor computer-models can predict total Pu production. In most cases however, reactor information is imperfectly known or of questionable validity, so a measurement-based method such as GIRM is essential to such a verification strategy.

In this article we concentrate on GIRM's estimation procedure and its associated uncertainty. We describe a simulation strategy to estimate its uncertainty, including the impact of local and global computer codes, and illustrate GIRM for a specific reactor.

Keywords: reactor code errors, simulation, uncertainty in estimated Plutonium

1. INTRODUCTION

In GIRM, samples from the graphite moderator are taken along the fuel channels in the reactor of interest, and isotopic ratios in the samples are measured with mass spectrography (TIMS for Uranium or Pu isotope ratios and SIMS for Boron isotope ratios). These isotopic ratios are converted to local plutonium production rates (i.e. grams of Pu produced per cm of fuel rod) using a lattice physics code (WIMS). Finally, a 3D regression model is used to estimate a Pu fluence field for the reactor which is then integrated over the fuel channels to estimate the total Pu produced. The 3D regression model used in a specific application is determined from a reactor physics code (such as KENO).

The basic scheme has been subjected to several feasibility studies and experimental tests (see [1], [2], and [3]). Two previous error analyses of this methodology have been conducted. The earliest analysis, Reference [1] evaluated uncertainties associated with a "generic" reactor. The other, Reference [3] evaluated actual measurements taken from a British reactor, Trawsfynydd. Because the GIRM methodology has changed substantially in recent years [4] and additional sources of error have been included (i.e. WIMS reactor

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

code error and 3D fluence model errors), it is timely to re-evaluate the total uncertainty with the GIRM-based estimate of total Pu production. The procedures described in the next section are illustrated by applying them to a "generic" reactor, which is described in [4].

2. METHODOLOGY

GIRM can be organized into three steps: (1) a planning step that assembles and evaluates information about the target reactor; (2) a sampling step that extracts graphite samples from the target reactor, and (3) an analysis/estimation step that measures the isotopic ratios and converts them to a Pu estimate.

In the planning step, information concerning a target reactor is gathered. This includes reactor core dimensions, fuel channel locations, control rod locations, coolant details, fuel rod specifications, and operating history. WIMS produces curves that relate the isotopic ratios to local Pu fluence. WIMS runs also generate a parametric evaluation of the main reactor variables such as temperature, density, and dimensions. A sampling plan is designed that describes where graphite samples will be taken from the target reactor.

The analysis step includes mass spectrographic measurements of the isotopic ratios of each sample, and statistical analysis to combine all the measurements to estimate total Pu production. In the first step in this statistical analysis, the measured isotopic ratios are converted to local Pu fluence (g/cm) estimates using curves calculated by WIMS.

The second step in statistical analysis fits a 3D linear regression model to the local Pu fluence estimates produced by the first step. The 3D regression model (called the global regression model), is produced from 3D reactor-physics solutions. After the global regression model has been fit to the local fluence estimates, it is easy to determine the total Pu production by integration. Standard propagation-of-error (POE) formulas also produce the uncertainty associated with this estimate in the form of a standard error. Uncertainties of the entire estimation procedure are verified using Monte-Carlo simulation. Because of deficiencies in the POE methods, we presently rely on simulation to produce the most comprehensive estimate of total uncertainty.

2.1. Pu-fluence from U/Pu Data and B data

At a given sample location, the TIMS mass spectrograph produces a vector of 5 isotopic ratios ($^{236}U/^{238}U$, $^{236}U/^{235}U$, $^{240}Pu/^{239}Pu$, $^{241}Pu/^{239}Pu$, $^{242}Pu/^{239}Pu$), while SIMS produces a single isotopic ratio for $^{10}B/^{11}B$. All measured isotopic ratios have an associated uncertainty, supplied by the chemical analysis. It should be noted that Boron is particularly suited to a low burn-up reactor. For a high burn-up reactor, the Boron measurements would be replaced by Titanium.

These isotopic ratios are used to estimate the local Pu-fluence, ϕ_{pu} at each sample location, and produce $se(\hat{\phi}_{pu})$. To accomplish this, a nonlinear regression model is used to relate Pu-fluence ϕ_i to the measured isotopic ratios, R_{ij}. The model is;

$$R_{ij} = H_j(\phi_i) + e_{ij} \tag{1}$$

where the index i identifies the sample location, and the index j the specific isotopic ratio measured. The function $H_j(\phi)$ represents the isotopic ratio curves produced by WIMS runs for the target reactor and the error term, e_{ij}, represents error. The least-squares estimator for the vector of ϕ_i is the value that minimizes the quadratic form*;

$$SSE_i = [R_i - H(\phi_i)]^T Cov(e_i)^{-1}[R_i - H(\phi_i)] \tag{2}$$

while its standard error is given by an asymptotic approximation used for non-linear regression (i.e., the inverse of Fisher information matrix).

The covariance of the regression uncertainties includes measurement error, and reactor physics error (RPE). In other words, the covariance matrix used by the regression is $Cov(e) = Cov(\text{Meas. Error}) + Cov(\text{Reactor Phy. Error})$. Including RPE in the regression weights improves the total Pu estimates by 50% because it is the dominant source of uncertainty in the data. RPE describes the effect that uncertainties in important reactor parameters (such as graphite temperature) have on this regression problem.

We note here that the se's produced by this algorithm do not account for calibration bias. Although the se of the result does include RPE, correlation between local estimates is not produced, which also can have an important effect on total Pu uncertainty and is the chief reason that Monte-Carlo generated uncertainties are better than POE uncertainties.

2.2. Global Regression and Estimate of Total Pu

Linear regression is used to fit a 3D fluence field to the local fluence estimates produced by the previous local estimation step. Let $\Phi(x;\beta)$ represent the fluence field model, with x representing a location in the reactor core, and β a set of unknown parameters that determines the shape of the fluence field. Linear regression is used to determine parameter estimates that produce a fluence field that is as close as possible to the local fluence values.

We call this "global" regression because it transforms the local Pu estimates into a fluence field that is defined at any point within the entire reactor core. The global regression model is described by; $\hat{\phi}_i = \Phi(x_i;\beta) + e_i$ with weight W_i associated with location x_i. The weight is determined by the se assigned to $\hat{\phi}_{pu}(x)$ from the local Pu estimation step, $W_i = se(\hat{\phi}_{pu}(x))^{-2}$. Weighted regression is then performed to produce estimates, $\hat{\beta}$, of the unknown parameter vector β, and its covariance matrix, $Cov(\hat{\beta})$. The form used for the 3D fluence field model is

$$\Phi(x;\beta) = \sum_{k=1}^{p} \beta_k \Psi_k(x). \tag{3}$$

The $\Psi_k(x)$ must be known functions, that are relatively good at approximating the target reactor fluence field. An estimate of the total Plutonium production, $\hat{T}_{pu}$, is then determined by integrating the estimated fluence field over all fuel channels in the reactor. The integration result is a linear combination of the unknown beta parameters; consequently, it is simple to calculate the standard error of the estimate, $\hat{T}_{pu}$, from $Cov(\hat{\beta})$.

*$R_i = (R_{11}, R_{12}, ...R_{1n})$ and similarly for $H(\phi)$, and e_i.

Let $(x_{1\ell}, x_{2\ell})$ represent the horizontal coordinates of the ℓ'th fuel channel, so that $\hat{T}_{pu}$ is given by;

$$\hat{T}_{pu} = \sum_{\ell} \int_{Z_{bottom}}^{Z_{top}} \Phi(x_{1,\ell}, x_{2\ell}, z_3; \hat{\beta}) dz_3 = \sum_{k=1}^{p} \hat{\beta}_k \sum_{\ell} \int_{Z_{bottom}}^{Z_{top}} \Psi_k(x_{1,\ell}, x_{2\ell}, z_3) dz_3 = \sum_{k=1}^{p} \hat{\beta}_k \omega_k \tag{4}$$

with the integration constants ω_k defined by $\omega_k = \sum_{\ell} \int_{Z_{bottom}}^{Z_{top}} \Psi_k(x_{1,\ell}, x_{2,\ell}, z_3) dz_3$.

An estimate for the total Pu can therefore be found by multiplying the beta estimates by the integration constants ω_k. In vector notation this reduces to; $\hat{T}_{pu} = \omega^T \hat{\beta}$ and the estimate of total Pu has $se(\hat{T}_{pu})^2 = \omega^T Cov(\hat{\beta})\omega$. These formulas produce an estimate of T_{pu}, and its se, which is the final objective of this statistical procedure. Incidentally, if the regression model $\Phi(x, \beta)$ can fit the true fluence field without any error, then the estimate is unbiased and the true total fluence, T_{pu} can be expressed as a linear combination of the ω_k and true β_k that resembles Eq. (4).

The adequacy of the global regression model depends upon the proper selection of the set of "basis functions", $\Psi_k(x)$. 3D reactor physics models are used to produce a "basis set" of reactor fluence profiles, based on reactor operating history. If one is fairly certain about the operating history, this set will be small, if one is less certain, the set will contain more profiles and uncertainty related to global regression will be correspondingly greater.

In the generic example presented in this article, the KENO code has been used to produce a "best estimate" for fluence, which is included in the basis set, along with 44 "eigen-function" profiles from an homogeneous-core solution. The homogeneous-core "eigen-function" profiles are the functions we would use when little reliable information is available for the target reactor. Thus the example global regression model is constructed from generic information and one "best estimate" profile.

2.3. Monte Carlo Simulation for Error Analysis

The steps for our simulation are: (1) Calculate a "true state" for the target reactor (the global Pu-fluence, $\Phi(x)$ at locations x_i) (see the next paragraph); (2) Add error to locations x_i that arises due to limitations of the sampler; (3) Sample the reactor parameters (such as coolant and fuel temperature, etc.) from distributions centered on true values with standard deviations estimated using linear approximations to results of auxiliary WIMS runs. The "true" isotopic/fluence curves are generated WIMS using these sampled reactor parameters; (4) Add error to the true isotopic ratios to create measured values at samples (with sample contamination, calibration, and random errors), (5) Run the simulated measurements through the Pu estimation algorithm, and finally (6) compare the estimated result to the "true" Pu value, and repeat many times.

The "true" global fluence is simulated by random selection from a set of "representative" 3D fluence fields. This set is related to, but not the same as the "basis" set used to construct the global regression model. The set of 3D fluence fields contains plausible (best-estimate) and extreme fluence shapes, calculated from what is known about the target reactor's operating history. For the generic reactor used here, this set consists of seven fields calculated by KENO. Of these seven fields, three represent best-estimates

versions of reactor fluence, while the other four represent extreme fields. The extreme fields are calculated with unusual control rod configurations which are possible, but it is unlikely that the reactor would be operated in this configuration for any length of time.

The local RPE is the result of errors in certain reactor parameters required to complete a WIMs calculation. The WIMs code result can be mathematically described by the function $R = H(\phi; \alpha)$, where R is a vector whose components represent the ratios of all of the measured isotopes in graphite, while ϕ represents the corresponding Pu-fluence. $H(\cdot)$ is also a function of reactor parameters, as represented by the vector α. There is some error associated with the best-estimate value for α used in the estimation procedure, and to account for this, α is considered a random vector, with a mean centered around the "best estimate" and standard deviation representative of the uncertainty associated with these parameters. In the Monte-Carlo, a set of "true" fluence curves is simulated by sampling α from this distribution and computing $H(\phi, \alpha_{sampled})$.

For the example reactor, the reactor parameter vector α represents the following; fuel pin radius, fuel temperature, graphite density, graphite temperature, graphite equivalent boron concentration, and WIMS code uncertainty. "WIMS code uncertainty" represents numerical errors produced in WIMS, estimated by comparing WIMS results to results from another code. It is supposed to represent the difference between the answer produced by WIMS in GIRM, and a computer code that could solve the problem without error.

As a final step, the Monte Carlo simulates SIMS and TIMS measurement errors. These consist of random, calibration, and contamination errors, which are added to the true isotopic ratios to produce a measured value. These operations are summarized by the formula $R_{meas} = (1-C)R_{irr} + CR_{nat} + e_{cal} + e_{ran}$ where, for example, R_{irr} is the true (irradiated) ratio of Boron in the sample, which is contaminated by $C\%$ of natural Boron (that has ratio R_{nat}). A calibration error of e_{cal} and a random error of e_{ran} is then added to this result to obtain the measured ratio, R_{meas}.

It is important to note that the simulated measurements are more complicated than those assumed for the regression model (Equation 1) used in the estimation step. The contamination can cause bias, while the calibration errors can cause correlations, and neither of these effects are accounted for in the regression model.

2.4. Sampling as an Optimization Problem

A sampling plan consists of a set of specified locations in a reactor. These locations are represented by $x_1, x_2, ...x_n$, with each x_i a 3D vector representing the coordinates of a sampling location. The Monte Carlo error analysis methodology can be used to select the best sampling plan from a limited set of candidate plans. However, Monte Carlo evaluation is unsuited (too slow) to be used as part of a sample optimization scheme.

To find the optimal sample design, we have utilized the global regression model described in the estimation step. This global regression model can be quickly evaluated to produce an approximate RMSE for a particular to a sample design. These results are used by an optimization algorithm to find the sample design with a small root mean square error. The optimization is typically done with constraints; Samples can only be taken in certain allowed channels, and the sample locations within a channel are fixed.

Our optimization evaluations for the example reactor resulted in a design with 100 sample locations, taken in 10 channels. Roughly, the design consists of three "half-planes" in three radial directions. An experimental design tool (ECHIP, available at echip.com) was also employed to evaluate the designs and identify "high-bias" sample locations, which were eliminated. This optimized design was used in the example evaluations presented in this article.

3. EXAMPLE RESULTS FOR A LOW-BURNUP REACTOR

This example provides an upper-bound case for uncertainties when GIRM is applied to a low-burnup reactor. This means the uncertainties in RPE, contamination, instrument error, etc. are large, but not unreasonably large.

3.1. Estimation Uncertainty for the Example

To determine Pu estimation error for the example, 1400 sets of measurements were simulated, resulting in 1400 "total Pu estimates," which were compared to the "true Pu" (250 Kg for the generic case). Table 1 illustrates a few of these results.

Table 1. Simulations for Generic Case

	Estimation Results				Est Error
	(Kg)	(Kg)	(Kg)		(Kg)
Run	$\hat{T}_{Pu}$	True T_{Pu}	$se(\hat{T}_{Pu})$	GOF	$\hat{T}_{Pu} - T_{Pu}$
1	253.80	250.00	1.22	4.46	3.80
2	247.66	250.00	1.40	3.88	-2.33
3	252.76	250.00	1.00	2.54	2.76
.	.	.	.	.	.
1400	252.08	250.00	1.59	3.27	2.08

$RMSE = 4.04Kg$ %RMSE= 1.62

Mean $se(\hat{T}_{pu})$=1.11Kg (Prop. Error SE)

In Table 1, each row summarizes one simulation. The total Pu estimate is in column 2 and the POE-based estimate of $(se(\hat{T}_{Pu}))$ is in column 3. Column 4 is a "Goodness of Fit" statistic computed by the fluence regression. When the data fits the statistical models presented in the previous sections, the GOF statistic should be approximately 1. Theoretically a value above 1.5 indicates significant lack of fit. Note that GOF statistic is typically higher than it should be. This is because the simulated data contains non-independent errors (i.e. contamination, RPE, measurement calibration errors) that are being detected by this statistic. The GOF statistic measures how severely the data deviates from the assumed model, which assumes independent, zero-mean errors. We expect the GOF statistic to be larger than its theoretical value for real data, because real data will have calibration and contamination problems to some degree. At the bottom of the table are the average RMSE and %RMSE over the 1400 simulations. As one can see from

the example, the POE-based se is too small; actual errors are more than 3 times larger in this example (average RMSE of 404 compared to average POE-based se of 111).

Table 2 decomposes the generic scenario RMSE into the three main components. Bias describes a general bias that may be due to any combination of global regression model misfit, measurement contamination, calibration bias, or RPE. Although significant, this bias is relatively small. Because bias is influenced by many inputs, it can be altered significantly by changing the above-mentioned inputs.

Table 2. Decomposition of RMSE for Generic Scenario

Source of Variation	Absolute (g)	Relative % of T_{pu}
Bias: Due to various sources.	74	0.30
Between Model SE: Global Model Uncertainty	111	0.44
Within Model SE: All Sources Except Global Modal Uncertainty	381	1.52
Decomposition of Within SE:		
–Random Measurement Error:	50	0.20
–Measurement Calibration Error:	30	0.12
–Contamination Error:	25	0.10
–Location Error	0	0
–Reactor Physics Error	375	1.50
Total RMSE*	404	1.62

$$* \; RMSE = \sqrt{\text{Bias}^2 + \text{SE}^2}$$

The between model error describes the bias in the global regression model when applied to the population of seven global fluence models used as truth. This error is the second largest component of error, but still relatively small. The major difference in these results is not in the se's (which are all approximately 1.50%), but in the bias. Biases range from -0.18 to 0.81 for the seven fluence fields. These seven biases correspond to a between model se of 0.44%, which is about half of the maximum, and describes the error in the population of "true state" fields.

The third and largest source of error, "within model" error, comprises all other sources of error in the data (contamination, measurement error, calibration error and local RPE). By switching specific sources of error off and re-running the Monte Carlo program, one can decompose this within model error further, into its major sources. The results of the Monte Carlo runs are also listed in Table 2. The decomposition of within model error shows that local RPE is the largest contributor to Total Pu error, producing an error of 1.50%. The next largest contributor is SIMS and TIMS random measurement error, producing a contribution of 0.20%. SIMs and TIMs calibration error produce a 0.12% error, which roughly corresponds to the magnitude of calibration error with respect to random error. The assumed contamination for SIMS and TIMS measurements produces a 0.10% error, which is the smallest of the measurement error effects. Finally, sampler location error is less than 0.01% and we have therefore set it to zero. Even a unrealistically

large location error of 2cm only increases the se in total Pu to 0.03%, so it is safe to eliminate location error as a serious source of estimation error.

RPE is the largest error source (1.50%), and the contributions of the key components of RPE are fuel pin radius (1.05%), fuel temperature (0.30%), graphite density (0.24%), equivalent boron concentration (0.16%), graphite temperature (0.63%), and specific power (0.17%), and WIMs Code (0.25%).

Figure 1 illustrates the effect of different sources of error. In the first five plots, the local errors are illustrated for each of the measured isotopic ratios. In these plots, the dashed line represents the "true" relationship between fluence and the ratio, while the solid line represents the relationship used in the local estimation procedure. The difference between the two curves is local RPE. The points scattered about the dashed line represent measurement error. The local estimation procedure uses the solid curves to produce a best estimate for Pu-fluence at each sampling location.

The last plot in Fig. 1 represents the results of the global regression procedure. In this plot the estimated fluence is plotted against true fluence at each sample location. Heasler et al. [4] includes more details, including some additional worst-case results in which all error sources were increased to extremely large values, and and in one example, the RPEs were forced to vary in the same direction. So, for example, both fuel pin radius and temperature would be forced to have positive errors. These addition results resulted in RMSEs of 2% to 6% (the 6% is arguably worse than worst-case).

4. DISCUSSION

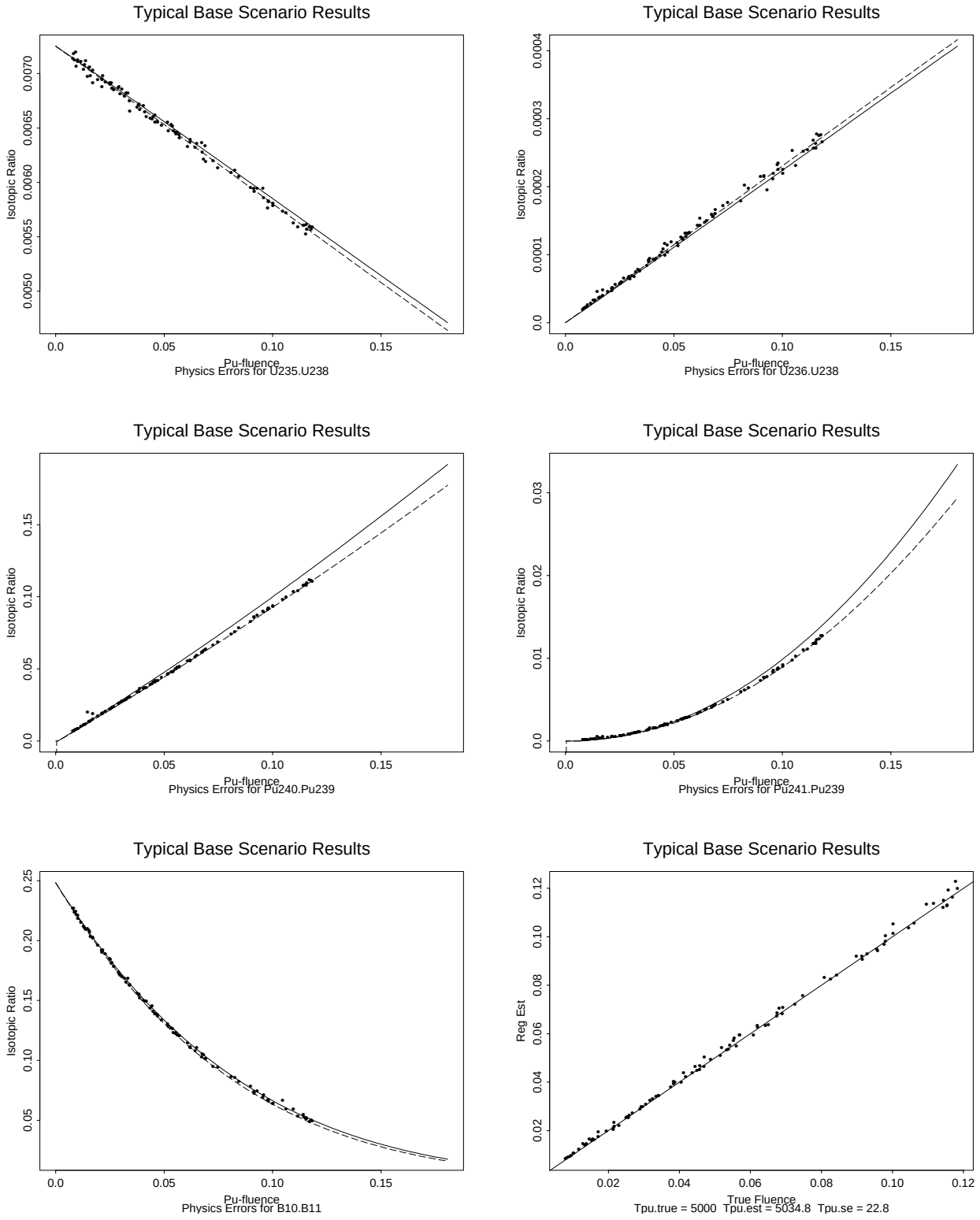
We presented GIRM and a simulation method for assessing its RMSE. Local and global RPEs have been considered, with local RPEs appearing to be the largest error source. Although we focused on estimating Pu production for a specific example reactor, clearly the approach can be applied to almost any graphite reactor.

Although some basic knowledge of the reactor operating parameters should be available, the methodology can be applied to reactors for which no detailed operating history is available. If a sufficient number of samples (say 100 or so) can be taken, total Pu production can be estimated to within a few percent. If, on the other hand, a basic shape of the fluence profile can be calculated from operating history, a better estimate of total Pu production can be obtained with fewer samples. For example, in the Trawsfynydd exercise [3], estimates were produced from approximately 30 samples.

GIRM is a unique blend of reactor physics models and empirical data. Any error associated with the local (WIMS) model or global 3D model has important consequences for the resulting total Pu estimate. The simulation quantifies the effects of error sources, and allows us to evaluate sampling schemes. Benchmark comparisons [3] add to our confidence in the performance claims here. However, RPE remain the dominant error source and it is important to understand potential pitfalls.

The local RPE considered here are due to improperly specified parameters; implicitly we have assumed that the model is truth, provided the parameter values (such as fuel pin radius) are accurately specified and that numerical accuracy is good. In experimental

Figure 1. Local and Global Estimation Results



data, any outlying ratios in curves such as in Fig. 1 are assumed to be due to sample contamination or other measurement problems. The fact that other models give similar results provides some assurance that pure model error is minimal, but the main assurance is the experimental data such as in [3].

The impact of global (KENO) model error in predicting the 3D fluence field depends critically on the GIRM strategy involving fitting the local Pu estimates using a special set of basis functions, and sample size/location choice. Because there are constraints on the 3D shape of the flux (it can have spikes near control rods, but generally it is relatively smooth), it is reasonable to assume that not too large a basis set is required for a good fit (allowing accurate integration of the local Pu estimates). Also because of known constraints on flux shapes, we did not allow arbitrary true flux shapes. Instead, in the simulations, the true flux was randomly selected to be one of the seven shapes described. All seven shapes were relatively smooth, and it is possible in a new application (although highly unlikely) that the true shape would have sharp peaks and valleys, leading to understatement of the 3D model error. The use of basis functions to fit the 3D fluence field has several advantages compared to using the single best-guess field shape. For example, in the evaluation presented here if the true field shape were assumed to vary randomly around the best-guess field shape then the estimated impact of global model error would most likely have been too low. In addition, the sampling was optimized under the assumption that the true shape was among the seven choices. However, a linear combination of the basis functions can accurately approximate an arbitrary field if enough terms are used. In the example provided, because sufficient detail was provided for the target reactor to have a "best guess" shape that is close to the true shape, it is unlikely that the 3D shape could be even as misspecified as assumed in the simulation, so if anything, experts believe that Table 2 overstates, but not drastically, the impact of global model uncertainty. Therefore, performance claims have not yet been attempted for the situation where initial sample results indicated an extremely rough 3D flux field, leading to a request for more samples at key locations. So, performance claims for the situation in which test samples not used for fitting indicated a problem with the 3D fit would be desired on an as-needed basis.

REFERENCES

1. R. J. Talbert, B. D. Reid, P. G. Heasler, D. C. Gerlach, C. M. Heeb, and K. A. Pauley, Accuracy of Plutonium Production Estimates from Isotopic Ratios in Graphite Reactors, PNL-RTC-0693-REV1-DEL, Feb. 1995.
2. B. D. Reid, Morgan, E. F. Love, D. C. Gerlach, S. L. Peterson, Livingston, L. R. Greenwood, and McNeece, Graphite Isotope Ratio Method Development Report: Irradiation Test Demonstration of Uranium as a Low Fluence Indicator," PNNL-13056, Sept. 1999.
3. B. D. Reid, D. C. Gerlach, P. G. Heasler, and J. V. Livingston, "Trawsfynydd Plutonium Estimate," PNNL, Sept. 1997.
4. P. G. Heasler, B. D. Reid, C. J. Gesh, D. C. Gerlach, S. L. Petersen, C. K. Bayne, and T. L. Burr, Estimation Procedures and Error Analysis for GIRM, PNNL-14332, July 2003.

Hydrocarbon exploration risk evaluation through Uncertainty and Sensitivity analyses techniques

Paolo Ruffo [a], *Livia Bazzana* [a], *Alberto Consonni* [a], *Anna Corradi* [a], *Andrea Saltelli* [b] *and Stefano Tarantola* [b]

[a] ENI – E&P Division, Via Emilia 1, 20097 S. Donato Mil.se (MI), Italy;
[b] JRC – IPSC , 21020 Ispra (VA), Italy
E-mail: paolo.ruffo@agip.it, livia.bazzana@agip.it, alberto.consonni@agip.it, anna.corradi@agip.it, andrea.saltelli@jrc.it , stefano.tarantola@jrc.it

Abstract: The evaluation of the exploration risk in the oil industry is a fundamental component of the decision process related to the exploratory phase. In this paper the two basic components of the exploratory risk: trap *geometry* and trapped hydrocarbon quantities (*fluid*), are compounded in a single coherent uncertainty and sensitivity approach. The results clarify that the model geometry influences each Petroleum System Modeling step and that the *geometric* uncertainty is correlated with the *fluid* uncertainty. The *geometric* uncertainty evaluation makes use of geostatistical techniques, that produce a number of possible realizations of the trap geometry, all compatible with available data. The evaluation of the *fluid* uncertainty, through a Montecarlo methodology, allows to compute the possible quantities of oil and gas, generated in a basin and migrated from the hydrocarbon source location to each single trap. The final result is the probability distribution of oil and gas for each trap in the basin, together with other useful indicators like: the hydrocarbon filling probability map, the closure probability map, the drainage area probability map, the spilling paths probabilities, the trap filling scenarios.

Keywords: Hydrocarbon exploration, petroleum system modeling, uncertainty, sensitivity, risk evaluation

1. INTRODUCTION

In the hydrocarbon exploration activities the main goal is to find traps where oil and/or gas were accumulated and retained in quantities that are greater than a variable economic threshold. Neglecting all the economical variables that come into play, the basis of any drilling decision is associated to the presence/absence of hydrocarbons (hydrocarbon risk) in the potential traps of the basin. This evaluation is the result of the joint efforts undertaken by a team of geologists, geochemists, geophysicists and engineers, in trying to get the best picture of the prospects that may be potentially drilled.

This study can be schematically splitted into two components: the *geometric* one and the *fluid* one. The first one is the object of Basin Modeling (BM) [*] activity, that gives a 4D (in space and time) description of the basin status and evolution. The *fluid* one is the objective

[*] Historically the term "Basin Modeling" has been used with slightly different meanings. In this paper it is used to describe the rocks properties distribution in space and evolution in time. In practice this is strictly correlated to the fluid properties distribution and evolution and therefore a clear distinction between BM and PSM is partially subjective.

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

of the Petroleum System Modeling (PSM), that produces the history of the geologic processes that led to generation and accumulation of hydrocarbons in the current traps.

2. METHODOLOGY

Modeling geological processes is subjected to uncertainty because input data are scarce and imprecise and also because the used modeling algorithms are an approximation of the true geological processes. This observation implies that a probabilistic approach is needed to account for the lack and imprecision of knowledge, enabling at the same time to compute the "hydrocarbon risk" for each trap of a basin.

Regarding Basin Modeling, geostatistical techniques can be used to model the uncertainties of both geologic layer geometries and facies* properties. Geometric uncertainty is related to the process of depth converting interpreted seismic time reflections into depth reflectors, as a function of the uncertainty of seismic propagation velocities. The evaluation of the uncertainty of facies properties distribution refers both to seismic data attributes and to sedimentological interpretation. In this paper we are considering only geometric uncertainty, as the methodology to take into account facies uncertainty is quite complex to set up, in basins case.

As the PSM is an inverse problem and the data available is scarce and uncertain, we have to deal with multiple possible "realizations" of the basin model as well as of the petroleum system evolution. Even if all the "realizations" are fitted to known available data, this calibration process just reduces the space of possible solutions but it is unable to justify by itself the choice of a unique, or most probable, or optimal solution.

Each of the phases of the PSM (including Basin Modeling) contributes to the overall uncertainty and can be explored with a sensitivity approach. As shown in Ref. 1, where a brief summary of the different approaches is presented, almost all the papers deal with the uncertainty evaluation of only some of the phases of PSM. In fact, aside to the great amount of CPU time needed, the main difficulty in applying Uncertainty and Sensitivity Analyses (UASA) to the entire PSM is given by the complex management of the complete workflow.

Another great source of uncertainty is given by the assumptions (the conceptual model) that are practically needed in the inversion of the scarce data available. These alternative conceptual assumptions, or geological hypotheses, are dealt by means of scenario variables [2]. To drop a scenario, or one of the combinations of different scenarios, means to hide a component of the uncertainty and thus to increase the risk of biased choices.

Our approach consists in considering all possible combinations of scenario variables and in producing, for each scenario, as many realizations as needed for uncertainty and sensitivity analyses. With regard to uncertain input variables these can be numerical (continuous or discrete) or categorical (e.g. a set of functions or maps).

Moreover the non linearities hidden in the modeled geological processes are such that there are threshold values of some variable or combination of variables that may trigger one event (e.g. generation of gas, …). For this reason the use of UASA approaches that assume linearity or continuity hypotheses need great caution, so that, in our opinion, it is preferable to use a Montecarlo approach to the extent that is allowed by hardware constraints.

---

* A "facies" is a rock layer that differs from the others (as in composition, age, …) because of its formation .

2.1. Basin and Petroleum System Modeling

A general introduction to BM and PSM techniques can be found in many books (see for example Ref. 3), in the following only the main modeling steps are mentioned and the work-flow that has been applied is described (see Fig. 1).

The BM (for the purposes of this paper) consists of: the shape definition of the geologic structures; the spatial distribution of the geological properties of interest for each geologic layer in the basin model; the structural evolution of the basin during geologic time (in this study only vertical compaction of sediments due to overlying sediment load was taken into account); the definition of the history of the heat flow at the basis of the sediments (coming out from the basement, in the earth upper crust).

The PSM can be grossly subdivided into: the description of the evolution of the Pressure & Temperature (P&T) fields in the sediments; the history of the Generation & Expulsion (G&E) of the hydrocarbons (oil and gas, in the simplest case) from source rocks; the Migration of hydrocarbons from source rocks to reservoirs and the preservation of Trapping conditions of the hydrocarbons throughout the evolution of the basin (M&T).

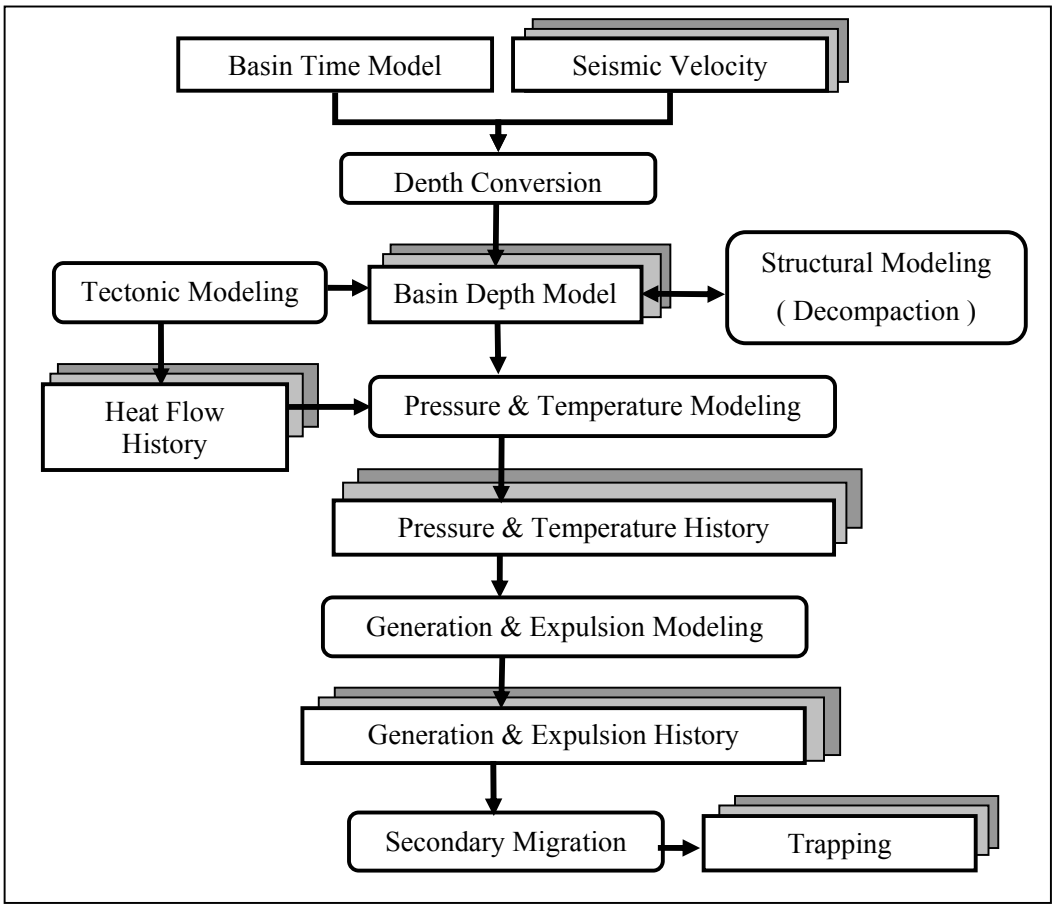


Figure 1. Workflow of the Basin Modeling and Petroleum System Modeling applied in this study. Boxes with rounded corners represent modeling steps. Rectangular boxes represent data. Repeated shadowed rectangular boxes represent multiple realizations/simulations of that data.

2.1.1. Basin Modeling

At the basin scale the depth model is suitably built using the *layer-cake* vertical depth conversion method from interpreted seismic time maps. This consists in the transformation of the time "thickness" into a depth thickness using the corresponding seismic layer velocity. In this way the depth model is constructed layer by layer, from top to bottom.

Velocity maps are obtained using geostatistical techniques because they allows to produce an optimal velocity map integrating different data sources: seismic velocities, Well velocity measurements and geological knowledge (called *a priori information*).

This approach has been applied in the current study to produce a basin model compound of 7 interfaces and 6 layers (see Fig. 2a). Layer 1 (dark blue) and layer 4 (brown) are the hydrocarbons source rocks (respectively "Source 1" and "Source 2"), while layer 3 (green) is the carrier for hydrocarbons migration into potential traps (in the same layer). In Fig. 2b the position of section AA' is shown together with the top of layer 3 and the nine major traps.

The description of the resulting Basin Model was completed with information about layers parameters (porosity, permeability, thermal conductivity, …) derived from Wells, literature and sedimentological studies. The application of a decompaction law for each layer of the basin, allowed to recover the basin evolution history, that is the 4D Basin Model. In practice this is a set of 3D models, one for each selected geological time step, describing the evolution of the 3D Basin Model through the geologic time. According to the 4D Basin Model, a heat flow history can be established from tectonic modeling and Well data [4], resulting in a set of heat flow maps, one for each selected geologic time step.

2.1.2. Thermal & Pressure Histories

The distribution of temperature within the sediments is controlled by: the regional geo-thermal regime (heat flow at the base of the sediments and paleo-temperature at the top), the thermal properties of the rocks and the fluid movements through rock pores. The geothermal regime changes during the basin evolution in connection with variations of the basal heat flow and paleo-temperature fluctuations.

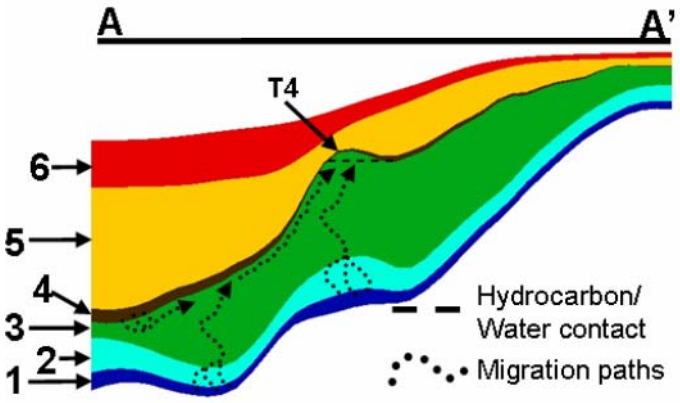


Figure 2a. Section of the basin model
Source Layers: 1 and 4. Carrier and Reservoir layer: 3.

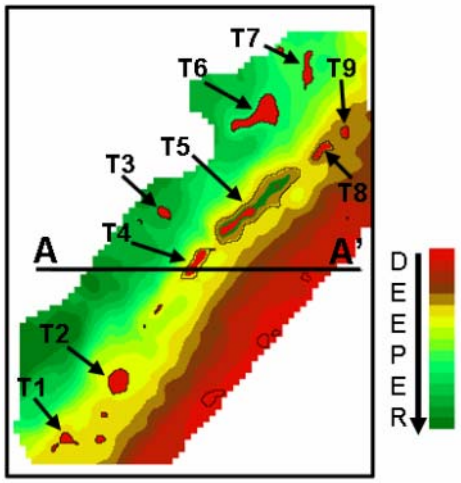


Figure 2b. Top of layer 3
Potential Traps: T1, T2, T3, … , T9

At the same time, as the basin geometry changes due to accumulation and compaction of sediments, the bulk thermal properties of rocks are varying too. The compaction, resulting from depositional overburden load, reduces rock pores volume and squeezes the water out, accordingly to the permeability of the layers. When a low permeability layer is involved, water pressure departs from hydrostatic equilibrium and an overpressure results. The duration of the pressure disequilibrium is controlled by the filtration velocity, that is by the overpressure gradient and the permeability of the layers. The fluid flow is modeled through a Darcy equation. The computation of thermal and pressure histories is performed solving the corresponding equations of energy and momentum conservation through a finite element approach. The resulting pressure and temperature fields are obtained for each geologic time step.

2.1.3. Generation & Expulsion Histories

The "source rock" is a part of a low permeability layer (the *source rock layer*) and it is such that within it there is an abundance of organic matter that is playing the role of kerogen. The initial kerogen degrades through a number of parallel reactions (*primary* cracking) into hydrocarbons, which are more stable components (oil and gases). This is followed by *secondary* cracking, which transforms oils and wet gases into lighter components (dry gases) and coke. Both *primary* and *secondary* reactions are assumed to be independent of pressure and their temperature dependence is given by an Arrhenius type equation.

Expulsion of the generated hydrocarbons from the source rock is a very complex and still poorly constrained process. It is modeled as a multiphase flow (hydrocarbons and water) within a low permeable porous medium (the source rock itself), controlled by the relative permeability of each fluid phase. In practice hydrocarbon flow is driven by the overpressure resulting by the compound effect of the sediment compaction and of the hydrocarbons generated (by the transformation itself).

The result of this modeling step is a set of expulsion maps, one for each selected geologic time step and for each hydrocarbon component (oil and gas in the simplest case). These maps represent the result of the expulsion out of the *source rock layer*.

2.1.4. Hydrocarbon Migration & Trapping

Once the hydrocarbons are expelled out of the source rock layer, they move (secondary migration) along permeable rocks until they reach a trap. The simplest approach to model secondary migration is based on the so-called 'ray-tracing technique'. This technique assumes that hydrocarbons move, because of buoyancy, just at the top of a permeable layer overlain by an impermeable one (acting as a seal) and following the steepest path. Subsequent processes, like spillage and leakage, may cause hydrocarbons to migrate out of traps. All the processes are modeled over the geologic time scale, taking into account the evolving basin geometry and changing properties of rocks and fluids.

The final results of the whole PSM workflow (modeling of temperature and pressure, generation and expulsion, secondary migration and entrapment) are the total amounts of hydrocarbons (oil and gas) that fill at the present time each trap of the basin area.

2.2. Uncertainty and Sensitivity Analyses

As already mentioned, this study addresses both *geometric* and *fluid* uncertainty and sensitivity evaluation in a 3D study. Even if it is possible to add more sources of uncertainty, this is the first time, to our knowledge, that such a complete approach is proposed.

2.2.1. Basin Modeling Uncertainty and Sensitivity Approach

The main sources of uncertainty of a basin depth model are the interpretation of seismic times and the estimation of seismic velocities. Disregarding the radical error of having misinterpreted seismic reflection (travel) times, the greatest uncertainty is associated to seismic velocities, as they are indirectly estimated from seismic signal coherency [5].

The geostatistical techniques used to produce an optimal velocity map, allow also to generate an infinite number of equally-probable simulated velocity maps. In this study the seismic velocity field of the most critical layers for the depth conversion was simulated using a specific geostatistical technique for probabilistic depth conversion [6].

As a result of this approach 100 simulated depth models have been computed, all of them geostatistically equally-probable but quite different from one another if compared with their possible effect on the PSM result. As 100 depth models were too many to be dealt with, a selection of 8 representative depth models was performed in such a way that they "optimally" sampled the "uncertainty space of the basin depth model". They are the ones that are the most different from one another in the main area of interest, that of trap T5, the biggest one.

A similar approach was used to select 4 heat flow maps, that is the selected ones are the most different from one another and they optimally sample the heat flow uncertainty space.

Both the depth models and the heat flow maps have been treated as scenario variables, which means that all the combinations among them have been considered, producing 32 different 4D basin models as input to the following PSM study.

2.2.2. PSM Uncertainty and Sensitivity Approach

In practice the P&T phase was run 32 times, one for each scenario, and for each of these P&T runs, the subsequent phases (G&E and M&T) were run 32 times. This produced 1024 runs with a LP-τ sample on all uncertain input variables associated to: Basin Model, P&T, G&E and M&T. This UASA approach extends that proposed in Ref. 7 where: the geometry of the Basin Model was kept fixed (instead of having 8 scenarios), the 3D P&T runs where only 8 (instead of 32), the G&E runs where 32 in total (instead of 1024).

Moreover this paper applies the method proposed in [8] as it allows to compute the uncertainty of the hydrocarbon trapped quantities (oil and gas) and evaluate, at the same time, the sensitivity indices of the first order, using the same set of 1024 runs. In order to analyze non linear effects between variables, a Neural Network (NN), trained on the set of 1024 runs, was then used to compute the sensitivity indices of the first and second order and the total effects, using the method of Sobol' with the extension described in Ref. 9 and Ref. 10.

2.2.3. Uncertainty and Sensitivity Evaluation

Fifteen uncertain variables were used, grouped in 4 subsets:

- Basin (2 variables) : 8 model *geometries* combined with 4 *heat flow* maps to get 32 Basin Models;
- Source (6 variables) : *Total Organic Carbon*, *Porosity-Stress* Curve, *Water Threshold Saturation*, for Source 1 (layer 1) and Source 2 (layer 4);
- Migration (4 variables) : *Expulsion Efficiency* from Source 1 and Source 2, *Leakage of Gas* and *Leakage of Oil* from a trap;

- Trap (3 variables) : *Net to Gross* ratio; *Water Irreducible Saturation*, *Thickness* of the reservoir layer.

The uncertainty evaluation of the trapped oil and gas quantities in each of the 9 main traps (Fig.2b) is based on LP-τ sampling of 13 uncertain variables for each of the 32 basin model scenarios. This means that the output variables are 18 (oil and gas volume in each trap).

At the same time first order sensitivity indices were computed in order to identify the driving factors of the G&E and M&T processes that cause an accumulation of oil and/or gas. As mentioned we have applied the method based on "State Dependent Parameter models" (SDP) [8] to estimate sensitivity indices of first order.

In order to estimate sensitivity indices of the second order it would be necessary to run a huge number of times the entire BM and PSM workflow, requiring an unaffordable CPU time. This suggested the use of a NN approach [11] to model the global PSM workflow. For this purpose the NN was trained on the 1024 uncertainty runs, using all the 15 input variables and the 18 output variables. The goal of this NN was the computation of Sobol' indices and in particular of second order indices, in order to better understand non linear interactions between variables.

3. RESULTS

The analysis of the results shows that, as expected, some non linear relationships among variables exist (Fig. 3a shows an example for an input and an output variable). Moreover some scenario yields "discontinuous" results, as (Fig. 3b) for model geometry 3 where trap T4 never contains oil and always gas, while all the other model geometries "convey" both oil and gas in trap T4 (in the Fig. 3b they are in gray).

The spilling from one trap to outer regions of the modeled area or to other traps is quite important in the correct evaluation of the hydrocarbon trapped quantities. This is better understood looking at the spilling scenarios, shown in Fig. 4a, where also the probabilities of each spilling path has been computed, using a statistics of 30 model geometries (including the 8 ones used for UASA). It is important to remark that the spilling paths coming out of a single trap are alternative cases, as only one of them could be the real one.

Taking a look to the hydrocarbon quantities that may have filled a trap, as is shown for trap T5 in Fig. 5a for gas and in Fig. 5b for oil, the total amount coming out from all the simulations summarizes the contributions due to the different spilling scenarios. For example the case "no spill-in" represents all the simulations where trap T5 was filled only directly from the drainage area, while the case "T4" is the set of all the simulations in which trap T5 was filled both directly from source and from the spilling of trap T4. For the filling of trap T5, only the 6 scenarios listed in Fig. 5a and in Fig. 5b are possible and again they represent alternative cases as only one of them could be the real one.

Other useful statistical results are the *probability maps* that summarize the *probability of presence* of a characteristic of interest for each point of that map. For example Fig. 4a shows the "Hydrocarbon Filling Probability" map that represents the probability that each point has to be inside an hydrocarbon (gas and/or oil) accumulation. While the "Closure Probability map" measures the probability that each point has to be inside a *closure*, defined as the maximum volume available to hydrocarbon accumulation for that trap (trap T5 in Fig. 4b) before that a spilling out will take place.

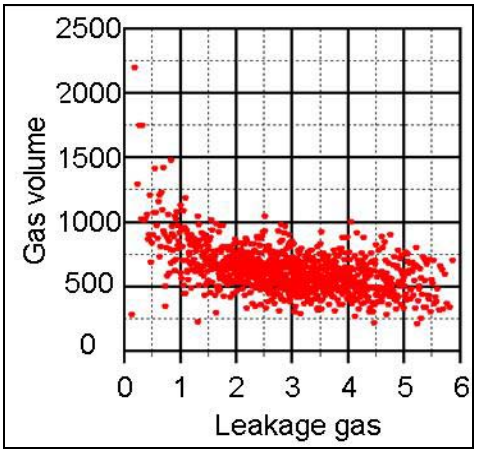


Figure 3a.
Trap T5 – Non linear relationship

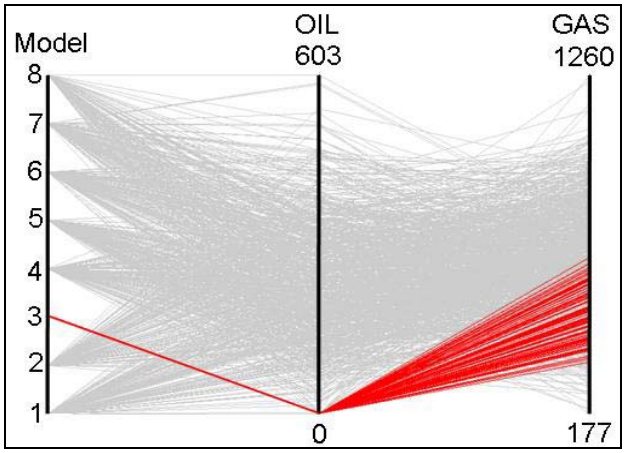


Figure 3b.
Trap T4 for Model 3: never oil and always gas

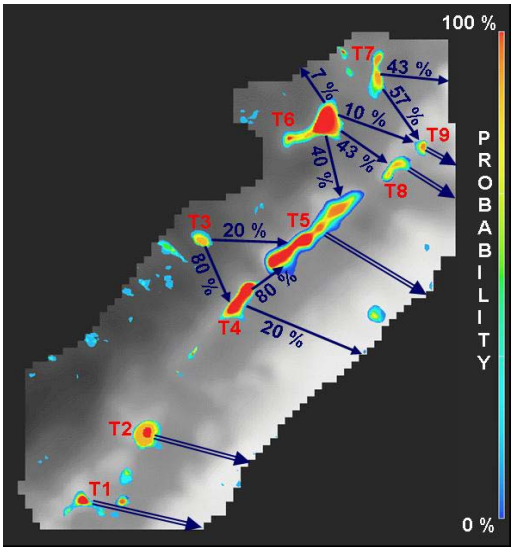


Figure 4a. Top Layer 3 in gray - Hydrocarbon Filling Probability in color - Spilling Scenarios probabilities as vectors connecting Traps.

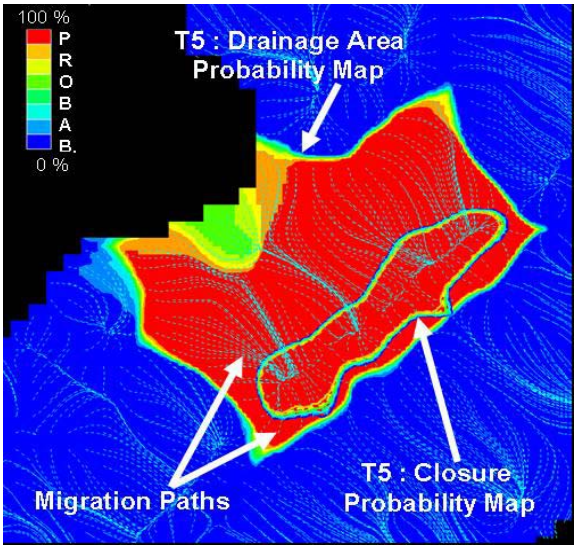


Figure 4b. Zoom on Trap T5 – Migration paths in dashed light blue - Drainage Area and Closure Probability Maps in color.

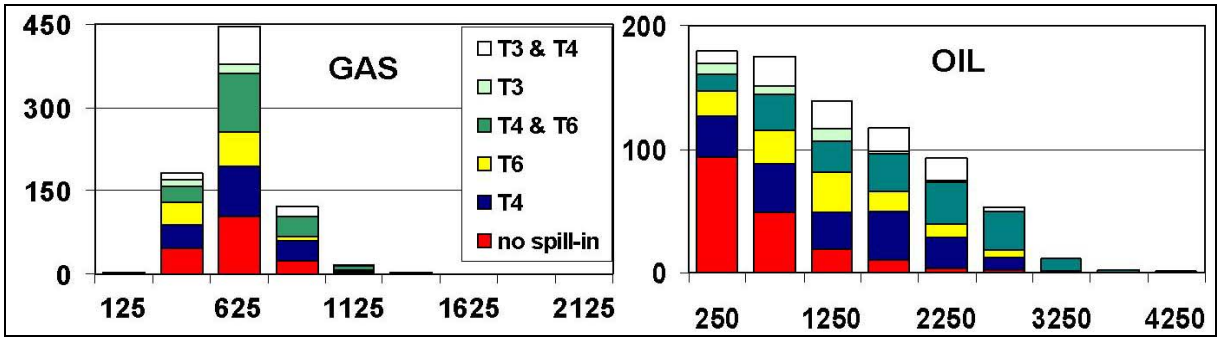


Figure 5a. Trap T5 – Volumes of Gas

Figure 5b. Trap T5 – Volumes of Oil

Colors are related to the different trap spilling scenarios (e.g. T4 means: gas/oil from spilling of T4)

The Filling Probability can be computed even if the model geometry is fixed, while the Closure Probability requires a number of different model geometries (in the example 30). The same is true also for the "Drainage Area Probability map", that describes the probability that a point has to be inside the drainage area of a trap (as trap T5 in Fig. 4b).

As regards the trapped hydrocarbon quantities, the result of the study is that all the traps are filled by gas and, as only traps T4 and T5 have a median value of the oil probability distribution greater than zero, only these actually contain both oil and gas. To be precise only trap T6, among others, has the 3rd quartile of the oil probability distribution greater than zero.

The analysis of the first order sensitivity indices (computed with the SDP method) can be only shortly summarized. The observation just mentioned on gas prone traps is reflected in the sums of the first order indices, in fact these are less than 0.40 for the (improbable) oil accumulations of these traps. In all other cases the sum of first order indices of hydrocarbon accumulations (gas or oil) is greater than 0.70, with the exception of trap T6 which is in the middle with a value that sum to 0.56 for the oil accumulation.

According to the sensitivity analysis, based on (reliable) first order indices, the most critical parameter for the gas prone traps (T1, T2, T3, T6, T7, T8, T9) is the Model Geometry. For traps T4 and T5 the most critical parameter for gas accumulation is related to Trapping Conditions (Net/Gross and Gas Leakage, respectively), while for oil accumulation it is related to Model Geometry for trap T4 and to Source parameters (Porosity-Stress curve of Source 2) for trap T5.

A double comparison was done between the first order indices computed with SDP method on real data (from model simulations) and those computed with SDP and Sobol' methods on data produced by the trained NN. The result is that the first two most critical parameters are found all the same as far as gas accumulations are concerned and also oil accumulations but only for traps T4 and T5. This justify the use of second order indices, computed on NN data, to understand possible correlations among variables. In synthesis this analysis reveals as critical the interactions among Model Geometry and : Source parameters (TOC, Porosity-Stress Curve), Migration parameters (Expulsion Efficiency, Gas Leakage) and Trapping parameters (Net/Gross, Reservoir Thickness).

4. DISCUSSION

The current study was possible thanks to the porting of the G&E modeling phase on a parallel platform, in practice a cluster of Linux workstations that amounts to 24 CPU (including a Linux cluster of 16 CPU). The CPU time needed to perform the study, with this setting, was about 84 hours (72 hours only for the 1024 runs of the G&E phase).

Obviously the use of a NN could reduce the CPU time needed, but as mentioned above, this can not be a straightforward or blind choice, in fact the analysis of the results confirms that the NN has a tendency to linearize the model behavior and has difficulties in modeling categorical variables. This do not exclude that a more advanced NN could give better results.

As the sensitivity study has suggested, the Model Geometry plays an important role in the different modeling phases, its effect on the results is non linear and it may also introduce discontinuities in the space of results. All this claims for a more systematic use of Model Geometry uncertainty, which means the ability to manage at least 32 model geometry simula-

tions, that in turn requires the setting of a semi-automated procedure in substitution for the manual one used in this preliminary study.

The scenario approach is, as already mentioned, the only way to properly manage an interpretative uncertainty, but as the analysis of the results has shown (see Fig. 5a and Fig. 5b) it is not possible to identify *a priori* a set of parameters that will produce a *pessimistic* result or an *optimistic* one. On the contrary we have to look for *pessimistic/optimistic* cases only *a posteriori*, applying a careful analysis of the behavior and of the characteristics of the basin model and of the petroleum system under study.

5. ACKNOWLEDGEMENTS

Authors wish to thank ENI – E&P Division for the permission to publish this paper. They are particularly grateful to ENI colleagues: Paolo Bozzoni, Alfredo Vanelli and Enrico Carena, for the extensions of the PSM routines to the probabilistic approach, and to the JRC colleague: Marco Ratto for his assistance with the sensitivity analysis technique *SDP models*. All the software used in the study are proprietary of ENI, with the exception of : the engine of the secondary migration (based on the software SEMI [12] by Sintef Research) and UASA routines (by JRC).

REFERENCES

1. P. Ruffo. Probabilistic Petroleum System Modelling. An Overview. Abstract in the Proceedings of the EAGE Workshop on *Multidimensional Geological Modelling: from Cutting Edge Research to Application in the Field*, Firenze, 2002.
2. P. Ruffo, S. Tarantola, A. Corradi, and D. Grigo. Uncertainties evaluation in petroleum system modelling. Abstract in the Proceedings of IFP International workshop on *Quantitative Basin Evaluation, a Tool for Lowering Exploration Risks*, Paris, 2001.
3. D.H. Welte, B. Horsfield, and D.R. Baker (Eds.). *Petroleum and Basin Evolution*. Springer-Verlag, Berlin, 1997.
4. Y.Y. Podladchikov, S.M. Schmalholz, D. Cavazza, and P. Ruffo. Palaeo-Heat Flow Assessment using Automatic Inverse Modelling of Sedimentary Basin Formation. Abstract in the Proceedings of the EAGE Workshop on *Multidimensional Geological Modelling: from Cutting Edge Research to Application in the Field*, Firenze, 2002.
5. O. Yilmaz. *Seismic Data Processing*. Vol. 2 of Investigations in Geophysics. SEG, Tulsa, 1987.
6. P. Ruffo, and V. Scola; Structural Model Uncertainty of a Prospect - A Geostatistical Approach. Abstract in the Proceedings of the 54th EAGE Meeting, Paris, 1992.
7. S. Tarantola, A. Corradi, P. Ruffo, and A. Saltelli. Global sensitivity analysis techniques for the analysis of the oil potential of sedimentary basins. Long Abstract in the Proceedings of *3rd International Symposium SAMO*, Madrid, 2001.
8. M. Ratto, S. Tarantola, A. Saltelli, and P.C. Young, Accelerated estimation of sensitivity indices using State Dependent Parameter models, *in these proceedings*.
9. A. Saltelli, Making best use of model evaluations to compute sensitivity indices, *Computer Physics Communications*, 145: 280-297, 2002.
10. A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*. John Wiley & Sons publishers, Probability and Statistics series, to appear in March 2004.
11. B. Muller, and J. Reinhardt. *Neural Networks – An Introduction*. Springer-Verlag, New York, 1991.
12. M. Hamborg, A. Tømmerås, Ø. Sylta, T. Kjennerud, and H. Borge. *SEMI 4.23 Users manual*. SINTEF Petroleum Research, Trondheim, 2003

Model Validation in Chemical Process with Multiple Steady States

C. M. Villa, J. P. Masy, F. A. Castillo, L. H. Thompson, and J. W. Weston

The Dow Chemical Company, 2301 N. Brazosport Blvd, Freeport TX 77541, U.S.A.
E-mail: cmvilla@dow.com, jmasy@dow.com, facastillo@dow.com, lhthompson@dow.com, jwweston@dow.com

Abstract: Pilot plant work is considered a reasonable intermediate step when chemical processes are moved from lab to commercial scale. Unfortunately, pilot plants are not easy or cheap to operate and modelers are faced with the problem of getting the most information out of the least expensive pilot plant experimental plan. We have developed a kinetic model for a catalytic reaction mechanism using lab scale experiments in an Accelerating Rate Calorimeter. Under isothermal conditions in a continuously stirred tank reactor, the catalyst exhibits three steady states, one of which is unstable. Our kinetic model is now part of steady state and dynamic models for reactors in the pilot plant. In this work, we compare the performance of our mechanistic models against that of empirical models obtained by analysis of the chosen statistical experimental design, we point out the potential pitfalls of the empirical models, and we suggest ways around those pitfalls for systems with complex dynamic behavior. The combination of mechanistic and empirical modeling offers a balanced approach. Mechanistic models provide insight that is useful in dealing with complex dynamic behavior, while empirical models allow analysis and optimization of system responses for which mechanistic models are not available.

Keywords: pilot plant, process scale up, state multiplicity, catalytic reaction mechanism, chemical reactor, reaction ignition and extinction, mechanistic and empiric models

1. INTRODUCTION

Using only lab data to design and build a commercial chemical plant is considered risky since limitations that are not present at the smaller scale can become dominant at the larger scale. Mass and heat transfer limitations, for example, play an increasing role during plant scale up. Pilot plants are useful in making the transition form lab to commercial scale. Unfortunately, pilot plants are viewed more as means to generate product samples for market testing than as devices to use for model validation, despite the great need to have models that can be scaled up reliably.

In our system, choosing the operating conditions and the catalyst level that will favor the desired stable steady state is difficult, since process economics demand low catalyst levels and this constraint pushes the system towards the low activity region. In other words, a delicate balance has to be established between process economics and the ability to control the process.

In the absence of a kinetic model for a strongly non-linear system like the one considered in this work, a purely empirical approach to experimental planning and plant scale up is at risk of failing or it has to be constrained to small regions in parameter space where linear approximations are valid and other stable steady states are far enough from the desired stable steady state. Unfortunately, these regions are likely to be of little or no value from a practical viewpoint.

2. METHODOLOGY

2.1. Accelerating Rate Calorimeter

In preparation for pilot plant experiments, kinetic modeling work was initiated in an Accelerating Rate Calorimeter (ARC). The ARC was developed at the Dow Chemical Company in the late 70's as a tool for thermal hazard evaluation [1]. Our system involves two reactants and a catalyst. Reactant A is nonvolatile while reactant B is highly volatile. The reaction is strongly exothermic and, for safety reasons, only low concentrations of B are allowed in pilot plant and commercial reactors. In a typical experiment, a spherical container with a mass of 16 grams would be loaded with a sample containing 20 parts per million catalyst, 4 grams of A, and 1 gram of B. The temperature profile observed during the reaction period is shown in Fig. 1. The ARC is particularly useful in systems where strongly exothermic reactions are possible and where a pressure build-up occurs because of the release of gaseous substances. In addition to providing a safeguard against thermal hazards during the phases of production, storage, and transport of chemicals, the ARC is also a valuable tool in thermokinetic analysis. It is with this application in mind that we selected the ARC.

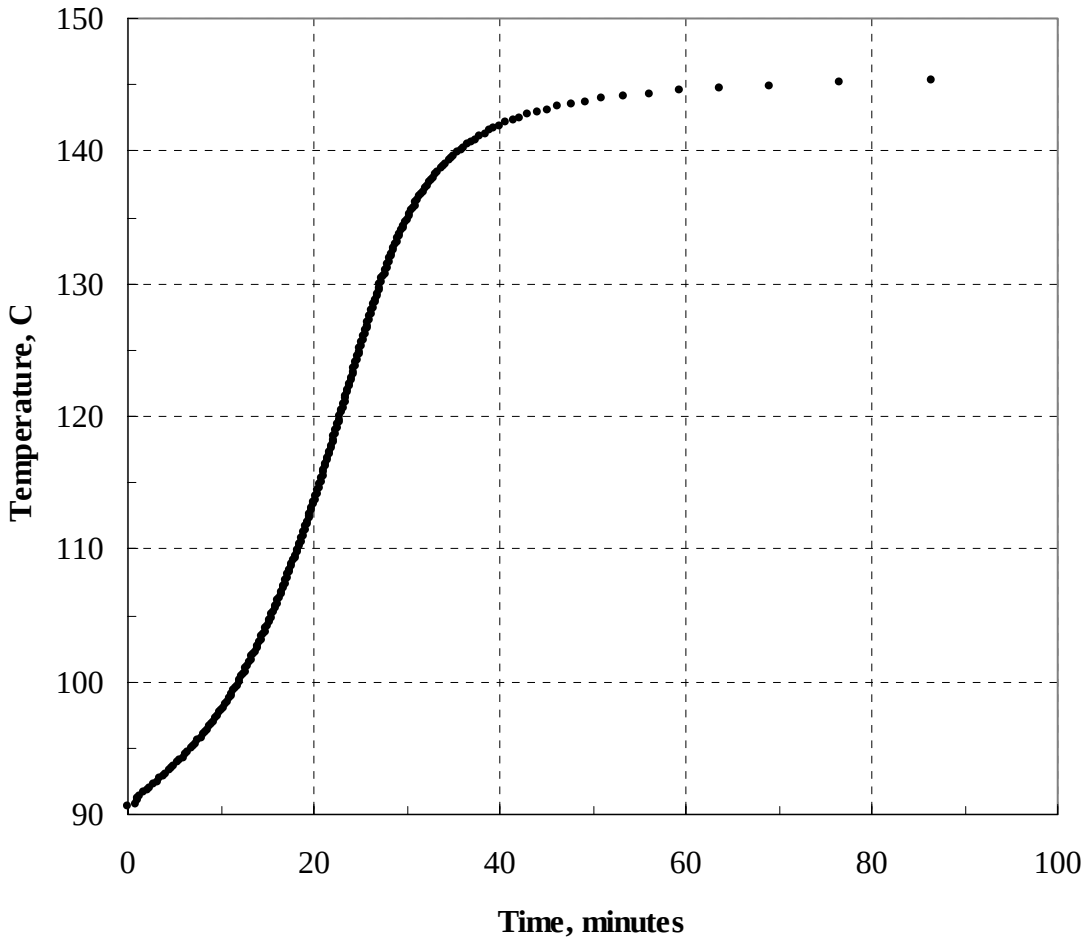


Figure 1. Temperature profile in ARC during exothermic reaction between 4 grams of A and 1 gram of B in the presence of 20 ppm catalyst

With catalyst concentration and initial weight fraction of B as factors, ARC experiments were chosen at the following conditions:

Run Number	Catalyst Level, ppm	Initial weight fraction of B
1	10	0.2
2	60	0.2
3	10	0.4
4	60	0.4

Temperatures observed in ARC experiments were transformed to weight fractions of B during the reaction by making use of Eq. (1), which assumes fixed values for heat capacities and heat of reaction.

$$w_B = w_{Bi} - \frac{\left(m_m C_{pm} + m_s C_{ps}\right)\left(T - T_i\right)}{m_m \Delta H_{rxn}} \tag{1}$$

where w_B is weight fraction of B at a chosen temperature, w_{Bi} is weight fraction of B at the initial temperature, m_m is sample mass, C_{pm} is heat capacity of sample, m_s is sphere mass, C_{ps} is heat capacity of sphere, T is temperature, T_i is initial temperature, and ΔH_{rxn} is heat of reaction.

2.2. Pilot Plant

The pilot plant uses a continuous stirred tank reactor (CSTR) followed by a plug flow reactor (PFR). Additional tanks are available for storage of catalyst, reactants, and product. Near Infrared sensors provide measurements of the concentration of reactant B at points of interest. The pilot plant has been designed for temperatures in the range from 90 to 130 ºC and overall flowrates as high as 50 kilograms/hour. Unlike ARC experiments, the pilot plant experiments were conducted isothermally and with a continuous flow of materials. The factors chosen for study were in this case catalyst concentration and temperature in the CSTR. Conversion of B in the CSTR was chosen as the main system response.

3. RESULTS

3.1. Kinetic Mechanism

The results of ARC experiments show an important trend for this system: when the temperature of the sample reaches a high enough value, often in the neighborhood of 145 ºC, the reaction nearly dies. Based on this, we have suggested a kinetic mechanism where the catalytic sites are assumed to exist in two forms: fast and slow. When the catalyst is fresh, all sites are fast. Exposing the catalyst to reactant B and to high temperatures promote the irreversible transformation of fast sites into slow sites. This mechanism leads to Eqs. (2) and (3) for the rates of consumption of B in the main reaction and the rate of site transformation, respectively.

$$R_M = \left(k_P \theta_P + k_S \left(1 - \theta_P\right)\right) \frac{4 w_B^2}{\left(1 + 4 w_B\right)} \tag{2}$$

$$R_T = k_T \theta_P \frac{4w_B^3}{(1+4w_B)} \tag{3}$$

where θ_P is the fraction of fast catalytic sites, k_P is the reaction coefficient for the main reaction at the fast sites, k_S is the reaction coefficient for the main reaction at the slow sites, and k_T is the reaction coefficient for site transformation.

The rate expressions in Eqs. (2) and (3) can be used in macroscopic mass balances [2] describing the behavior of the reacting system in the ARC. The necessary balances are given by Eqs. (4) and (5).

$$\frac{dw_B}{dt} = -w_C R_M \tag{4}$$

$$\frac{d\theta_P}{dt} = -R_T \tag{5}$$

where t is time, and w_C is the concentration of catalyst.

In the system of Eqs. (1) through (5), the three reaction coefficients depend on temperature, as described by the Arrhenius expression: $k_i = k_{io} e^{-E_i/RT}$. There will be two adjustable parameters for each reaction coefficient, k_{io} and E_i, for a total of six adjustable parameters. As illustrated by Eq. (1), temperature is not an independent variable but is linked to the weight fraction of B. Fig. 2 shows the fit of this model to the experimental data obtained in the ARC, with conversion defined for the ARC runs as $X_B = 1 - w_B / w_{Bi}$.

3.2. CSTR Model

Macroscopic mass balances provide again a description of the dynamic behavior of the reacting system in a CSTR. Eqs. (6) through (8) correspond to balances for B, catalyst, and fast catalytic sites, respectively, assuming the density of the mixture is independent of composition and the reactor is always full:

$$\tau \frac{dw_B}{dt} = w_{Bf} - w_B - w_C \tau R_M \tag{6}$$

$$\tau \frac{dw_C}{dt} = w_{Cf} - w_C \tag{7}$$

$$\tau \frac{d\theta_P}{dt} = (1-\theta_P)\frac{w_{Cf}}{w_C} - \tau R_T \tag{8}$$

where τ is reactor residence time, w_{Bf} is weight fraction of B in the feed stream, and w_{Cf} is concentration of catalyst in the feed stream. At steady state, the accumulation terms in the left hand sides of Eqs. (6) through (8) will be zero.

3.3. Steady State Multiplicity

At steady state, the system described by Eqs. (2), (3), and (6) through (8) leads to Eq. (9), which allows calculation of the concentration of catalyst needed in reaching a chosen degree of conversion of B, defined for the CSTR as $X_B = 1 - w_B / w_{Bf}$.

$$w_{Cf} = w_{Bf} X_B \frac{1 + \tau k_T \dfrac{4 w_{Bf}^3 (1 - X_B)^3}{(1 + 4 w_{Bf} (1 - X_B))}}{\tau \left(k_P + k_S \tau k_T \dfrac{4 w_{Bf}^3 (1 - X_B)^3}{(1 + 4 w_{Bf} (1 - X_B))} \right) \dfrac{4 w_{Bf}^2 (1 - X_B)^2}{(1 + 4 w_{Bf} (1 - X_B))}} + c(1 - w_{Bf}) \qquad (9)$$

The second term on the right hand side is used to correct for the amount of catalyst that is killed by impurities. These are suspected to come in mostly through reactant A. Fig. 3 shows curves created using Eq. (9) at two temperatures. Reactor residence time is 5 hours and the weight fraction of B in the feed is 0.8. The curve at the lower temperature illustrates the behavior one would typically expect, with a monotonic increase in the degree of conversion of B as the concentration of catalyst is increased.

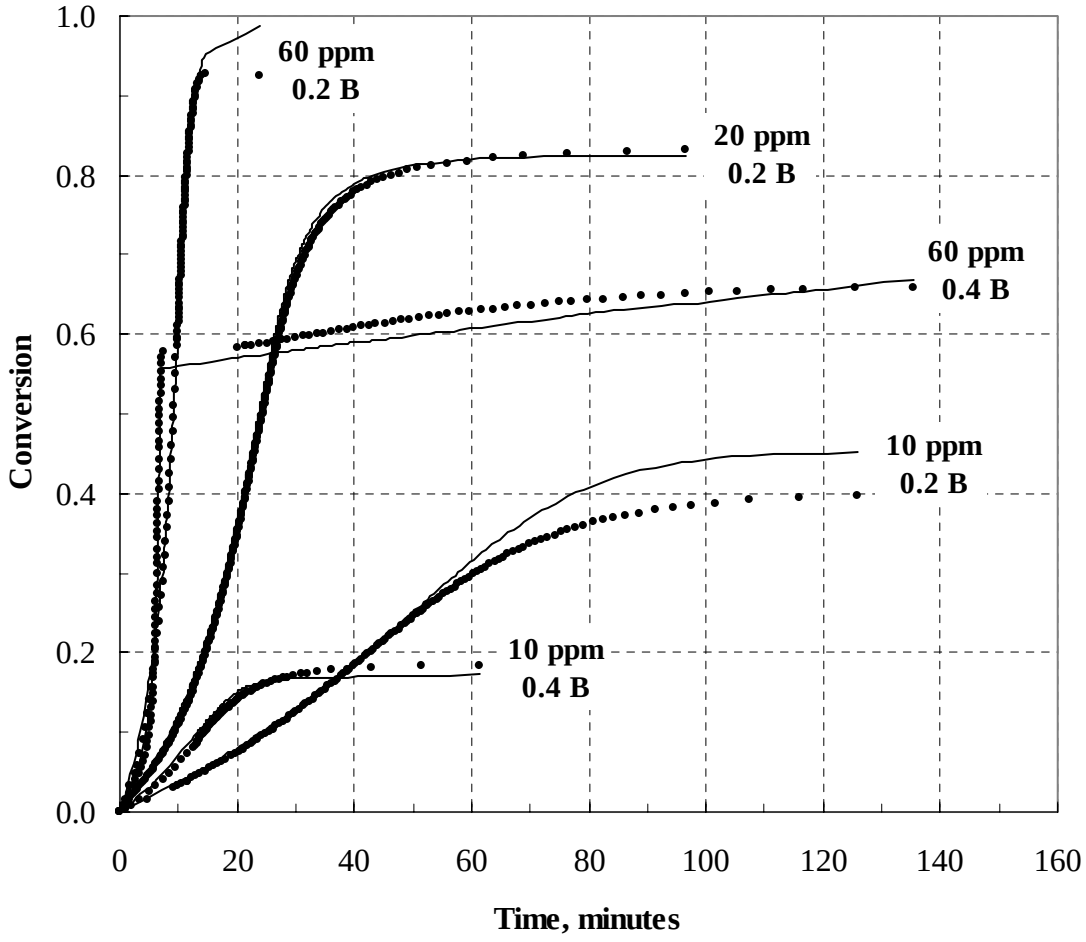


Figure 2. Experimental results and model predictions for conversion of B in ARC runs

The curve at higher temperature reveals that as many as three steady states are possible for a given concentration of catalyst. State multiplicity is thus more likely to appear at higher temperatures, when site transformation has a more noticeable effect on the main reaction.

Steady state multiplicity has been studied in great detail for non-isothermal reactors [3]. Isothermal reactors also display this kind of behavior [4]. Reaction mechanisms similar to ours have been reported in biochemical systems [5].

3.4. Model Validation

Eq. (9) was used in choosing the set of operating conditions for pilot plant experiments. Pilot plant experiments were allowed to reach steady state by running them for at least 25 hours, or five times the chosen residence time of 5 hours. The feed weight fraction of B was approximately 0.8 in all runs. These conditions were expected to lead to significant monomer conversion in the pilot plant, since conversions lower than 0.75 were considered unsafe. Slight differences in the type and quality of the reactants used in pilot plant experiments, when compared to the reactants used in ARC experiments, led however to lower catalyst activity than expected.

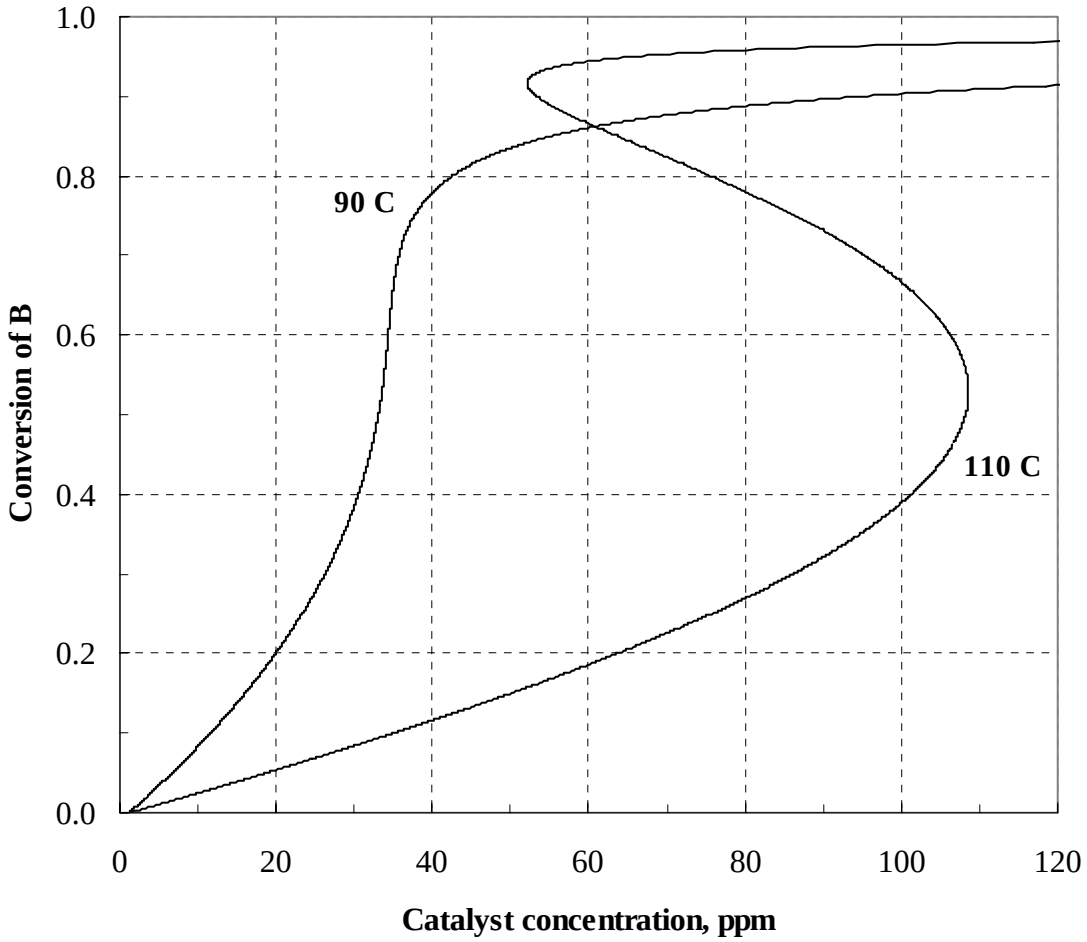


Figure 3. Effect of catalyst concentration on conversion of B at low and high temperatures

As implied by the first few pilot plant experiments, the estimates supplied by ARC experiments for the parameters of Eq. (9) could not accurately describe the behavior of the pilot plant. Because of these differences, the original experimental plan was not followed strictly and modifications were needed. The first modification involved an overall increase of 20 ppm in the concentration of catalyst. Additional complications led to further modifications in the experimental plan. Fig. 4 compares the runs of the original plan to the actual runs executed. Missing from the later set are some of the runs at low catalyst concentration and high reactor temperature. Various attempts at executing these runs were unsuccessful since the conversion of reactant B would quickly move away from conversion levels considered safe and the pilot plant had to be shut down. This is consistent with the trend displayed in Fig. 3 as the temperature goes up.

Despite all the complications, Eq. (9) was a good starting point since it provided reasonable estimates of the conditions to expect at steady state for all runs. When things did not go as expected, Eq. (9) still helped in locating problem causes and in choosing corrective actions. Additional help was provided by the full dynamic model described by Eqs. (2), (3), and (6) through (8).

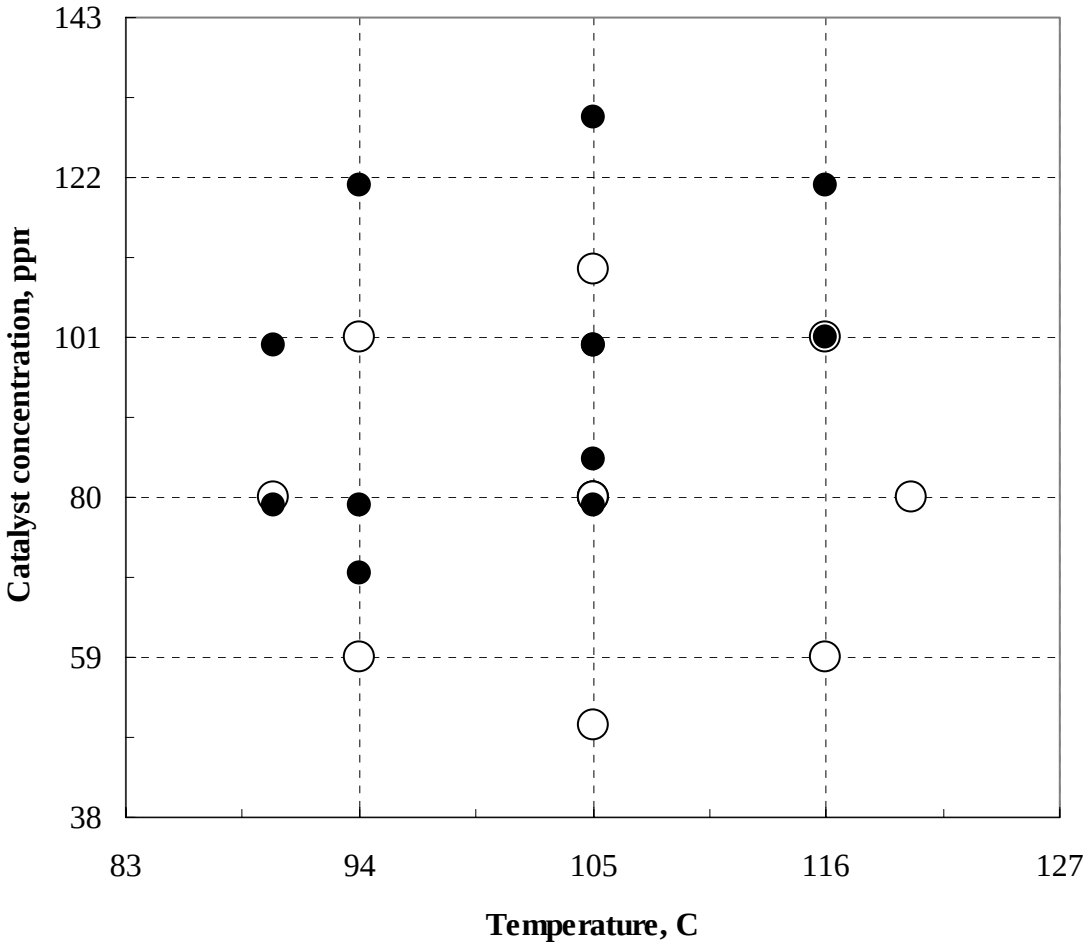


Figure 4. Planned (O) and actual (●) runs of experimental design for the pilot plant

3.5. Mechanistic vs. Empirical Models

Fig. 5 compares predicted and observed values of conversion of B during pilot plant experiments. One set of predictions was generated using Eq. (9) and the parameter estimates obtained by fitting the ARC data. This model has a tendency to underestimate the activity of the catalyst when it is present in high concentrations and to overestimate it at low concentrations. The other set was generated by a linear empirical model whose parameters were fitted to the pilot plant results.

The fit provided by the linear empirical model is excellent. Even though an improved fit is obtained when the parameters in Eq. (9) are adjusted to match the pilot plant results, the predictive ability of this mechanistic model can only get as good as that of the linear model, but not better. Add to this the fact that there are product properties for which only the empirical approach is possible and the advantages of this approach become clear. Unfortunately, in its current form the empirical model knows nothing about time. Eq. (9), on the other hand, includes the residence time as a variable. The dynamics of the system are modeled by Eqs. (2), (3), and (6) through (8), with the corresponding steady state solution provided by Eq. (9). One may argue, however, that data from dynamic pilot plant experiments could be used to include time as an independent variable in the empirical model.

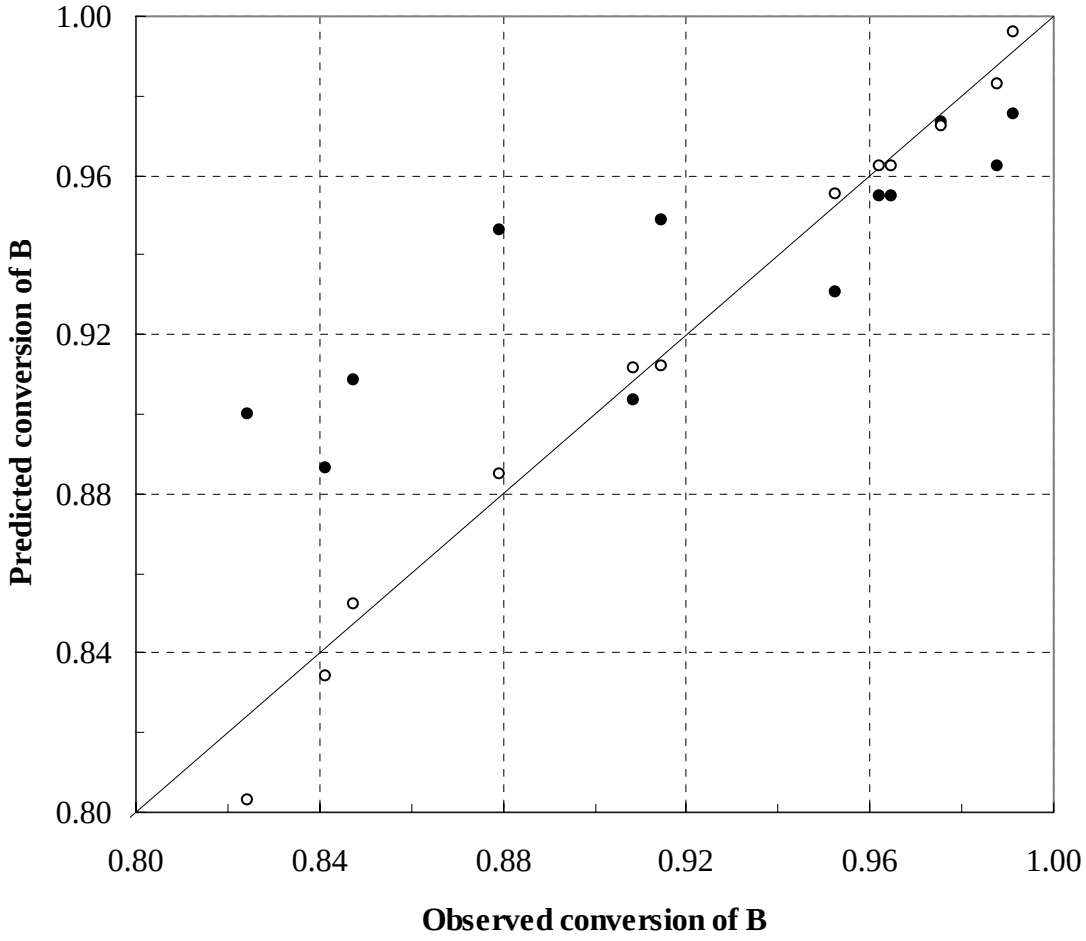


Figure 5. Fits generated by empirical (o) and mechanistic (●) models of conversion of B

The most troublesome aspect of the empirical model is the way it handles conditions of low catalyst activity, which correspond to high temperature and low catalyst concentration. At 116 ºC and 79 ppm, for example, a conversion of B of 0.89 is predicted by the empirical model. In the pilot plant, however, the conversion had already dropped to 0.75 and would have continued dropping if a plant shut down had not been ordered. The chosen conditions thus fall in a region that the mechanistic model can handle better than the empirical model, as illustrated by Fig. 6.

An improved set of parameters for Eq. (9) was used to create the corresponding curve in Fig. 6. More impurities are expected in pilot plant experiments than in ARC experiments and for this reason a higher catalyst level at zero conversion is observed in Fig. 6 than in Fig. 3. This level, like the prediction of 0.5 for the conversion of B at 79 ppm are not supported by experimental data since conversions lower than 0.75 were not allowed in the pilot plant. Thus, while the mechanistic model is in principle capable of generating predictions over the entire conversion range, it can only be built by using high conversion data. ARC and dynamic experiments should provide, however, the information needed in validating a given kinetic mechanism and set of kinetic parameters for the mechanistic model.

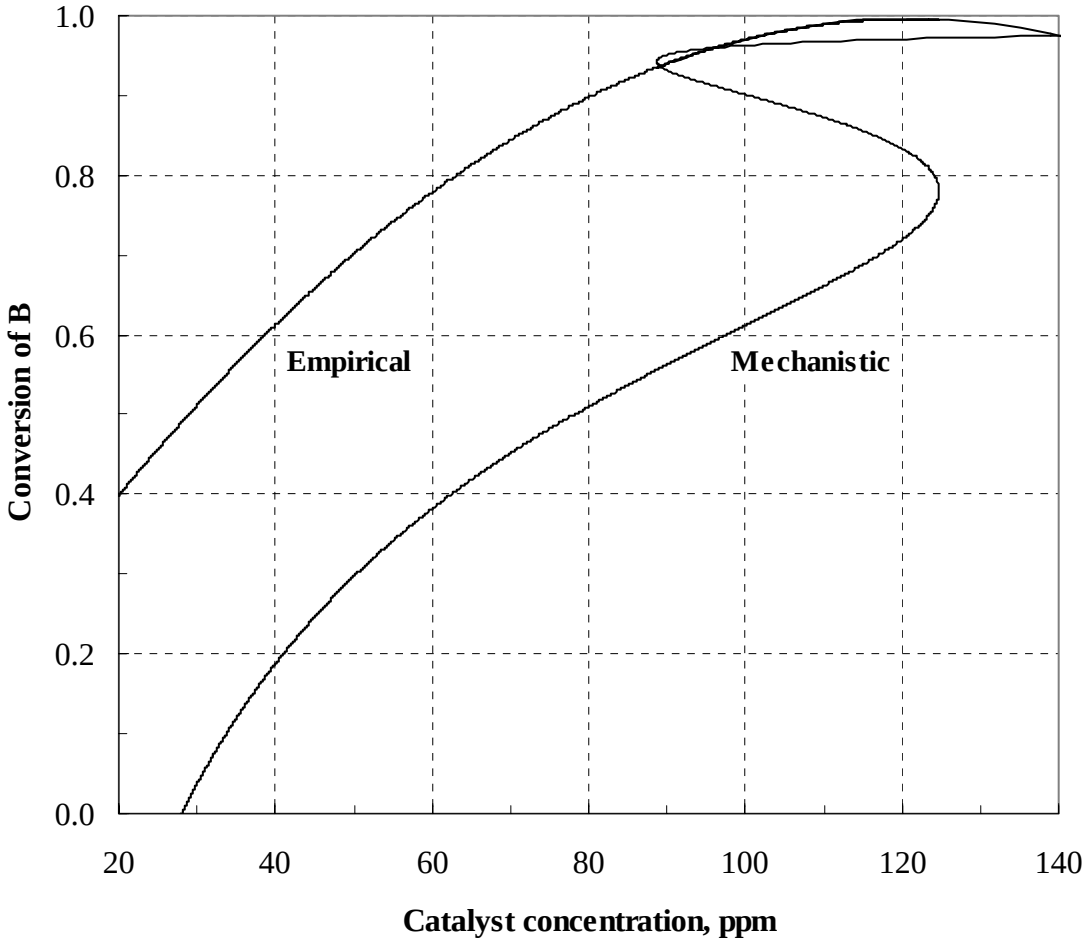


Figure 6. Conversion of B predicted by empirical and mechanistic models at 116 ºC

4. DISCUSSION

Our validated reactor models are going to be used as part of a process intensification strategy that will search for optimum operating conditions of commercial plants and will choose product grade transition strategies for those plants. At the commercial scale, the energy balance plays a more important role and the combination of this balance with the reaction mechanism for the catalyst will introduce complications that process control systems will have to deal with. Those systems should be designed to avoid reaction extinction episodes or the cycling behavior that results from process configurations and process conditions that favor sustained oscillations.

We continue to develop empirical and mechanistic models for reactant conversion and for product properties of interest in this system. We view the approaches as complementary, rather than contrary. Empirical models can provide excellent fits over the experimental region and they are often the only available option. Mechanistic models, when properly validated and when extrapolated with care, can help in the analysis of possible scenarios at the commercial scale even before a commercial plant becomes available.

One of the key aspects of this work has been the early detection of state multiplicity during ARC experiments. Had it not been for such detection, execution of the pilot plant experimental plan would have been plagued by far too frequent reaction extinction episodes. Thanks in part to this knowledge we have been able to push our experiments to the lowest catalyst concentration and thus more economic operating region without considerable waste of resources.

We are dealing with a strongly exothermic reaction and a highly volatile reactant. We are thus constrained to low concentrations of unreacted B in pilot plant reactors and in larger reactors. This represents a major limitation in model validation since the low conversion region is rich in information but unreachable. Despite such limitation, we already have steady state and dynamic reactor models that are very helpful in evaluating multiple catalysts and products, and should be very helpful in facilitating the transition to commercial scale reactors.

REFERENCES

1. D. I. Townsend and J. C. Tou. Thermal hazard evaluation by an accelerating rate calorimeter. *Thermochimica Acta*, 37: 1 – 30, 1980.
2. R. B. Bird, W. E. Stewart, and E. N. Lightfoot. Transport Phenomena. John Wiley & Sons, New York, 1960.
3. A. Uppal, W. H. Ray, and A. B. Poore. On the dynamic behavior of continuous stirred tank reactors. *Chem. Eng. Sci.*, 29: 967 - 985, 1974.
4. L. F. Razón and R. A. Schmitz. Multiplicities and instabilities in chemically reacting systems — a review. *Chem. Eng. Sci.*, 42: 1005 - 1047, 1987.
5. P. A. Ramachandran, B. D. Kulkarni, and A. Sadana. Analysis of multiple steady states of complex biochemical reactions. *J. Chem. Tech. Biotechnol.*, 31: 546 – 552, 1981.

Response surfaces and sensitivity analyses for an environmental model of dose calculations

Bertrand Iooss[a], *Nicolas Devictor*[a] *and François Van Dorpe*[b]
[a] *CEA Cadarache, DEN/DER/SESI/LCFR, 13108 Saint Paul lez Durance, France;*
[b] *CEA Cadarache, DEN/DTN/SMTM/LMTE, 13108 Saint Paul lez Durance, France*
E-mail: bertrand.iooss@cea.fr, nicolas.devictor@cea.fr, francois.vandorpe@cea.fr

ABSTRACT

A parametric sensitivity analysis is carried out on a radiological impact software describing the radionuclides transfer to the man following a chronic gas release of a nuclear installation. An effective dose received by age group can thus be calculated according to the duration of the release. Due to the large number of input parameters (more than fifty for each output variable) a methodology is proposed. The generation of one thousand Monte-Carlo simulations allows to calculate correlation coefficients between input parameters and output variables, which give a first overview of important factors. Least-squares multiple regression is used to construct response surfaces in polynomial form for each output variable. Then using these polynomials, we calculate the global sensitivity indices of Sobol by the Monte-Carlo method. We show the application of this method to one site of study and to one reference group, for two radionuclides: iodine 129 and uranium 238.

Keywords: Uncertainty, sensitivity analysis, environmental transfer, radiological impact

1. INTRODUCTION

The present study presents the global sensitivity analysis of a radiological impact software called GASCON. GASCON (developed by CEA/DAM/DASE) is dedicated to chronological atmospheric releases and dosimetric impact which is used for CEA facilities safety assesment. This software evaluates the doses received by a population (called reference group) exposed to the cloud of radionuclides and *via* the food chains. It takes into account the interactions which exist between the man, the plant and the animal, the different ways of transfer (wind, rain, ...), the distance between emission and observation, the time passed between emission and calculation, ...

Various stages in the analysis of a process (software, measurement, experiment, ...) introduce potential errors, in particular in the construction of the various models: real phenomenon with the physical model, physical model with the mathematical model, and mathematical model with the numerical model. The principal sources of uncertainties

Further author information: (Send correspondence to B.I.)
B.I.: Telephone: +33 (0)4 42 25 72 73
N.D.: Telephone: +33 (0)4 42 25 30 05
F.V.: Telephone: +33 (0)4 42 25 79 18

are in the approximation made by the modeling of the physical phenomenon, the approximations made on the parameterization of the model, the input data and the input parameters. The uncertainty analysis is used to evaluate the confidence interval or the probability distribution of the result. The global sensitivity analysis is used to quantify the influence of the input parameters uncertainties on the output variables uncertainties. Recent studies have applied different uncertainty analysis and sensitivity analysis methods to environmental models (Helton [6], Campolongo and Saltelli [2], Hedin [5]).

The results provided by GASCON are in the form of annual effective doses (Sv/yr) received by a reference group, divided into three age compartments: adult, child and baby. We also distinguish three operating cycles of the release installation: one year, ten years, fifty years. In our study the reference group is a village near the release installation, and we consider two radionuclides ^{129}I and ^{238}U. There is thus eighteen output variables.

The main ways of exposure taken into account in GASCON are:

- ◇ external exposures: radioactive cloud and soil deposits;
- ◇ internal exposures: inhalation, ingestion of plants contaminated by direct way (foliar transfer by contact with the radioactive cloud) and indirect way (soil deposit then root transfer), ingestion of contaminated animal productions.

Some input data are specific of the studied radionuclide or of the studied site (meteorological conditions, soils nature, feed rations, ...). We have deduced from the literature the variation ranges of parameters considered for the sensitivity analysis, which are:

- ◇ dose factors for external irradiation, effective ingestion, effective inhalation;
- ◇ transfer factors to animal productions (milk, meat of cow, ewe, goat, pig ...);
- ◇ factors of soil-plant transfer (vegetables, cereals, fodder, ...);
- ◇ translocation factors (fruits, vegetables, cereals, ...);
- ◇ sorption coefficients K_d (sands, silts, clays and organic matter);
- ◇ dry deposit rate for each radionuclide;
- ◇ local feed rations of the reference group for the various age compartments (vegetables, fruits, cereals, milk, meat, egg, ...), and animal feed rations (grass, hay, corn) related to the products eaten by the reference group.

The following section presents the four stages of our methodology: uncertainty analysis *via* Monte-Carlo calculations, fast sensitivity analysis with correlation coefficients between input and output variables, construction of response surfaces requiring negligible computation times, calculations of Sobol sensitivity indices. In the third section this methodology is applied to the GASCON software using specific nuclear installation and reference group. We discuss the result of this approach and conclude in the last section.

2. METHODOLOGY

2.1. Uncertainty analysis

The general objective of our uncertainty analysis is to evaluate uncertainty on a computation result Y taking into account uncertainties on the input parameters X_i $(i = 1, \ldots, N_p)$.

It is necessary for each input parameter to evaluate a probability density function (by expert opinion or by a data statistical analysis). The results of the uncertainty analysis is conditioned with the choices of these probability densities. To propagate uncertainties, we use the pure Monte-Carlo strategy: random generation of N samples of input parameters, then software calculation for each sample.

Because of our lack of knowledge, we choose the uniform law for the distribution of all input parameters. Moreover, it requires only the bounds of the parameter variation ranges. However for the majority of the GASCON parameters, an order of magnitude separates the minimal and nominal values (min $\sim$ nominal/10) and the nominal and maximal values (max $\sim$ nominal $\times$ 10). Thus, if we choose the uniform function on [min; max], the majority of the simulated values will be included in the interval [nominal; max]. To counter this problem, each simulation proceeds in the following way:

- $\star$ we simulate a uniform random variable u on $[0; 1]$;
- $\star$ if $u \leq 0.5$: the simulation value is $2u(\text{nominal} - \text{min}) + \text{min} \in [\text{min}; \text{nominal}]$;
- $\star$ if $u > 0.5$: the value is $(2u - 1)(\text{max} - \text{nominal}) + \text{nominal} \in]\text{nominal}; \text{max}]$.

From the Monte-Carlo simulations, we obtain for each output variable the elementary statistical parameters (average, minimum, maximum , standard deviation, variation coefficient, skewness and Kurtosis coefficients) and the probability distribution. From the distributions, we can observe the spreading out of the output variables, the confidence intervals, the multiplicity of modes, ... Statistical comparison tests can also be made. In our study, we analyze nine output variables for each radionuclide, and we deduce by the Kolmogorov-Smirnov test (Saporta [11]) which variables are statistically similar.

2.2. Sensitivity analysis with correlation coefficients

The global sensitivity analysis is used to quantify the contribution of each input parameter to the response variability. The linear correlation coefficient ρ (or Pearson coefficient) between two random variables X and Y is the simplest sensitivity index. If ρ is close to $+1$ or -1, the assumption of linearity between X and Y is valid. If Y is an output variable and if there are several input variables X_j $(j = 1, \ldots, N_p)$, the correlation coefficients between Y and each X_j have not quantitative sense, but can reveal the linear character of some dominant variables (Saporta [11]).

If the behavior of Y compared to each parameter is overall linear, it is possible to obtain quantitative measurements of their influence from the standard regression coefficients (linear regression connecting Y to the X_j). To obtain a measurement of the linearity of the relation between Y and X_j, we use the partial coefficients of correlation p_j. In opposition to the standard regression coefficients, the partial correlation coefficients allow to eliminate the influence from the other variables (Saltelli et al. [9]). If the relation between X and Y is not linear, the correlation coefficients of the ranks (or Spearman coefficients) can be used. By replacing the values of parameters $X^1, \ldots, X^N$ and of output variable $Y^1, \ldots, Y^N$ by their rank, the assumption of linearity is thus replaced by the assumption of a monotonous relation.

The regression or correlation coefficients are related to linear or monotonous assumptions. Moreover, they study only the relations between the output variable and an input

parameter independently of the other parameters. However, many problems are neither linear nor monotonous, and reveal physical dependences between parameters. Global sensitivity analyses more adapted to these conditions are available, but they are definitely more expensive in computing times (Saltelli et al. [9]). In order to test them on GASCON, it is necessary to simplify this software and to replace it by a response surface.

2.3. The response surface method

The response surface method (Box and Draper [1], Kleijnen [8]) is used to build a function which simulates the behavior of a physical or chemical phenomenon in the field of variation of the influential parameters, starting from a certain number of experiments. In our study, an experiment is a calculation by the GASCON software. Building a response surface (RS) aims to obtain a mathematical model representative of the studied software, having good capacities of prediction, and whose computing time to evaluate an output variable is negligible. Such a RS will be thus effective for the uncertainty and sensitivity analyses, requiring several thousands of simulations.

To build a RS, it is necessary to have the software H which models the studied phenomenon, a sample D of N points $(\mathrm{x}(i), z(i))$, where $\mathrm{x}(i)$ is a vector of the N_p random input parameters and $z(i) = H[\mathrm{x}(i)]$ $(i = 1 \dots N)$ is the software response, and a family F of functions $f(\mathrm{x}, \mathrm{c})$, where c is a vector of parameters (parametric regression) or indices (nonparametric regression) which makes possible the identification of the various elements of F. There are multiple RS families (Hastie et al. [4]): polynomials, interpolating radial functions, splines, generalized linear models, partial least squares, neural networks, support vector machines, ... In this work, we use only polynomials, because they have a simple physical interpretation and the majority of the GASCON equations are linear formulas compared to each variable.

In general, we use the technique of least squares to obtain the best representing f_0 in the family F. We minimize the function $\sum_{i=1}^{N}\{z(i) - f[\mathrm{x}(i), \mathrm{c}]\}^2$ in relation to the parameters c, to obtain c_0 and the RS $f_0(\mathrm{x}) = f(\mathrm{x}, \mathrm{c}_0)$. The RS quality of approximation is given from a statistical analysis on a construction basis, whereas the quality of prediction is deduced from a prediction basis. A simple method to qualify a RS is to compare on the two bases some indicators obtained from the RS with those obtained directly with software H. In our study, we initially compare their average, standard deviation, minimum and maximum. In addition, a regression analysis allows to determine the share of variability of the output variable explained by the model. Other possibilities to validate RS are the cross-validation or bootstrap techniques (Hastie et al. [4]). For simplicity, we just present two statistics which give global measurements of correlation between two data sets A and B: the Pearson correlation coefficient ρ and the coefficient of determination R^2 which writes

$$R^2(A, B) = 1 - \frac{\sum_{i=1}^{N}(A_i - B_i)^2}{\sum_{i=1}^{N}(\overline{A} - A_i)^2} \ , \tag{1}$$

where $\overline{A}$ is the average of A. The coefficient R^2 represents the fraction of the variation compared to the average explained by the smoothed model, i.e. the percentage of output variables explained by the response surface. In our case, A is the software response $A_i = z(i) = H[\mathrm{x}(i)]$ and B is the model (RS) response $B_i = f[\mathrm{x}(i), \mathrm{c}_0]$.

These criteria are global and it is possible that the data adjusted are not homogeneous. It is the case when the studied variable covers a broad range of variations with multiple orders of magnitude. In this case, the contributions of the low values to the R^2 measurement are negligible. To cure this problem, the study of residual statistics gives some indications of the regression accuracy. The residuals ϵ have to be centered with a standard deviation σ_ϵ small compared to the GASCON standard deviation σ_A. We also examine the average and the standard deviation of the relative residuals $\epsilon_i/B_i = (A_i - B_i)/B_i$.

2.4. Global sensitivity analysis

We consider methods of variance analysis which aim at determining the weight of the variance of the response $Y = f(X)$ resulting from a variable or a group of variables (McKay [7], Saltelli et al. [9]). Their objective is to calculate the global sensitivity index St_i defined as the sum of all the sensitivity indices implying parameter X_i. The Sobol method allows a relatively simple evaluation of the terms S_i, S_{ij}, ... (Saltelli et al. [9]).

In practice, we can evaluate Sobol indices by a Monte-Carlo method, which require a very significant number of simulations, typically $N_s = 10000$ to estimate an index of a parameter (S_i, S_{ij}, ... or St_i). This justifies the use of response surfaces to minimize the computing times. Although the calculation of Sobol index with FAST method is definitely less expensive, we use the Monte-Carlo method because we obtain a realistic confidence interval on the Sobol index by repeating the index calculation. This information is essential if we want to rigorously classify the influence of the various input variables. In our study, we carry out $N_{ic} = 200$ calculations of each Sobol index. Moreover, the Monte-Carlo method calculates all the Sobol indices (S_i, S_{ij}, ... or St_i), which bring information on the interactions between the input parameters. In our study, we just calculate S_i and St_i to measure the influence that the variable X_i has while acting alone. The Monte-Carlo method calculates these two indices using the same N_s simulations (Saltelli [10]), whereas FAST calculates them starting from two different sets of simulations.

For the model $Y = f(X)$ where X is a vector of N_p parameters, we need $N_{ic} \times N_s \times (N_p + 2)$ evaluations of f to calculate the first order indices S_i and total indices St_i for all the parameters X_i, and to allocate a confidence interval to them. In our study, $N_s = 10000$ and $N_{ic} = 200$. The value of N_p depends on the model of response surface which is adjusted. For GASCON, we try to take into account no more than ten parameters in each response surface. For $N_p = 10$, there will be 2.4×10^7 calculations of f.

3. RESULTS

The GASCON software is applied to a French nuclear research center. The gas release is fixed at a symbolic value of 1 Bq/year which does not represent a realistic release. This also induces non realistic effective dose rates. We perform one thousand independent Monte-Carlo simulations of the GASCON software (30 seconds per simulation). In Figure 1, the distributions of the 18 output variables are represented. For a given radionuclide, the output variable distributions seem very similar. We carry out statistical tests between the coherent output variables (same radionuclide, same age compartment or same operating cycle) by the Kolmogorov-Smirnov test which evaluates if there are or not statistically

significant differences between the two distributions on the degree of confidence 95%. For ^{129}I, there is no difference at 95% between all the distribution couples, except between (adult, operating cycle of 50 years)/(baby, operating cycle of 1 year) and (adult, operating cycle of 50 years)/(baby, operating cycle of 10 years). For ^{238}U, all distribution couples have differences at 95%, except for (adult, 1)/(adult, 10), (child, 1)/(child, 10), (baby, 1)/(baby, 10). In summary, the variables at ten years have approximately the same distributions than the variables at one year, and for ^{129}I the variables "child" have the same distributions than the variables "adult". Thus for ^{129}I and ^{238}U, it is sufficient to study four output variables: (adult, 1), (adult, 50), (baby, 1), (adult, 50).

Table 1 shows the largest Pearson and Spearman correlation coefficients between input and output variables. All the output variables reveal five important input parameters: the dose factor of effective ingestion *ingeff*, the human feed ration of goat's milk *ra_gmilk*, the goat feed ration of grass *ra_grass*, the dry deposit rate *dep*, and the transfer factor to the goat's milk *gmilk*. For ^{238}U, some additional parameters appear: the goat feed ration of cereals *ra_cer* and the transfer factor of the vegetables fruits *vegfr*.

ρ ρ^S	*ingeff*	*gmilk*	*ra_gmilk*	*ra_grass*	*dep*	*ra_cer*	*vegfr*
Ad_I_1	33 55	09 16	31 46	28 41	21 38		
Ad_I_50	33 56	09 16	31 45	28 41	21 39		
Bb_I_1	32 54	09 16	31 47	28 42	21 37		
Bb_I_50	32 54	09 16	31 47	28 42	21 38		
Ad_U_1	28 55	20 24	21 29	18 26	15 20	08 11	
Ad_U_50	36 62	17 19	18 21	15 20	20 29		10 19
Bb_U_1	25 53	21 33	21 37	18 33	15 26	08 10	
Bb_U_50	27 59	20 29	21 32	18 28	16 21	08 08	

Table 1. Correlation coefficients (in %) of Pearson ρ and Spearman ρ^S between the output and input variables (selected if ρ and ρ^S are higher than 8%). The notation Ad_I_1 means (adult, ^{129}I, one year of release).

At present, for each output variable, we want adjust a response surface by a polynomial model obtained by multiple regression. By selecting and combining the parameters found in Table 1, the best results are obtained when the regressions are made according to certain food chains, which are linear combinations of the various terms contributing in each chain. For example, the food chain of the goat's milk is

$\alpha_1 * ingeff * gmilk * ra\_gmilk * ra\_grass * dep + \alpha_2 * ingeff * gmilk * ra\_gmilk * ra\_grass * dep^2 + \alpha_3 * ingeff * gmilk * ra\_gmilk * ra\_hay * dep + \alpha_4 * ingeff * gmilk * ra\_gmilk * ra\_hay * dep^2$, where α_i $(i = 1, \ldots, 4)$ are regression coefficients.

For all the output variables in ^{129}I and for the variables (^{238}U, baby, 1) and (^{238}U, baby, 50), the response surfaces are polynomials based on the food chains of the goat's milk and the ewe's milk (the same than the goat's milk by replacing the goat by the ewe). For the variables (^{238}U, adult, 1) and (^{238}U, adult, 50), response surfaces include also the effective inhalation term $\alpha_1 * inheff$. For (^{238}U, adult, 50), we add the food chain of the pig's meat and the food chain of the vegetable fruits by indirect transfer:

$\alpha_1 * ingeff * vegfr * ra\_vegfr * dep + \alpha_2 * ingeff * vegfr * ra\_vegfr * dep^2$.

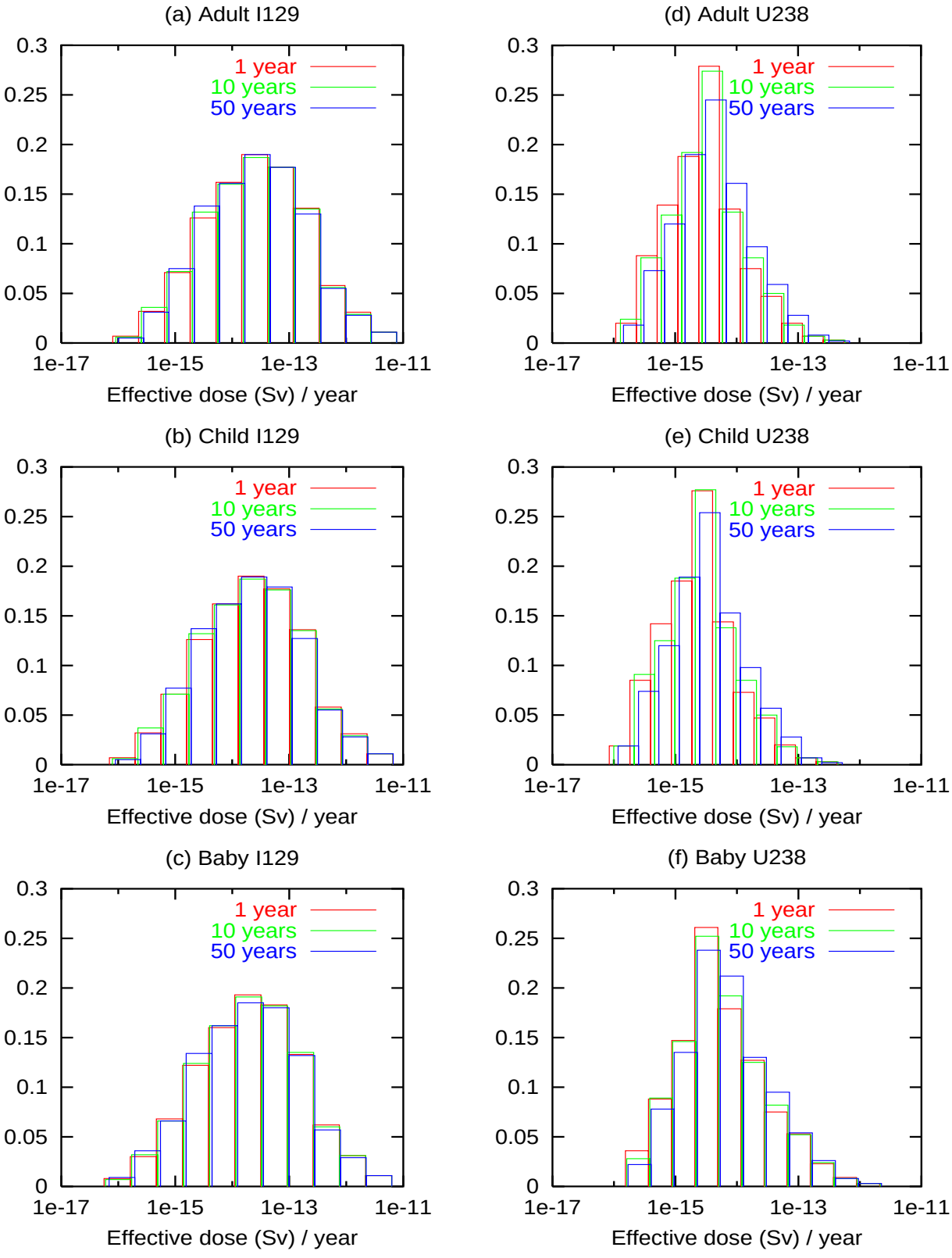


Figure 1. Output variable distributions (Sv/year).

The statistical validity of the RS is studied on a basis of construction (of size 2/3 of the data base) and on a basis of prediction (of size 1/3 of the complete data base). For ^{129}I, the statistics of the R^2 and ρ are excellent: they are equal to 99% for all the variables and on the two bases. The responses in ^{129}I of the GASCON software are almost entirely explained by the terms of the food chain of the goat's milk. For ^{238}U, the statistics R^2 and ρ are also satisfactory (all higher than 92%). It is necessary to integrate in the model the food chains of the goat's milk, the ewe's milk, the effective inhalation, and the pig's meat. In addition, at fifty years, the food chains of vegetables by indirect transfer seem essential. It is noticed that the dry deposit rate *dep* operates linearly and quadratically in the food chains. This is due to the fact that the relations in GASCON utilize *dep* like a power of another factor.

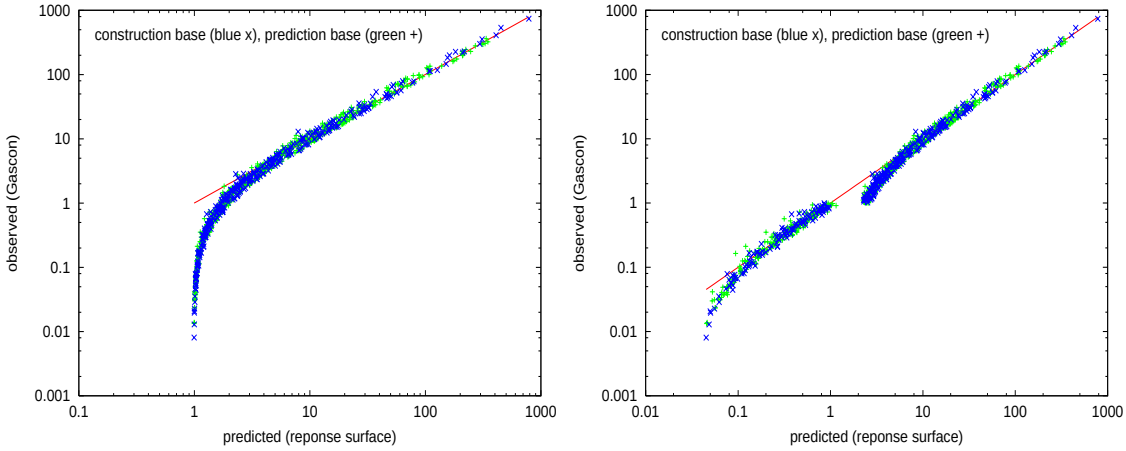


Figure 2. For the adult, ^{129}I and one year of release, comparisons between response surface and GASCON calculations for the two bases (construction and prediction). Unities have to be multiplied by 10^{-14} Sv/year.

It thus seems that the RS are valid but the relative statistics balance our judgement: the relative standard deviations σ_ϵ/σ_A are worth roughly 10% for ^{129}I and 20% for ^{238}U. This lets suppose that the adjustments are not good everywhere. The calculation of the averages and standard deviations of the relative residuals confirms this judgement. For ^{129}I, the representativeness of GASCON by the RS is on average of -30%. For ^{238}U, the representativeness of GASCON by the RS is on average of -15%. The Figure 2 (left) makes it possible to locate the problem thanks to a comparison on the Ad_I_1 variable between GASCON and RS in logarithmic scale. It is noted that the high values ($> 10^{-14}$ Sv/yr) are well adjusted, whereas the low values are completely over-estimated by the RS. To cure this, we propose to separate the construction and prediction bases in two parts each one (one with values higher than 10^{-14} Sv/yr and the other with values lower than 10^{-14} Sv/yr). We choose the same factors of regression for the calculation of the RS. The adjustments are presented in Figure 2 (right). The two RS correctly explain the data in each field of variation with the same factors of regression. This is confirmed by the averages and standard deviation of the relative residuals on the basis of construction which are equal to -13.7% and 22.2% for the raised values, and -4.6% and 22.2% for the

low values, instead of -28.6% and 36.2% on all the variation domain. This problem of regression on a field of several orders of magnitude can be solved in a more satisfactory maneer by using the technique of weighted least squares. In the minimization of the functional, we allocate larger weights to low values. This makes it possible to have a homogeneous response surface on all the field of variation.

Now, the GASCON software can be replaced by the RS (polynomial model) which can be used to calculate Sobol indices by extensive Monte-Carlo computations. Figure 3 gives for each output variable the Sobol indices of the most influential variables, with their error bars. Having repeated 200 times the Sobol calculations, the average values are good estimates of the true Sobol indices. We conclude that for ^{129}I, the most influential parameters are the dose factor of effective ingestion and the feed ration of the goat's milk. For ^{238}U, the most influential parameters are the dose factor of effective ingestion, the transfer factor of the goat's milk and the feed ration of the goat's milk.

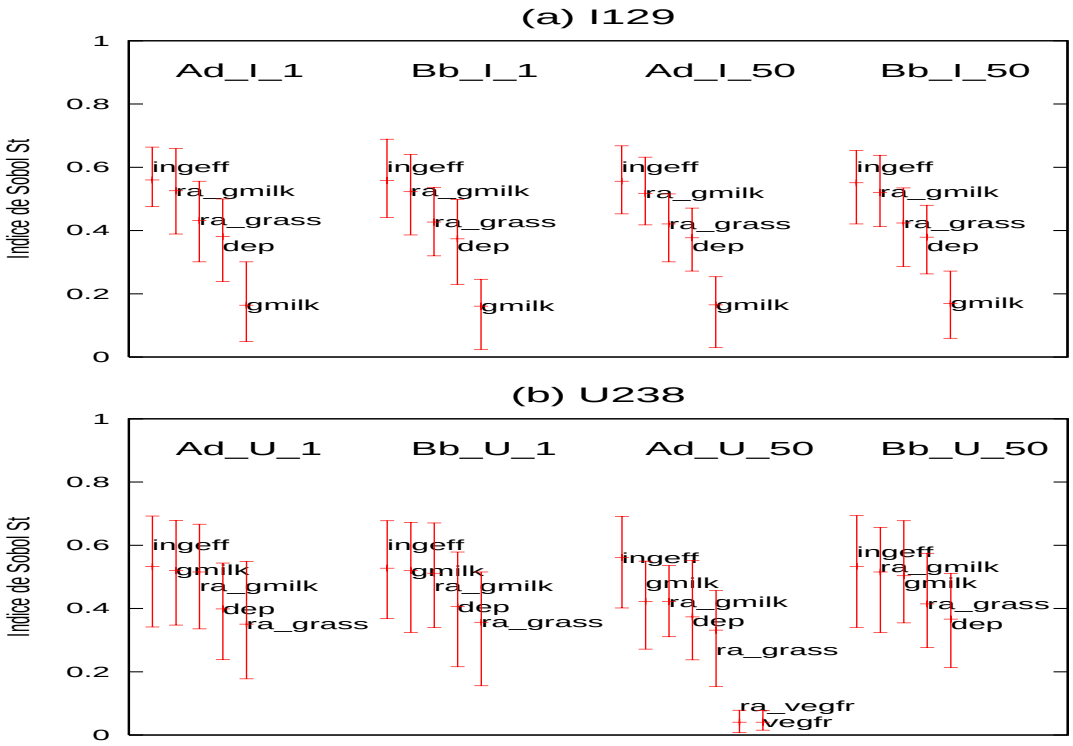


Figure 3. Sobol indices for each parameter. Uncertainty bar represents the minimum, average and maximum indices obtained with 200 Sobol calculations.

By carrying out calculations of the Sobol indices on the RS obtained by distinguishing high values ($> 10^{-14}$ Sv/yr) and low values ($< 10^{-14}$ Sv/yr), we find exactly the same results in the field of high values. In the field of low values, the classification is similar except that the dry deposit rate is placed in first position at equality with effective ingestion. It is thus considered that our results are valid in the field of the high values and are approximately correct in the field of the low values.

4. CONCLUSION

The four steps of our methodology (Monte-Carlo simulation, correlation coefficients analysis, response surfaces, Sobol indices) have allowed to quantify the influence of input parameters on the GASCON software response (annual effective dose received by the man), for a specific nuclear installation, a specific population, and for two radionuclides (^{129}I and ^{238}U). During the correlation coefficient analysis, the calculations of standard regression coefficients would allow to have more information on important input parameters. During the response surface construction, other statistical validation methods like cross-validation or bootstrap technique could also be useful (Hastie et al. [4]). In this work, the response surfaces found are in simple polynomial form explicit for the physicist understanding. For the approximation of software simulating more complex phenomena, more elaborated and not explicit response surfaces, like neural networks (Devictor and Martinez [3]), can be used.

ACKNOWLEDGMENTS

This work was supported by the MRIMP project of the "Risk Control Domain" which depends on CEA/Nuclear Energy Division/Nuclear Development and Innovation Division.

REFERENCES

1. G. Box and N. Draper. *Empirical model building and response surfaces,* J. Wiley, New York, 1987.
2. F. Campolongo and A. Saltelli. Sensitivity analysis of an environmental model: an application of different analysis methods, *Reliability Engineering and System Safety,* 57:49-69, 1997.
3. N. Devictor and J.-M. Martinez. Non linear regression methods in uncertainty and sensitivity studies and reliability computations, *In Proceedings of ESREL '2000,* Edinburgh, UK, May 2000.
4. T. Hastie, R. Tibshirani and J. Friedman. *The elements of statistical learning,* Springer, 2002.
5. A. Hedin. Probabilistic dose calculations and sensitivity analyses using analytica models, *Reliability Engineering and System Safety,* 79:195-204, 2003.
6. J. C. Helton. Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal, *Reliability Engineering and System Safety,* 42:327-367, 1993.
7. M. D. McKay. Non parametric variance-based methods of assessing uncertainty importance, *Reliability Engineering and System Safety,* 57:267-279, 1997.
8. J. Kleijnen. Sensitivity analysis and related analyses: a review of some statistical techniques, *Journal of statistical Computer Simulation,* 57:111-142, 1997.
9. A. Saltelli, K. Chan, and E. M. Scott (Eds). *Sensitivity analysis,* J. Wiley, Wiley Series in Probability and Statistics, 2000.
10. A. Saltelli. Making best use of model evaluations to compute sensitivity indices, *Computer Physics Communication,* 145,:280-297, 2002.
11. G. Saporta. *Probabilités, analyse des données et statistique,* éditions Technip, 1990.

Simulation Error Models for Improved Reservoir Prediction

A. O'Sullivan, M. Christie

Institute of Petroleum Engineering, Heriot-Watt University,
EH14 4AS, Edinburgh, Scotland;
E-mail: alannah.osullivan@pet.hw.ac.uk, mike.christie@pet.hw.ac.uk

Abstract: Successful reservoir prediction requires an accurate estimation of parameters to be used in the reservoir model. This research focuses on developing error models for simulation error within the petroleum industry, enabling accurate parameter estimation. The standard approach in the oil industry to parameter estimation in a Bayesian framework includes inappropriate assumptions about the error data. This leads to the parameter estimations being biased and over confident. An error model is designed to significantly reduce the bias effect and to estimate an accurate range of spread. A 2D viscous fingering example problem will be used to demonstrate both construction of the error model, and the benefits gained in doing so.

Keywords: Error model, parameter estimation, simulation error, likelihood, viscous fingering

1. INTRODUCTION

The accuracy of reservoir forecasting is limited by both the accuracy of the input data and of the method chosen for solving the system. If the data is accurate, it is expected the use of a detailed geological model to solve the flow equations representing the system will provide more accurate results compared with a reduced model. The sheer size of a reservoir makes it impossible to physically measure the properties, such as porosity and permeability in each cell for a detailed model. However an accurate description of the reservoir is needed to make an accurate forecast, leaving the true parameter values to be estimated. This problem of accurate parameter estimation is common to many applied scientific disciplines, such as weather forecasting, contaminant transport and reservoir simulation. This paper focuses on developing simulation error models for accurate parameter estimation research in the oil industry.

History matching in a Bayesian framework allows likely regions of parameter space to be identified. By simulating oil production with initial estimates from possible parameter values and comparing the simulated results with early production data, Bayesian analysis provides a formal framework for revising the parameter estimate values, [1]. Coarse models are favoured in the history matching process for time saving purposes, however this produces poor estimates for parameter space as simulation errors are introduced from using an approximate model. To over come this problem of simulation error, the idea of using an error model has been introduced. An error model is based on using simulation results from a fast model, such as an up-scaled model, together with statistical error data which is collected using a limited amount of detailed model data. The goal of an error

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

model is to produce parameter estimations with an accuracy comparable to that of a full field model, yet with a speed similar to that of a coarsely gridded model.

This paper is organised as follows. We start with a review of how parameter estimation is typically carried out under a Bayesian framework, and examine the associated assumptions. In section 3, a viscous fingering example will be set up to show the importance of an error model. This problem will also be used to show how an error model is constructed and demonstrate the benefits gained in doing so. Finally, factors input to the error model are examined in more detail to optimize results with respect to the amount of computation.

2. HISTORY MATCHING IN A BAYSIAN FRAMEWORK

Bayesian analysis is applied in the history matching process to determine parameter estimates for use in reservoir simulation and prediction. Bayes theorem, (1) is used to find appropriate regions of parameter space.

$$p(m|O) = \frac{p(O|m)p(m)}{\int p(O|m)p(m)} \tag{1}$$

The posterior probability, $p(m|O)$, is calculated from a combination of the prior and the likelihood function, $p(O|m)$. The prior is set by initial knowledge or beliefs of the parameters. Correctly calculating the likelihood function, (2), is key in defining an accurate posterior distribution. In the oil industry, the likelihood commonly assumes a Gaussian distribution for the error data.

$$p(m|O) = \exp(-M) \tag{2}$$

The probability p for the observation O to occur assuming the model m is correct, is measured in terms of likelihood function. The misfit, M in Equation (2), measures the mismatch between observed and simulated data (3).

$$M = \langle o - s|C^{-1}|o - s\rangle \tag{3}$$

For a given property to be estimated, o represents the observed data, and s the simulated data. Together with the inverse covariance C^{-1} the misfit is defined. The inverse covariance constructed in full represents spread of both data and simulation error, C_d and C_s respectively. Tarantola [2], shows under Gaussian assumption the observational error and simulation error can be combined by addition of the covariance matrices, (4).

$$C = C_d + C_s \tag{4}$$

In this paper we are concerned with quantifying simulation error only. Studies regarding simulation error have been recorded in the literature, [3],[4],[5],[6].

2.1. The Misfit Function

Typically in the oil industry, the misfit is simplified to a least squares model (5). This approach is characterised by a sum of squares of the difference between observed and simulated data and has a single value, σ^2, to represent the variance of this difference.

$$M = \sum_{i=1}^{n} \frac{(o_i - s_i)^2}{2\sigma^2} \tag{5}$$

From Gaussian statistics, this least squares method assumes the data values are independent, that is, the error is randomly spread.

Measurement of rock properties are limited by both the equipment and the user, which contributes error to the final solution. These types of data error are time independent so the least squares approach is valid for this case. Finite difference methods used in simulation progress using the solution from previous steps with an adjustment involving a time and flux combination. Thus the solution follows an evolution in time, and as such, the simulation error will also [7]. As simulation error is correlated in time the least squares method does not hold when used in conjunction with a coarse model.

By using a single value for σ^2, time dependent variance is not represented. Too small a value for σ^2 can give an over confident prediction in parameter estimation. Using an approximate reservoir model can give consistently wrong results causing heavily biased parameter predictions when using a standard least squares misfit.

3. VISCOUS FINGERING EXAMPLE

The following viscous fingering example is a relatively simple yet realistic problem for demonstrating parameter estimation. This 2D model has a set of injector wells spanning the left hand side of the reservoir, and a set of producers down the right hand side. Gas is injected into the reservoir, forcing the oil toward the producer wells. Due to the difference in mobilities between the two fluids, the injected gas fingers through the oil, Figure 1. These fingers cause early breakthrough, reducing recovery, [8],[9],[10].

The challenge is to determine a probability distribution for the viscosity of the oil given the concentration data, Figure 2, of the fluids output at the producer wells. The concentration data represents measured history data although the reservoir in this example is synthetic, thus the data is also. The data is treated as if the viscosity used in the simulation is unknown. The prior assumption is the viscosity value lies in the range $\mu = 5 \ldots 25$.

There are two main ways to solve this challenge. Using a fine grid model of the reservoir, we expect to find an accurate estimation for the oil viscosity. However, a fine grid model is too time consuming to be practical, so we compare this method with using a coarse grid model. The coarse grid model introduces errors due to using a limited amount of information but is extremely fast and simple.

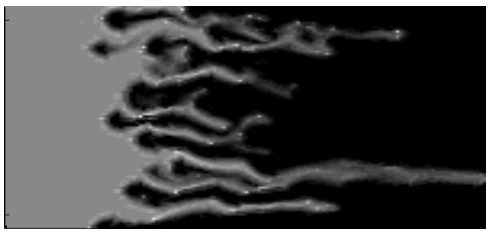


Figure 1. Viscous fingering.

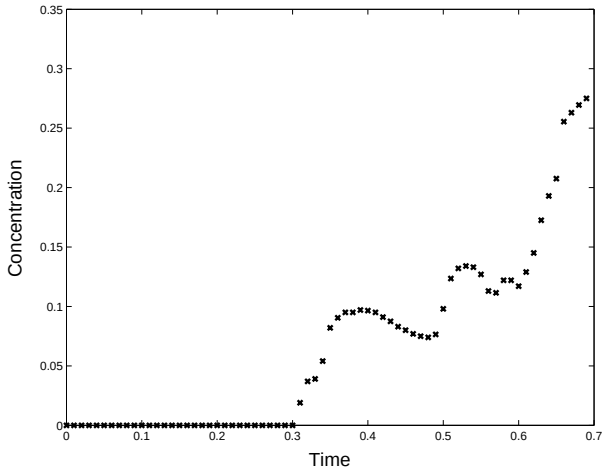


Figure 2. Concentration data for unknown viscosity value.

3.1. The Fine Grid Solution

The exact permeability field for the reservoir is unknown, thus a number of fine grid simulations are used, obtaining a mean solution with some degree of uncertainty. This method produces reasonable results, however the simulation time is long taking of the order hours on a standard work station, depending on the level of detail used in the model.

Figure 3 shows three realisations each (from sets of 20) for viscosity of $\mu = 5$, 10, and 15, as well as the unknown viscosity data. Although the results vary, Figure 3, the unknown viscosity looks to be somewhere between $\mu = 10$ and 15. At this stage, if we were to obtain a smaller range for the true viscosity value, we would proceed with further fine grid simulations, trying viscosity values in the range $\mu = 10 \ldots 15$. For each viscosity value chosen for simulation, each of the 20 realisations would need to be used in the simulations. It is easily seen that this method of parameter estimation is too time consuming and is thus impractical in an industry based situation.

For this viscous fingering example, the true value of the 'unknown' viscosity is $\mu = 13$. The fine grid solution above, although time consuming, has proven accurate.

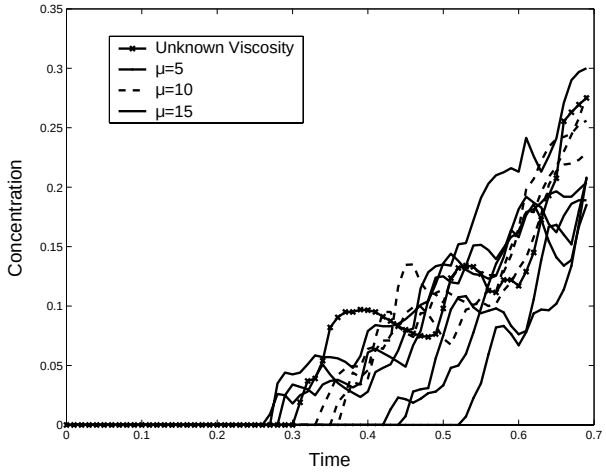


Figure 3. Fine grid approximation to the unknown data.

3.2. The Coarse Grid Solution

A coarse grid model usually consists of an up-scaled version of the fine grid model. The averaging behaviour of the up-scaling process means the full details of the reservoir are not available, hence different results are generated when compared with the fine grid solution in this viscous fingering example. Here the Todd & Longstaff model is used to approximate the average concentration at the production well in lieu of an up-scaled model [11],[12].

The Todd & Longstaff model is a well-known model which approximates the average response of the viscous fingering in the reservoir. Like up-scaled models, the Todd & Longstaff model is also subject to inaccuracies. The fractional flow in the Todd & Longstaff model is found simply from equation (6), where c is the average concentration and M_{eff} is the effective viscosity ratio.

$$f(c) = \frac{c}{c + (1-c)/M_{\text{eff}}} \tag{6}$$

M_{eff} is defined by equation (7), where M is the true viscosity ratio, μ_o/μ_s.

$$M_{\text{eff}} = (0.78 + 0.22M^{\frac{1}{4}})^4 \tag{7}$$

In this case, $\mu_s = 1$. The effective viscosity is found from using coefficients which are determined by fitting to Blackwell's experiment [13]. This model is extremely fast to solve and very simple to use.

The solution from the Todd & Longstaff model needs to be in a comparable form to the fine grid solution, which is given as concentration with time. Concentration is input to equation (6) as a linear drop across the reservoir from 1 to 0, defining the fractional flow. The derivative of the fractional flow describes the speed at which the fingers travel, [11], and can therefore be expressed as,

$$\frac{df}{dc} = \frac{x}{t}. \tag{8}$$

The Todd & Longstaff model is scalable, giving $x = 1$ at the producer wells. This enables t to calculated as $1/(df/dc)$. This construction of the Todd & Longstaff model does not contain the details of the permeability field. This is evident in Figure 4 where the Todd & Longstaff model fails predict a value for viscosity near the true value, which was noted in section 3.1 to be $\mu = 13$.

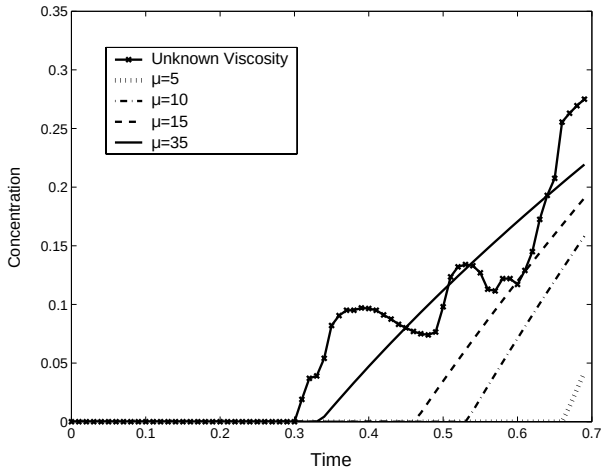


Figure 4. Todd & Longstaff model.

4. THE ERROR MODEL

The goal of an error model is to reduce bias and over confidence in predictions introduced from using a coarse model in the history matching process, so that accurate parameter estimates can be obtained by using limited fine grid simulations. The proposed way to achieve these goals is to include the mean error and time dependent covariance in the history matching process. An error model is incorporated into a solution by way of the misfit function defined in Bayes theorem. In this paper, three misfit definitions will be studied. The first is the least squares misfit, which has already been described as inappropriate for simulation error,

$$M = \sum_{i=1}^{n} \frac{(o_i - s_i)^2}{2\sigma^2}.$$

The second takes the same form as the least squares misfit, and also includes the mean error,

$$M = \sum_{i=1}^{n} \frac{(o_i - s_i - \bar{e})^2}{2\sigma^2}. \tag{9}$$

The third definition includes both mean error and full covariance,

$$M = \frac{1}{2}(o - s - \bar{e})^T C^{-1}(o - s - \bar{e}). \tag{10}$$

If the fine grid simulation is taken as truth, then the simulation error (11), can be calculated from the fine grid simulation minus the coarse grid (Todd & Longstaff) result,

$$e_j = FG_j - CG_j. \tag{11}$$

In equation (11), j is the realisation number. Figure 4 shows 20 fine grid simulations

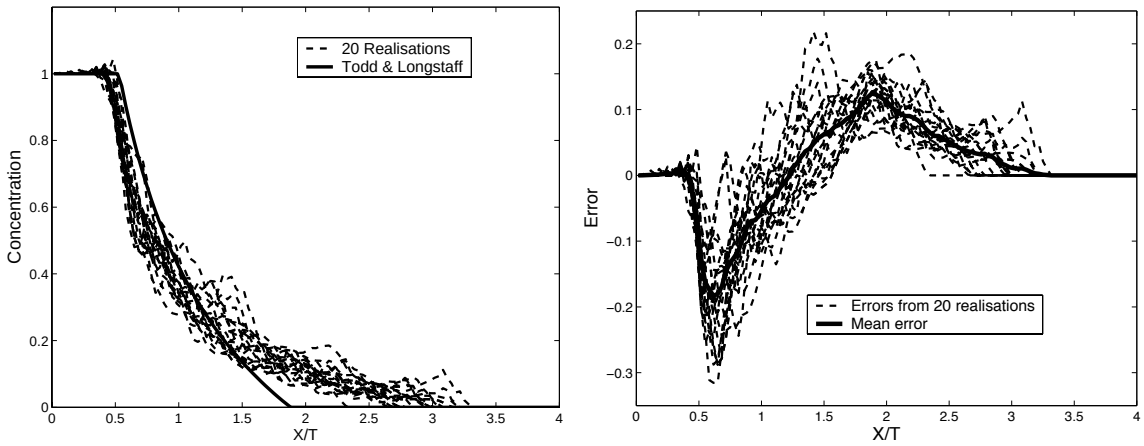


Figure 5. The Todd & Longstaff approximation and mean error.

and the Todd & Longstaff approximation. Todd & Longstaff has an early break through point and is late in sweeping all the oil to the producer well. However it approximates the average behaviour of the concentration quite well. Figure 5 shows the error plots, that is, e_j from equation (11). The curve in bold is the mean error (12), for the chosen viscosity. The covariance is also calculated from the 20 realisations, (13), [14]. Three sets of mean error data and covariance data for error are calculated for viscosity values μ =5, 10 and 15 and these points are known as the *base points* for this example.

$$\bar{e}(\xi) = \frac{1}{n}\sum_{j=1}^{n} e_j(\xi) \tag{12}$$

$$C(s,t) = \frac{1}{n-1}\sum_{j=1}^{n}(e_j(t) - \bar{e}(t))(e_j(s) - \bar{e}(s)) \tag{13}$$

4.1. Interpolated Data

The next stage in constructing the error model is to approximate the mean and covariance between the base points, limiting the fine grid realisations to just the base points of the error model. This means Bayes theorem can be applied over regions for which no fine grid simulations have been run. In this case, MATLAB's linear interpolation function was used. Figure 6 shows the interpolated errors at integer values between the base points.

Observing the peaks and troughs of the interpolated error data, it is seen that this scheme is not ideal. However outside the peaks and troughs, the interpolation scheme performs well, and is for now, used in the model. The interpolation scheme will be reviewed in section 6. For viscosities greater than $\mu = 15$ the data has been set to that of $\mu = 15$, satisfying the prior, 3. Now we have sufficient data for the error model.

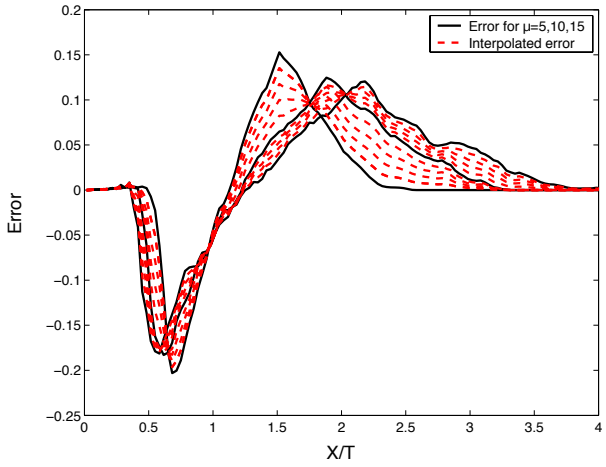


Figure 6. Interpolated data.

5. RESULTS

The three misfit definitions defined in the previous section, will be used to define the likelihood functions, giving 3 probability distribution curves. Recall the true value of the viscosity of the 'unknown data' is $\mu = 13$. The three definitions are referred to as

- LSQ
  Least squares model, (5)
- PME
  Least squares with mean error included, (9)
- FEM
  Full error model, which includes mean error and covariance, (10)

The prior model states the viscosity is given by a uniform probability distribution function in the range $\mu = 5 \ldots 25$. Using the data with 'unknown' concentration the three likelihood functions are calculated, predicting the probability distribution functions shown in Figure 7. The likelihoods were obtained by the coarse grid model and the interpolated fine grid model being defined at 0.1 viscosity increments up to 0.7 pore volumes injected. The LSQ misfit predicts a maximum likelihood of $\mu = 25$ for viscosity. The true value is $\mu = 13$, showing the LSQ model has given a heavily biased result. The other two misfit models show the bias effect to be almost completely removed.

The main difference between these two distributions is the variance, or width of the probability distribution. The PME model used a guessed variance of $\sigma^2 = 0.01$, which gives a fairly wide likelihood curve. The FEM model predicts close to the actual truth with a tighter distribution.

A benefit to having a more exact range for variance lies in the history matching process. When using a single value for variance, parameter space maybe defined either unnecessarily large, extending computation time as the prior contains unnecessary information,

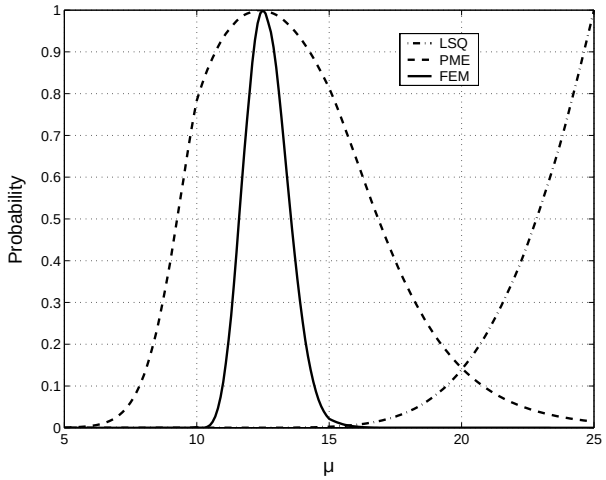


Figure 7. Comparison of likelihood functions, mean of 20 realisations.

or too narrow so that some important parts of parameter space are neglected. Using a correct variance we can home in on the correct region of parameter space more quickly.

5.1. Overall Performance

Figure 8 shows the overall performance of the different misfit functions. The x-axis shows the true viscosity that was input to the Todd & Longstaff model, and the y-axis shows the predicted value or maximum likehood prediction. The least squares approach over predicts the viscosity value in each case and quickly reaches the prior restriction set in the model, $\mu = 25$ for viscosity. The remaining two misfits perform better than the least squares approach, lying closer to the ideal case. Figure 8 shows the FEM to predict only slightly closer to the ideal case than PME, that is, it appears there is no noticeable advantage in using the full error model. If this is the case, it would make sense to use PME misfit to reduce the computing time. Figure 9 shows error bounds (p10 and p90 estimates) on the maximum likelihood predictions for the two mean error models. Also plotted on both graphs is the ideal case. It is clear the full error model shows a smaller range of viscosity predictions than the PME case, making the full covariance calculation worthwhile.

5.2. Compared with Detailed Simulation

The averaged set realisations provided as production data was accurately shown to have a maximum likelihood value of near $\mu = 13$ when tested with the FEM. The likelihood function for this realisation can be compared with the likelihood function created from using detailed simulation as simulation data.

Fine grid simulations were run for viscosities at integer values between $\mu = 5 \ldots 15$, and the error computed (11). The errors at these integer values are then interpolated with MATLAB's linear interpolation function at 0.1 increments to produce fine grid data over a range of viscosity values. It has been shown in section 3.2 that using a coarse

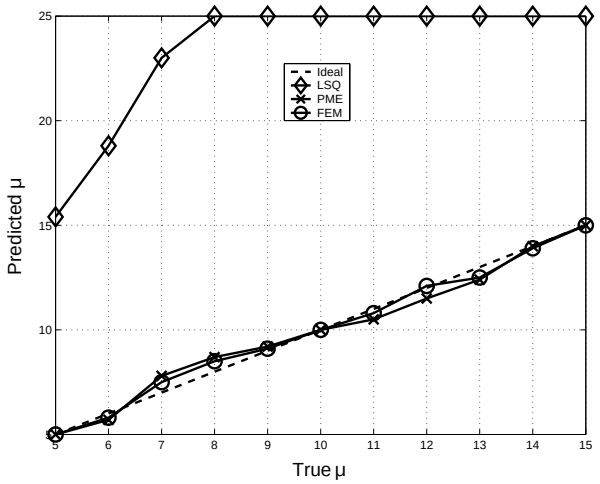


Figure 8. Overall performance of the misfit functions.

grid model in simulation gives a significant model error which is correlated in time. This correlation stemmed from the coarse grid model predicting early break through and being late in sweeping the oil to the producer well. When fine grid simulation is used as the simulated data this error correlation does not exist, as for this example the observed data and the simulated data were created with the same grid size and set of statistics. The likelihood for the detailed simulation case has therefore been created with least squares model alone as an error model for simulation is not required. The LSQ likelihood function for fine grid simulation is shown in Figure 10 with the FEM likelihood result.

A calibration curve is plotted to demonstrate how close the two likelihoods are in their predictions. This is done by calculating the cumulative probabilities from the likelihood curves. Figure 10 plots the cumulative data for the likelihood belonging to the FEM on the x-axis, and the cumulative data for the detailed simulation on the y-axis. The resulting curve is close to the ideal case, which is shown as the straight line. If the two likelihood functions were indistinguishable the calibration plot would show this as a straight line. This is represented in the calibration plot Figure 10 as the ideal case. The actual calibration plotted in the figure does not stray too far from the ideal case.

6. ISSUES

The results from the previous section are promising and the rest of the paper is dedicated to studying details of the error model. We investigate whether similar results can be obtained for less computation effort. Also other aspects are investigated such as the interpolation method.

6.1. Base Points

The first part of the error model to be revisited is the number of base points. Previously 3 base points were chosen, at $\mu = 5$, 10 and 15. As the base points are the part of the model that requires some fine grid simulation, it is the most computationally expensive

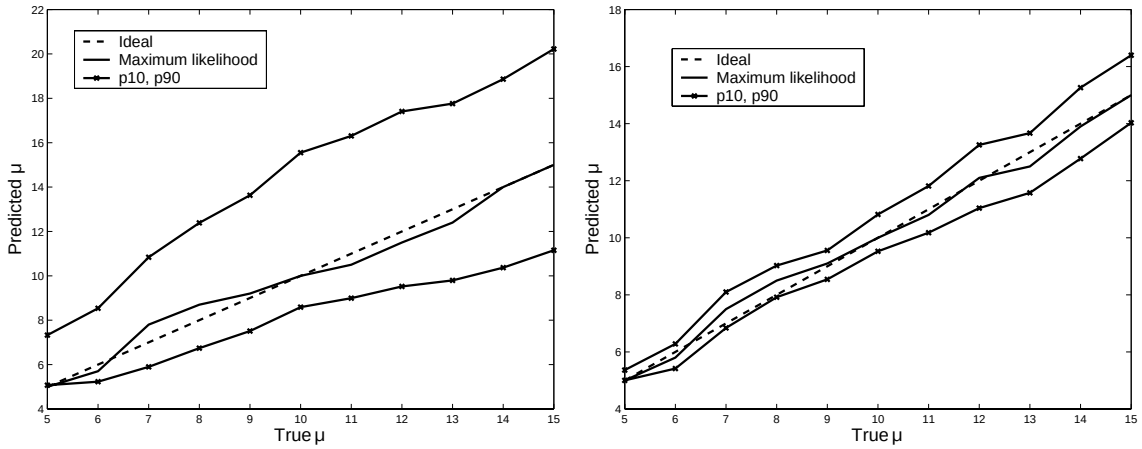


Figure 9. Error bounds for PME (left) and FEM (right).

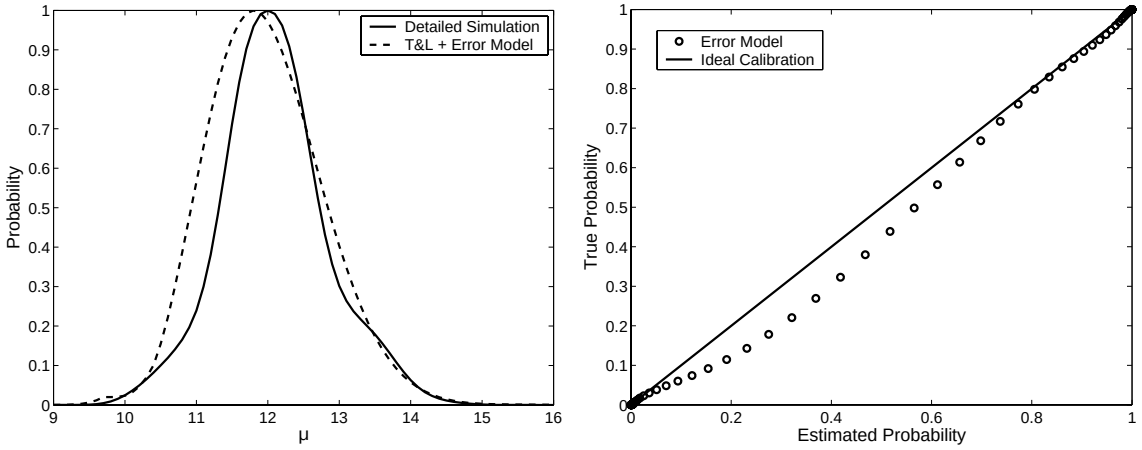


Figure 10. Comparison of Todd & Longstaff model and detailed simulation likelihoods, mean of 20 realisations (left) and calibration plot based on mean of 20 realisations (right).

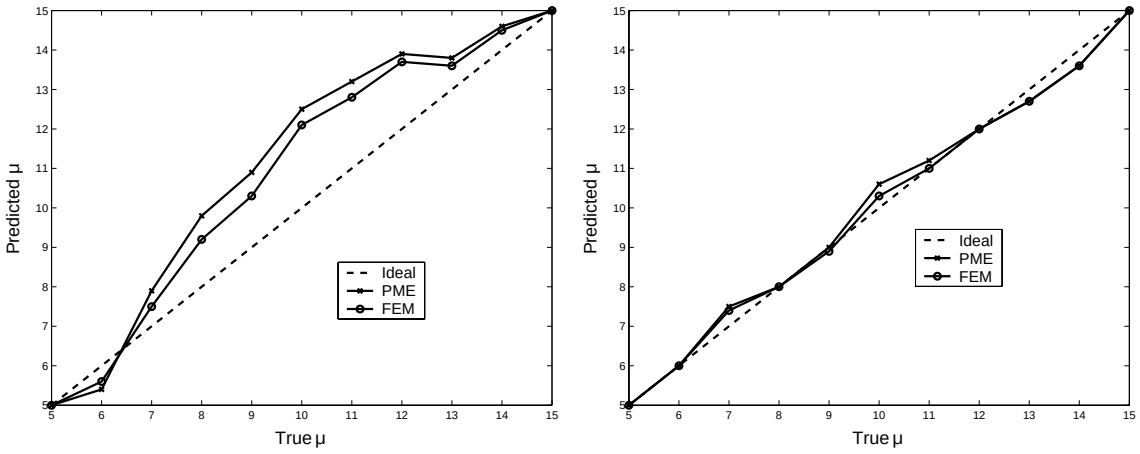


Figure 11. Overall performance: two base points (left) and four base points (right).

part. For this reason, including more or fewer base points than 3 is studied to see the effect on parameter estimation.

First two base points, at $\mu = 5$ and 15 are examined, Figure 11. Only the PME and FEM are compared, as the LSQ model is not competitive against these two models. Near the base points, PME and FEM are able to predict viscosity well. Away from the base points, both likelihoods begin to show signs of bias. The two models have similar behaviour and again as in the case with 3 base points, the FEM is slightly more accurate in predicting the maximum likelihood value. While bias effects are visible away from the base points, they are not as extreme as in the LSQ model.

Figure 11 shows predictions from both models from using 4 base points, at $\mu = 5, 8, 12$ and 15. Again the FEM narrowly obtains more accurate predictions. Overall there doesn't appear to be much gained in using 4 base points over 3. As the fine grid realisations are expensive to compute, it would be advised not to compute them unless necessary.

Table 1 shows results for PME and FEM with 2, 3 and 4 base points. All scenarios correctly predict at the base points, as that is how the error model is defined. Between the base points, typically the likelihoods slightly over predict rather than under predict and this is most noticeable with just two base points.

6.2. How Many Realisations Should be Used?

Different realisations created by a random permeability field generator give rise to differing predictions. This study has used a set of 20 realisations to compute the likelihoods at each base point. Although the results are not presented in this paper, it was found that using different sets of 20 realisations gave differing results for parameter predictions. As a consequence, we investigate how many realisations are required in simulation for optimum results. Figure 12 shows likelihoods created by using increasing numbers of realisations. The first likelihood curve has used 5 realisations in the error model, the second 10, and so on. As more realisations are used, the maximum likelihood is converging to a value,

Table 1. Comparison of 2,3 and 4 base points.

	PME			FEM		
True	2	3	4	2	3	4
5	5.0	5.0	5.0	5.0	5.0	5.0
6	5.4	5.7	6.0	5.6	5.8	6.0
7	7.9	7.8	7.5	7.5	7.5	7.4
8	9.8	8.7	8.0	9.2	8.5	8.0
9	10.9	9.2	9.0	10.3	9.1	8.9
10	12.5	10.0	10.6	12.1	10.0	10.3
11	13.2	10.5	11.2	12.8	10.8	11.0
12	13.9	11.5	12.0	13.7	12.1	12.0
13	13.8	12.4	12.7	13.6	12.5	12.7
14	14.6	14.0	13.6	14.5	13.9	13.6
15	15.0	15.0	15.0	15.0	15.0	15.0

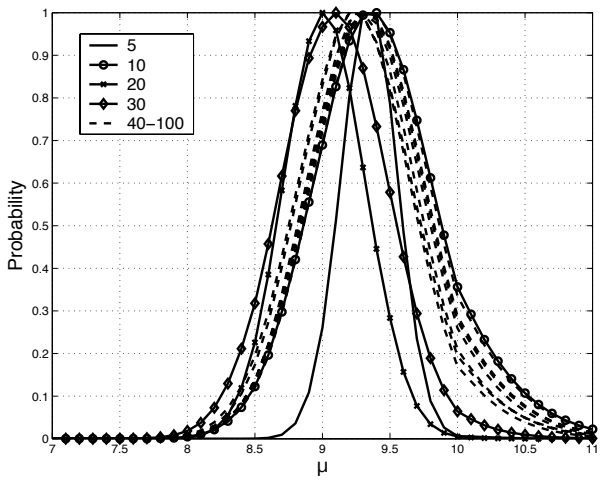


Figure 12. Likelihoods for increasing numbers of realisations.

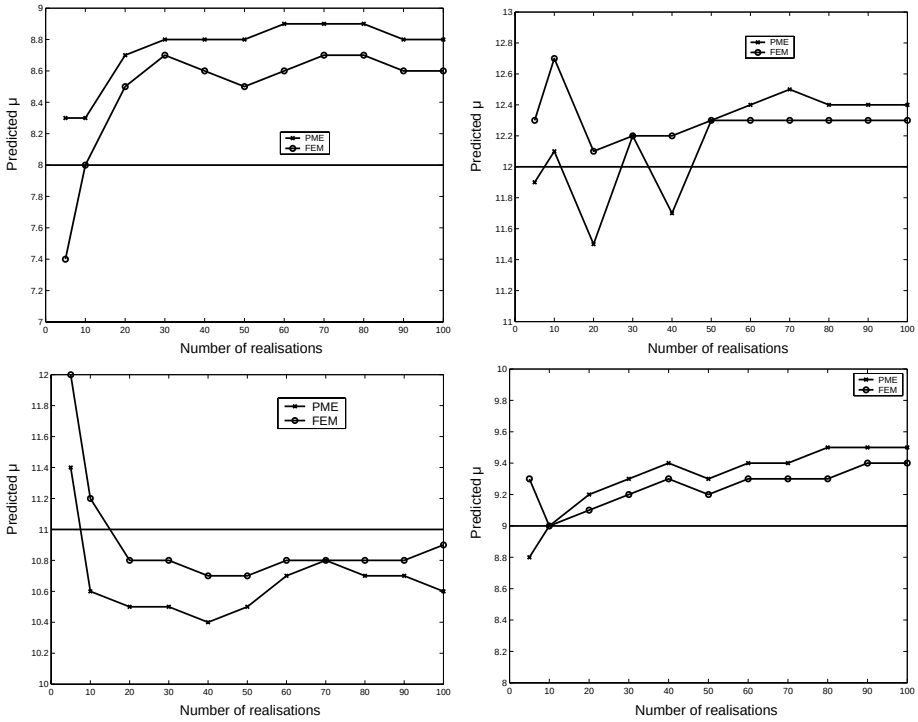


Figure 13. Convergence of likelihood functions, $\mu = 8, 9, 11$ and 12.

however it is not to the true viscosity, showing there is still a slight presence of bias. Figure 13 shows as the number of realisations used increases, the predicted viscosity tends toward some value. Although some bias is still present, there is not much variation in prediction with the full error model after 20 realisations. 20 realisations seems an appropriate number in this case.

6.3. Interpolation Scheme

As stated earlier, the interpolation scheme used thus far is MATLAB's linear interpolation scheme, Figure 14. This interpolation scheme did not adequately represent the peaks and troughs for the error data. An interpolation scheme has been created to produce a fair representation for the expected form of the error, Figure 14. This interpolation scheme focused on 3 points. The first point being where the mean error plot left zero, A. The second at its maximum, B, and the third, the end point, C. Between $\mu = 5$ and 10, the necessary number of evenly separated points were added, giving new A, B, C for intermediate viscosities. To link the points, the shapes of the two outer curves were followed. Some smoothing was necessary. The same was done for between $\mu = 10$ and 15 and the final effect is shown in the Figure 14.

The next step is to compare the maximum likelihood estimates with both schemes. The results are shown in Figure 15. By eye, looking at the limited examples, one interpolation scheme does not appear more favourable over the other. A more thorough investigation took place, studying the $5 \ldots 100$ sets of realisations with both interpolation schemes. 121

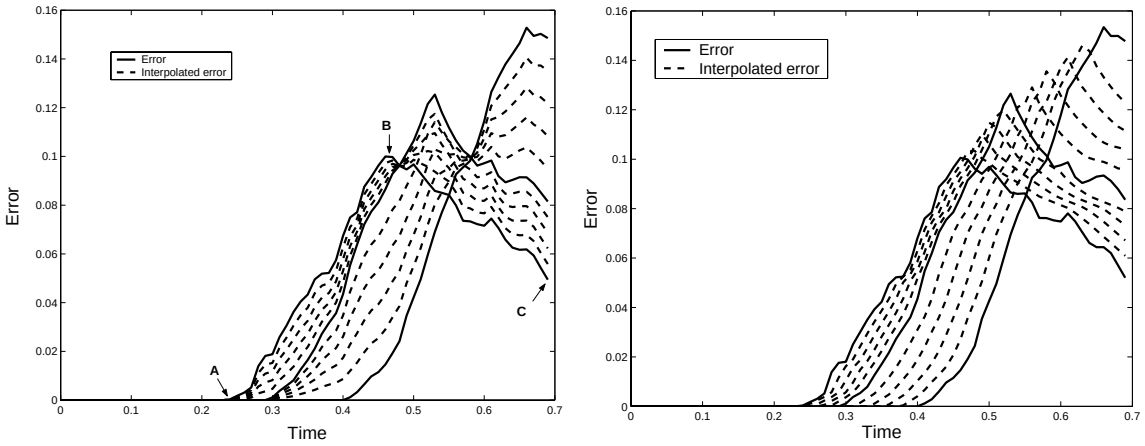


Figure 14. MATLAB's interpolation scheme (left) and improved interpolation scheme (right).

Table 2. Comparison of interpolation schemes.

5...100	PME	FEM
MATLAB	19	30
IMPROVED	57	40
EQUIVALENT	45	51

comparisons are made, 11 different sets of realisations for 11 viscosities. The results are summarised in Table 2, showing on how many occasions each method performed best or if they were equivalent. The improved scheme produced a more accurate estimate more often than MATLAB's linear interpolations scheme. They performed equally well on a number of occasions.

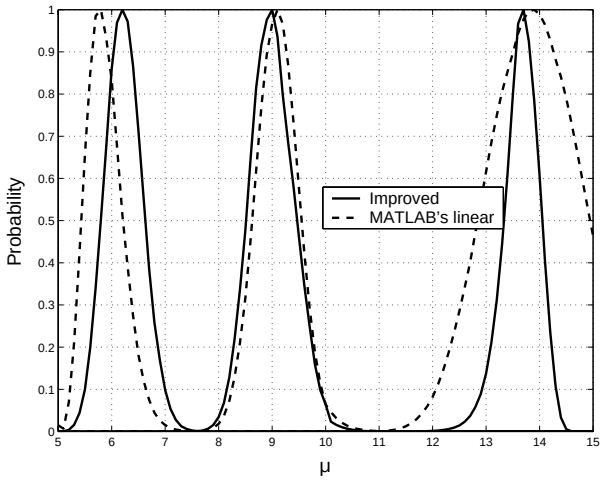


Figure 15. Comparison of interpolation schemes.

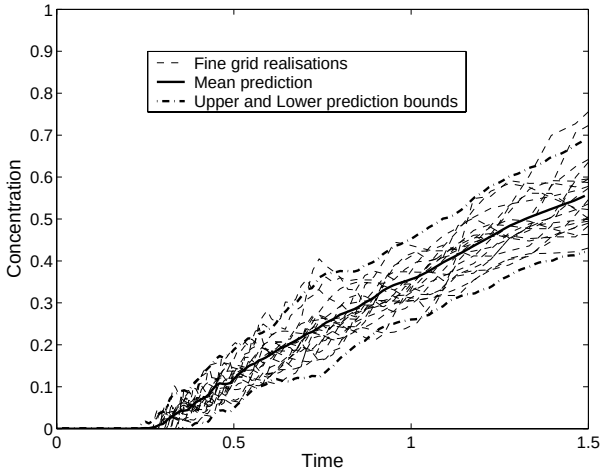


Figure 16. Error model prediction.

7. PREDICTIONS

The error model has shown to significantly reduce the bias introduced from using a coarse model, giving an accurate estimation of the parameter with limited effort. The purpose of determining properties for the system in this way is to be able to use the values to make accurate predictions for oil recovery. The next step is to check how well the error model is able to make a prediction.

The concentration data output from the simulator was defined for each cell in terms of $\xi = X/T$. The data is converted to units of time by realising the length of the reservoir is 1 when compared it to the scalable Todd & Longstaff model, thus $T = 1/\xi$, section 3.2. Hence, we can look forward in time. The 20 fine grid realisations for $\mu = 13$ are extended to $T = 1.5$ and are shown as black dotted curves on Figure 16. To see how the error model solution compares with these fine grid simulation results, the Todd & Longstaff result was calculated for $\mu = 13$ and the error model applied. The error model data consists of mean error and covariance for a range of viscosity values. To apply the error model to the coarse grid solution, the mean error from the error model for $\mu = 13$ is added to the coarse grid solution. This removes the bias incurred from the Todd and Longstaff solution. This is shown by the black solid curve in the figure. Next, $\pm 2\sigma$ for $\mu = 13$ from the error model was added to the newly calculated mean solution. These are bands are the spread of results predicted from using the error model with the coarse grid solution. These bands do encompass most of the fine grid results, making an accurate prediction.

8. CONCLUSIONS AND DISCUSSION

Using a coarse model in simulation for parameter estimation is necessary for a practical computation time. Using a least squares misfit in combination with a coarse grid model introduces simulation error, giving biased and overconfident parameter estimates. This provides a motivation for error modelling in reservoir prediction.

A 2D viscous fingering example was introduced, providing a means for explaining how an error model should be constructed and tested for parameter estimation. The parameter to be estimated is the viscosity of the oil. The error model was constructed in the following way.

- 20 fine grid realisations were generated at 3 base points, $\mu = 5, 10$ and 15.
- A coarse grid approximation known as the Todd & Longstaff model was calculated for the same three viscosity values.
- The simulation error for each viscosity in the example is found by subtracting the coarse grid solution from the fine grid solution.
- The mean and variance of the error for each base point is calculated and MATLAB's linear interpolation function is used to predict the mean error and variance at intermediate viscosity values.
- The error data is added to the coarse grid solution producing a fast simulation model with the bias removed by including mean error and overconfidence avoided by using a fair representation of the spread of error.

A number of variables were studied to minimise the amount of work put into the model with maximised results.

- It was found for the example given that 3 base points were sufficient. There was no significant gain in accuracy for the work put in with using 4 base points. Fewer than 3 base points increased bias effects.
- The number of realisations used for fine grid simulations was varied from $5 \ldots 100$. 20 realisations were more than adequate, with no further improvement gained from using 100 realisations.
- The interpolation scheme was improved, allowing peaks and troughs of the error data to be accurately resolved. This had a positive effect on parameter estimation, predicting values closer to the actual value than with the initial scheme.

Finally the error model was used to make a prediction. The coarse grid result was shown to make an accurate prediction when the error model data was added. The mean error significantly reduced the bias effects while the covariance gave a realistic spread from the mean prediction.

ACKNOWLEDGMENTS

The authors wish to thank the sponsors of the Uncertainty Project at Heriot-Watt University for financial support: Anadarko, BG, BP, CononcoPhillips, DTI, ENI, JOGMEC, Landmark, Shell and Statoil.

REFERENCES

1. D. Sivia. *Data Analysis: A Bayesian Tutorial*. Oxford Science Publications, Oxford, 1996.
2. A. Tarantola. *Inverse Problem Theory, Methods for data fitting and model parameter estimation*. Elsevier Science Publishers, Amsterdam, 1987.
3. B. DeVolder, J. Glimm, J. W. Grove, Y. Kang, Y. Lee, K. Pao, D. H. Sharp, and K. Ye. Uncertainty quantification for multiscale simulations. *J. Fluids Engineering*, 2002.
4. J. Glimm, S. Hou, Y. Lee, D. Sharp, and K. Ye. Prediction of oil production with confidence intervals. *SPE 66350*, 2001.
5. J. Glimm, S. Hou, H. Kim, Y. Lee, D. Sharp, K. Ye, and Q. Zou. Risk management for petroleum reservoir prediction: A simulation-based study of prediction. *Computational Geosciences*, 2001.
6. J. Glimm, S. Hou, Y. H. Yee, D. H. Sharp, and K. Ye. Solution error models for uncertainty quantification. *AMS*, 2002.
7. J. Glimm and D. H. Sharp. Prediction and the quantification of uncertainty. *Physica D*, 1999.
8. M. Blunt and M. Christie. How to predict viscous fingering in three component flow. *Transport in Porous Media*, 1993.
9. M. Christie, A Muggeridge, and J. Barley. 3d simulation of viscous fingering and wag schemes. *SPE Res. Eng.*, pages 19–26, 1993.
10. F. Fayers, M. Blunt, and M. Christie. Accurate calibration of empirical viscous fingering models. *Oil & Gas Science and Technology*, 46, No. 03:311–325, 1991.
11. M. Blunt and M. Christie. Theory of viscous fingering in two phase, three component flow. *SPE Advanced Technology Series*, 1993.
12. B. Rubin, J. Barker, M. Blunt, and M. Christie. Compositional reservoir simulation with a predictive model for viscous fingering. *SPE 25234*, 1993.
13. R. Blackwell, J. Rayne, and W. Terry. Factors influencing the efficiency of miscible displacement. *Trans AIME*, 216:1–8, 1959.
14. J. Jensen, L. Lake, P. Corbett, and D. Goggin. *Statistics for petroleum Engineers and Geoscientists*. Elsevier, Amsterdam, 2000.

Local and Global Uncertainty Analysis of Complex Chemical Kinetic Systems

J. Zádor, I. Gy. Zsély and T. Turányi

Department of Physical Chemistry, Eötvös University (ELTE),
Pázmány P. stny., Budapest H-1065, Hungary
E-mail: zador@chem.elte.hu, zsigy@vuk.chem.elte.hu, turanyi@elte.hu

Abstract: Computer modelling plays a crucial part in the understanding of complex chemical processes. Parameters of elementary chemical and physical processes are usually determined in independent experiments and are always associated with uncertainties. Two typical examples of complex chemical kinetic systems are the combustion of gases and the photochemical processes in the atmosphere. In this study, local uncertainty analysis, the Morris method, and Monte Carlo analysis with Latin hypercube sampling were applied to an atmospheric and to a combustion model. These models had 45 and 37 variables along with 141 and 212 uncertain parameters, respectively. The toolkit used here consists of complementary methods and is able to map both the sources and the magnitudes of uncertainties. In the case of the combustion model, the global uncertainties of the local sensitivity coefficients were also investigated and the order of parameter importance based on local sensitivities were found to be almost independent of the parameter values within their range of uncertainty.

Keywords: local uncertainty analysis, Morris method, Monte Carlo method, atmospheric chemistry, combustion modelling, mechanism validation

1. UNCERTAINTY ANALYSIS METHODS

Uncertain parameters used in a model give rise to uncertainties in simulation results. Highly nonlinear models tend to magnify the uncertainty of some parameters and damp the uncertainty of others. There is a wide range of methods for uncertainty analysis [1], which differ from each other in their applicability to different types of models, in the scope of information provided, and in the level of sophistication and computational demand. For chemical kinetic models, the most comprehensive task is the conversion of the joint probability density function (*pdf*) of the parameters into the *pdf*'s of the simulation results. A more modest request is the estimation of the variance of results from the variance of parameters. In this work, several types of uncertainty methods were used and the uncertainty indicators obtained from them were compared.

1.1. Local Uncertainty Analysis

Local uncertainties were calculated by combining local sensitivity coefficients s_{ij} [2] with uncertainty estimates of the input parameters [3]. An individual contribution $\sigma_j^2(c_i)$ of rate coefficient k_j to the total uncertainty of concentration c_i can be expressed as:

$$\sigma_j^2(c_i) = \sigma^2(\ln k_j)\left(\frac{\partial c_i}{\partial \ln k_j}\right)^2, \qquad (1)$$

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

where $\sigma^2(\ln k_j)$ is the variance of the logarithm of rate coefficient k_j. The overall variance $\sigma^2(c_i)$ of the output concentration c_i is:

$$\sigma^2(c_i) = \sum_j \sigma_j^2(c_i). \tag{2}$$

The individual contribution of k_j can be expressed as percentage s_{ij} %:

$$s_{ij}\% = \sigma_j^2(c_i)/\sigma^2(c_i) \times 100. \tag{3}$$

The main drawback of this method is that the linearity assumption is applied for a highly nonlinear chemical kinetic model.

1.2. Morris Method

Morris method is a screening method [1]. Screening methods are relatively cheap, compared to Monte Carlo (MC) type methods, but are investigating the model on a global range, i.e. the input parameters are varied over the whole range of their possible values. In the Morris method, the uncertainty is characterised by a value called *effect*, which is assigned to each uncertain parameter for each investigated output result. This effect is calculated several times, by varying the input parameter set according to a given algorithm. The results of the Morris analysis are usually shown on a graph, where the horizontal axis refers to the mean of the calculated effects, while the vertical axis represents the standard deviation of the effects.

This procedure enables the selection of important parameters, by evaluating the model with various input parameter sets. Besides importance, information on the type of the effect of the parameter is also obtained: it is possible to distinguish parameters with linear effects from parameters with nonlinear or interaction effects. The drawback of this method is that it does not provide information on the magnitude of the uncertainty of the output variables. The other weakness of this method is that it does not take the shape of the *pdf* of the parameters into account.

1.3. Monte Carlo Simulations with Latin Hypercube Sampling

The above methods are computationally cheap, but are not able to provide the exact and unbiased *pdf* of the output values. For this reason, Monte Carlo type simulations were also carried out. To keep the number of runs as low as possible, Latin hypercube sampling was applied. This sampling covers the parameter space with minimal sample size and in an unbiased manner [1]. The number of runs was 3000 in all calculations.

1.4. Assignment of Uncertainties to Parameters

Great attention was paid to the careful selection of input uncertainties. Uncertainty factors were collected from chemical databases [4-7], which are critically evaluated and are frequently updated. These factors were converted to the variance of the parameters using the method described in article [3]. If no uncertainty factor was found for a reaction, then a thorough literature search was carried out and this factor was estimated. The *pdf* of the parameters were also established; lognormal distribution was assumed for rate coefficients, normal distribution for heat-of-formation data, and uniform distribution for parameters of other type (e.g. channel ratios).

2. UNCERTAINTY ANALYSIS OF A PHOTOCHEMICAL AIR POLLUTION MODEL

2.1. Brief Description of the Model

Detailed uncertainty analysis was carried out on the photochemical degradation model of ethene that is implemented in the Master Chemical Mechanism version 3 (MCMv3) [8]. The MCMv3 is an explicit chemical mechanism, containing the photochemical degradation scheme of more than 120 volatile organic compounds, and incorporating approximately 10000 reactions of 2500 species. The initial compounds in our model were ethene (C_2H_4) and nitrogen oxides (NO_x). The submechanism of ethene degradation contains 141 reactions of 45 chemical species. For brevity, in this paper the results concerning the uncertainty of ozone (O_3) concentration are discussed only. Ozone is one of the most important photochemical pollutants and the prediction of its concentration from the initial concentration of pollutants is crucial. More results are presented in a recently submitted paper [9].

2.2. Experimental Results

The methods of uncertainty analysis were developed for comparison with measurements made in the European Photoreactor (EUPHORE) at Valencia, Spain. This is a so-called smog chamber, where the chemical compounds are injected into a tent having Teflon walls, and their concentration–time profiles are followed by state-of-the-art analytical instrumentation. There are three sources of error when chemical models are tested against smog chamber measurements: (i) measurement errors, (ii) errors introduced by chamber specific effects and (iii) errors and uncertainties in the model itself. Our work aimed to reveal the significance of the various error sources.

The ethene oxidation model was tested at two experimental circumstances: one with a high, the other with a low initial NO_x concentration. Fig. 1 shows that the slope of the calculated maximal ozone concentration depends on the ratio of the initial ethene and NO_x concentration. The two cases that we selected represent two fundamentally different regions according to this plot. In the low NO_x case, changing the initial ethene concentration does not effect the maximal ozone concentration, while in the other case ozone increases with increased initial ethene concentration.

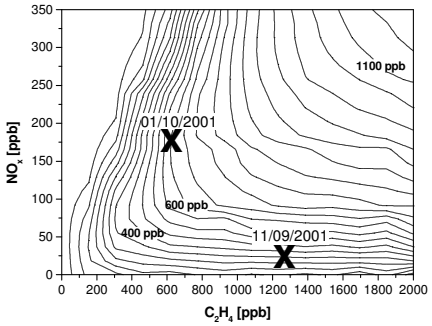
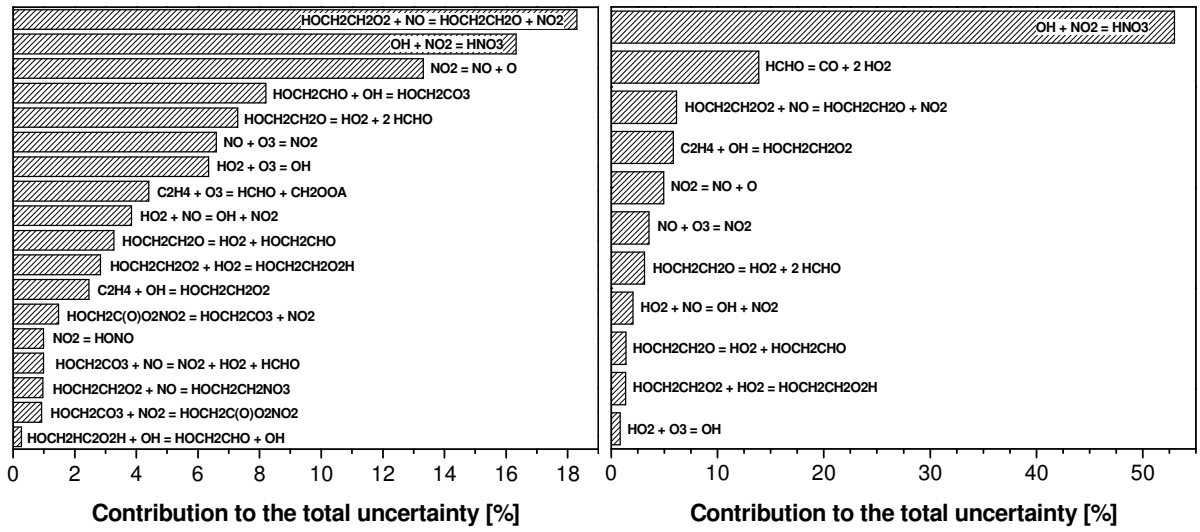


Figure 1. Simulated maximal ozone concentrations as a function of initial concentrations. The 'X' signs show the initial concentrations for the two investigated experiments.



(a) (b)

Figure 2. Contribution of the uncertainty of the rate coefficients to the uncertainty in the calculated ozone concentration at the end of the experiment at (a) low and (b) high NO_x conditions as determined by local uncertainty analysis.

2.3. Results for the Atmospheric Chemical Model

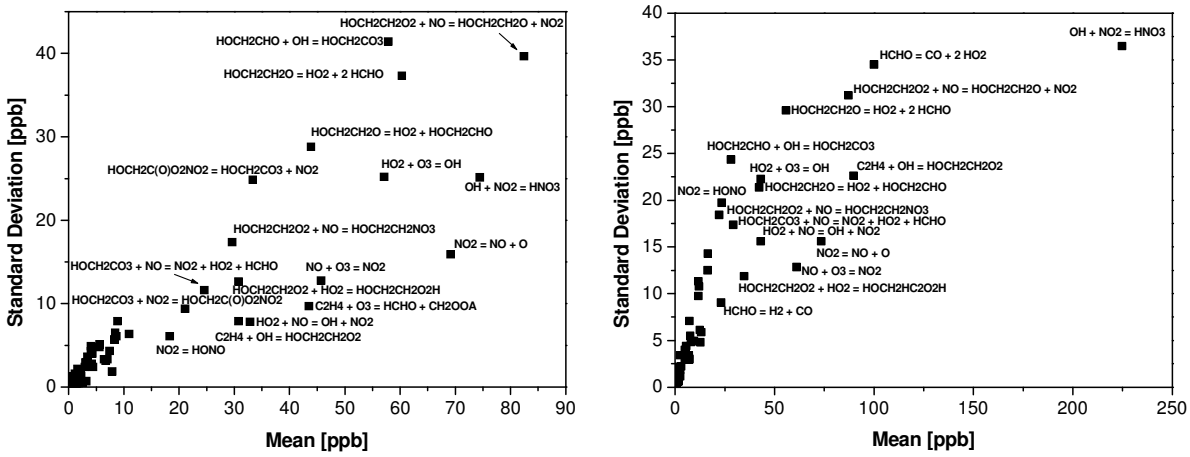
2.3.1. Local uncertainty analysis

Fig. 2 shows reactions with the highest uncertainty contribution s_{ij} % (see Equation (3)) to the calculated final O_3 concentration. The cut-off criterion was 1 % compared to the reaction having the highest uncertainty contribution. It is well visible that in the low NO_x case more reactions have significant contribution to the uncertainty of the ozone concentration.

In the low NO_x case reactions $HOCH_2CH_2O_2 + NO = HOCH_2CH_2O + NO_2$, $OH + NO_2 = HNO_3$ and $NO_2 = NO + O$ account for about 50 % of the total O_3 uncertainty. The overall 2σ uncertainty for ozone, calculated by Equation (2), is 20 %. In the high NO_x case, more than 50 % of the total uncertainty in ozone originates from reactions $HCHO + h\nu = 2HO_2 + CO$ and $OH + NO_2 = HNO_3$. The overall uncertainty of calculated ozone concentration at the final time was found to be 29 %.

2.3.2. Morris analysis

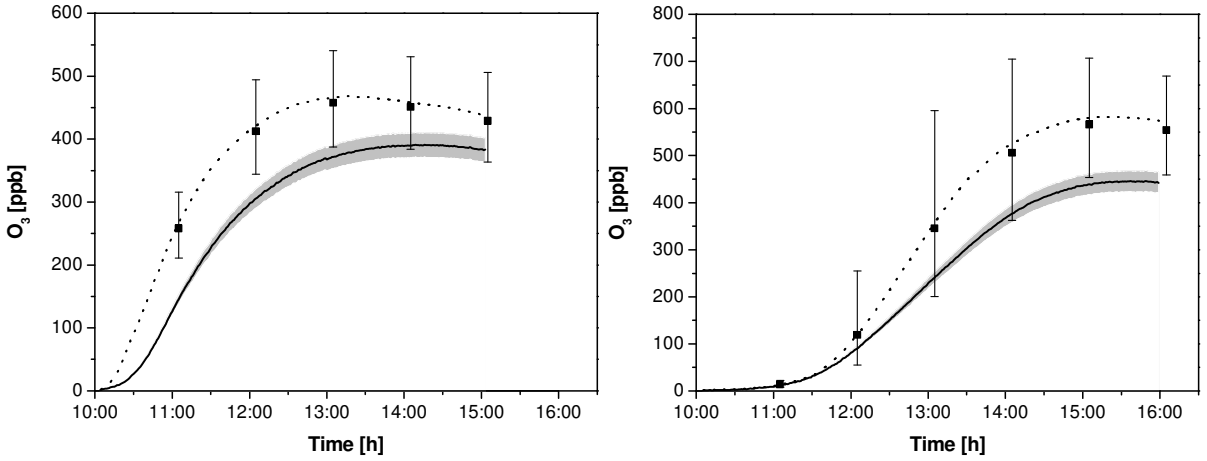
Fig. 3 shows the results of the Morris analysis for ozone. The mean values are in excellent accordance with the results of the local uncertainty analysis. The standard deviations provide interesting insights into the linearity assumption used, because ranking the rate coefficients according to their standard deviation differs from the ranking that results from the means. Under both conditions, the reactions of the $HOCH_2CH_2O_2$ and $HOCH_2CH_2O$ molecules are ranked higher on the standard deviation scale than on the mean scale. This can be due to the fact, that these parameters are important only in a period of the oxidation.



(a) (b)

Figure 3. The mean and the standard deviation effects, calculated with the Morris method for O_3 in the (a) low and (b) high NO_x case.

In both cases, less than 20 out of the 141 reactions are responsible for most of the uncertainties in the final ozone concentration. The mean values of the effects in the high NO_x case are twice those in the low NO_x case. However, the standard deviations are about the same, which suggests that the nonlinear behaviour is about the same for the two experiments. There is a significant correlation between the mean and the standard deviation: rate coefficients with great absolute effects tend to have high nonlinear effects as well.



(a) (b)

Figure 4. Comparison of modelled and measured concentrations of ozone and their associated uncertainties. Measurement and associated uncertainty (2σ): black line with grey band; model mean and its uncertainty (2σ): black dots and error bars; simulation with nominal parameter values: dotted line.

2.3.3. Monte Carlo simulations

Fig. 4 compares the uncertainty ranges of the measurements with that of the MC analysis results. The latter indicated that the distribution of the calculated ozone concentration is always nearly lognormal, while measurement uncertainties were assumed to be normally distributed. The overlap of the 2σ uncertainty limits of the measurement and model calculation is marginal, which suggests a systematic over-prediction of ozone concentration. There are difficulties in simulating both the rise time and the peak O_3 concentration. The uncertainty contribution of reaction OH + NO_2 = HNO_3 is high to the ozone concentration uncertainty (see Figs. 2 and 3). The rate coefficient of this reaction has been extensively studied recently [4], therefore it is expected that the uncertainty of the rate coefficient of this reaction will decrease significantly in the near future. In this case, uncertainty in the high NO_x case will drop significantly and consequently the overlap is likely to disappear.

3. UNCERTAINTY ANALYSIS OF A METHANE FLAMES

3.1. Brief Description of the Methane Oxidation Mechanism

Until recently, there are only few applications of uncertainty analysis to the investigation of complex combustion mechanisms. In our combustion calculations, a stationary, laminar methane flame was investigated and the simulations were performed with the Leeds Methane Oxidation Mechanism [10]. This chemical mechanism contains 175 chemical reactions and 37 chemical species. Our aim was to determine the uncertainty of simulation results caused by the uncertainty of thermodynamic and kinetic parameters. The investigated results included the concentration maximum of some important species (H, O, OH, CH, CH_2), the maximum temperature, and the laminar flame velocity (v_L); the latter number is characteristic to a freely propagating flame and is often used when model and measurement are compared [11]. The simulations were carried out with the CHEMKIN-II package [12] and with program KINALC [10]. In this paper results only for the stoichiometric case are presented.

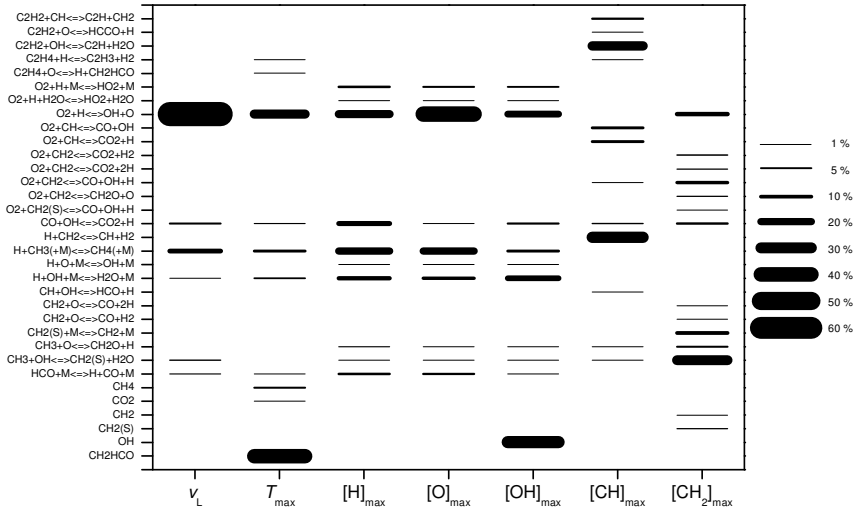


Figure 5. The percentage contributions are shown for those input parameters (rate coefficients of reactions and heat-of-formation of species), which contribute at least by 1 % to the uncertainty of at least one investigated result (see the horizontal axis). Uncertainty contributions are expressed in percentages, and the thickness of the line is proportional to the percentage value (see scale).

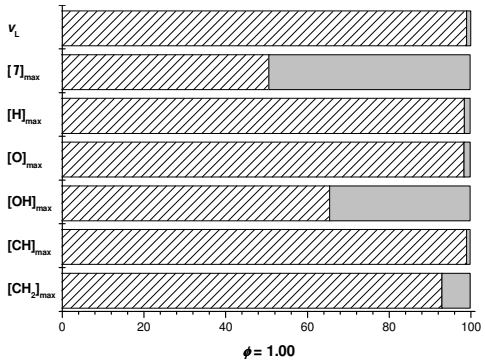


Figure 6. Kinetic (striped) and thermodynamic (grey) contributions to the total uncertainty of the monitored parameters calculated from local uncertainty analysis and expressed as percentages.

3.2. Results for the Combustion Model

3.2.1. Local Uncertainties

The results of the local uncertainty are summarised on a blob graph, show in Figure 5. It can be seen that only 33 out of the 212 parameters contribute at least with 1 % to the total uncertainty to any of the investigated parameters. Moreover, there are only few really important contributors, from which the most important is reaction of $O_2 + H = OH + O$.

In Figure 6, the contributions of the kinetic and thermodynamic parameter uncertainties are compared. For most of the investigated outputs, the effect of kinetic uncertainties is much greater than that of the thermodynamic ones. The two exceptions are the calculated flame temperature and the calculated OH concentration. Therefore, uncertainties arising from thermodynamic data cannot be neglected in a mechanism validation procedure.

3.2.2. Morris Method

Results of Morris method are in accordance with the local uncertainty analysis. The tendency that greater standard deviation belongs to greater mean effect is also observed in this case.

3.2.3. Monte Carlo Analysis

As a result of MC simulations, *pdf*s of the monitored output variables were obtained, from which a selection can be seen in Figure 7. These distributions have a high variability in both shape and width. The overall 1σ standard deviation for the laminar flame velocity is 12 %, for the temperature 0.1 %, while for the concentration of the CH radical is 46 %.

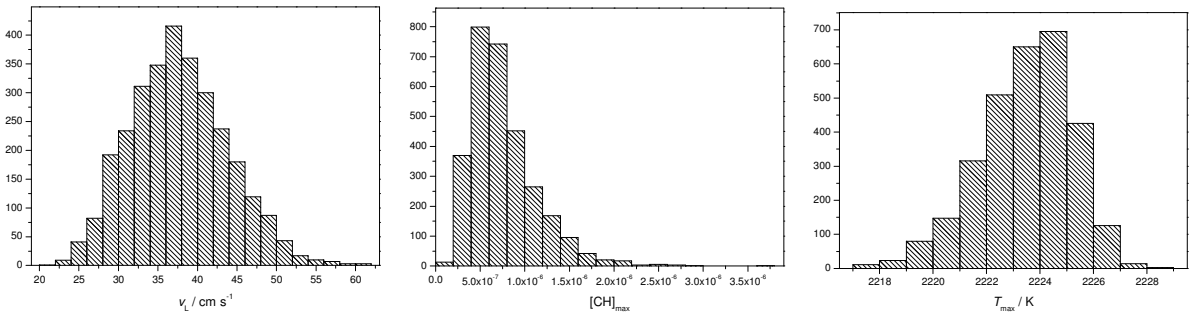


Figure 7. Distribution of some selected results in a stoichiometric methane flame, as obtained from MC simulations.

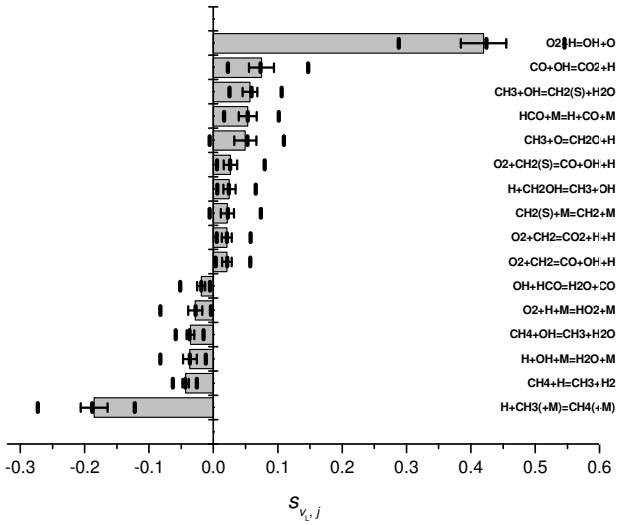


Figure 8. Result of global sensitivity analysis of the local sensitivity coefficients for the laminar flame velocity of a stoichiometric methane–air flame. Only those reactions are shown, whose rate parametric sensitivities are greater than 5% of that of the highest sensitivity one. Grey stripes refer to the local sensitivity coefficients at the nominal parameter set, small bars interconnected with a vertical line indicate the 1σ uncertainty interval of local sensitivity coefficients, and outer larger bars show the attainable minimum and maximum sensitivity coefficients at any parameter set within the uncertainty limits of parameters.

3.3. Global Uncertainty of Local Sensitivity Coefficients: A Numerical Approach

During the MC simulations, all parameters were varied simultaneously within their uncertainty limits and the first-order local sensitivity coefficients of rate parameters were calculated in each run. By processing the results, the global uncertainties of the local sensitivities were obtained. Figure 8 shows that for the laminar flame velocity sensitivities the 1σ uncertainty limits are relatively narrow. Looking at the possible extremes of the calculated local sensitivity coefficients, it can be seen that the sensitivity coefficients almost never change their sign. The small variation of the calculated sensitivity coefficients within the uncertainty range of parameters means that the rank order of importance of kinetic parameter as deduced by the local sensitivity coefficients is basically independent of the values of parameters within their range of uncertainty. This figure shows only a representative example, but very similar figures were obtained for the other variables and at other fuel-to-air ratios.

4. CONCLUSIONS

The two most significant areas of applications of large reaction mechanisms are the simulation of tropospheric chemical systems and the combustion of fuels. In these fields, the most important topics include the prediction of maximum generated ozone concentration at given conditions, and the simulation of methane flames. In this paper, we presented uncertainty analysis results for models of both types. Uncertainties of simulation results were calculated by local methods and Monte Carlo analysis, and also contribution of the various parameters to the uncertainty of the results were investigated by local sensitivity analysis and the Morris method. The surprising joint experience from the two calculations is that few parameters cause most of the uncertainties. The atmospheric chemical and the combustion models contained 141 and 212 uncertain parameters, respectively, and only about 30 parameters had noticeable

contribution to the uncertainty of any of the important results. This means that knowing better a few parameters only may significantly improve the quality of simulations in atmospheric chemistry and combustion science, which are among the most important fields of application of complex reaction mechanisms.

ACKNOWLEDGEMENTS

The authors acknowledge the helpful discussions with Prof. M. J. Pilling and Drs. V. Wagner, K. Wirtz, S. Tarantola and J. Tóth. This work was supported by OTKA T043770 and by the European Union project IALSI (contract No. EVR1-CT-2001-40013).

REFERENCES

1. A. Saltelli, E. M. Scott, and K. Chen (editors). *Sensitivity analysis*. Wiley, Chichester, 2000.
2. T. Turányi. Sensitivity analysis of complex kinetic systems - tools and applications. *J. Math. Chem.*, 5:203-248, 1990.
3. T. Turányi, L. Zalotai, S. Dóbé, and T. Bérces. Effect of the uncertainty of kinetic and thermodynamic data on methane flame simulation results. *Phys. Chem. Chem. Phys.*, 4:2568-2578, 2002.
4. R. Atkinson, D. L. Baulch, R. A. Cox, J. N. Crowley, R. F. J. Hampson, J. A. Kerr, M. J. Rossi, and J. Troe. Summary of evaluated kinetic and photochemical data for atmospheric chemistry (Web version). *Journal of Physical and Chemical Reference Data*, 2002.
5. S. P. Sander, D. M. Golden, M. J. Kurylo, G. K. Moortgat, A. R. Ravishankara, C. E. Kolb, M. J. Molina, and B. J. Finlayson-Pitts. Chemical kinetics and photochemical data for use in atmospheric studies, evaluation number 14. 2003.
6. D. L. Baulch, C. J. Cobos, R. A. Cox, P. Frank, G. D. Hayman, T. Just, J. A. Kerr, T. Murrels, M. J. Pilling, J. Troe, R. W. Walker, and J. Warnatz. Evaluated kinetic data for combustion modeling. *J. Phys. Chem. Ref. Data*, 24:1609-1630, 1995.
7. Ideal gas thermochemical database, ftp://ftp.technion.ac.il/pub/supported/aetdd/thermodynamics
8. S. M. Saunders, M. E. Jenkin, R. G. Derwent, and M. J. Pilling. Protocol for the development of the master chemical mechanism, MCM v3 (part a): Tropospheric degradation of non-aromatic volatile organic compounds. *Atmos. Chem. Phys.*, 3:161-180, 2003.
9. J. Zádor, V. Wagner, K. Wirtz, and M. J. Pilling. *Atmos. Environ.*, submitted.
10. Combustion simulations at the Leeds University and at the ELTE, http://www.chem.leeds.ac.uk/Combustion/Combustion.html, http://garfield.chem.elte.hu/Combustion/Combustion.html
11. I. Glassmann. *Combustion*. Academic, San Diego, 1996.
12. R. J. Kee, F. M. Rupley, and J. A. Miller, Chemkin-II: A FORTRAN chemical kinetics package for the analysis of gas-phase chemical kinetics, Sandia National Laboratories report No. SAND89-8009B, 1989

Techniques for Quality Assurance of Models in a Multi-Run Simulation Environment

M. Flechsig [a], U. Böhm [b], T. Nocke [c], C. Rachimow [a]

[a]Potsdam Institute for Climate Impact Research, 14412 Potsdam, Germany
[b]University of Potsdam, Institute for Physics, 14415 Potsdam, Germany
[c]University of Rostock, Institute of Computer Graphics, 18051 Rostock, Germany
E-mail: flechsig@pik-potsdam.de, boehm@pik-potsdam.de,
nocke@informatik.uni-rostock.de, rachimow@pik-potsdam.de

Abstract: In this paper we describe the multi-run simulation experiment environment SimEnv and its application in quality assurance matters for computer models. SimEnv has been developed to provide key working techniques for experimenting with complex models. This includes a wide range of simulation and model output evaluation methods in combination with corresponding visualization techniques. The SimEnv framework facilitates the easy execution of multi-run model simulation experiments for standardized, pre-formed experiment types which represent different sampling strategies of the model's input space. Further experiment types may easily be included, making SimEnv an open experimentation system. The coupling of models to the environment is supported by a simple interface, requiring only minimal model source code modifications. Uncertainty and sensitivity analyses are enabled in SimEnv by combining experiments available from the pool of pre-defined experiment types with interactive post-processing, applying sequences of related operators to both model output and reference data. Use of SimEnv as an experimental framework for models in global change research demonstrates the applicability of the approach to multi-input / multi-output problems with large amounts of spatio-temporal model output and emphasizes the importance of graphical result presentation and evaluation by appropriate visualization techniques.

Keywords: simulation environment, multi-run experiments, uncertainty analysis, sensitivity analysis

1. INTRODUCTION

Dealing with uncertainty and communicating it to decision makers and the general public is crucial in climate change research [1]. Recent papers address this issue for climate projections e.g., [2] on the basis of the findings of the Third Assessment Report of the Intergovernmental Panel of Climate Change. Identifying uncertainty in climate predictions requires comprehensive experiments for the diagnosis of the models used. The design of such models, simulation and evaluation are cornerstones in climate impact and global change research. In the past, chains of stand-alone model simulations were performed to derive from an input scenario (e.g., of greenhouse gas emission over time) of one model, outputs (e.g., climate change over time) then used as inputs to a succeeding model. The complete system can be studied and investigated this way. Nowadays, one of the challenges in global change research is the development of integrated models, which is being achieved mainly by the additional knowledge gained through feedbacks between the studied sub-systems on one hand, and through increasing computing power on the other.

Such complex simulation models are often based on legacy source code applications written in a programming language rather than in a model design language. They produce a large amount of (spatio-temporal) model output that has to be handled in the course of model validation, corroboration and/or scenario analyses. These aspects hamper the application of

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

quality assurance techniques to this kind of models, since source code is not always well known by model users and intensive code manipulations are normally beyond the scope of the work. Additionally, the computational costs for models in global change are often very high, which demands structured experimentation approaches.

2. GENERAL SIMENV APPROACH

SimEnv [3] has been developed to provide a toolbox-oriented simulation environment that enables the modeller and/or model user to deal with model-related quality assurance matters and scenario analyses for such models as described above. Both foci require flexible experiment design and model output evaluation to enable model inspection, validation / corroboration, uncertainty and sensitivity analyses without the necessity to change a complex model in general.

With respect to systems theory we consider a dynamic model **M** that can be formulated for the time dependent, time discrete, and state deterministic case - without limitation of generality - as

$$\mathbf{M}: \quad \mathbf{Z}(t) = \mathbf{ST}\ (\mathbf{Z}(t-\Delta t), \ldots, \mathbf{Z}(t-n*\Delta t), \mathbf{X}(t), \mathbf{T})$$
$$\mathbf{Y} = \mathbf{OU}\ (\mathbf{Z}(t), \mathbf{T}),$$

with

ST	state transition description
OU	output function
Z	state vector
X	input vector
T	parameter, initial value $\mathbf{Z}(t_0)$, and/or boundary value vector
t	time
Δt	time increment
n	time delay
Y	output vector

In the following, **z** and **t** are components of the vectors **Z** and **T** respectively.

The basic idea for the system design of SimEnv is to study **M** in dependence on numerical changes of a subset **t** of the parameter, initial value, and/or boundary value vector **T**:

$$\mathbf{z} = \mathbf{M}\ (\mathbf{t}),$$

where **z** is normally associated with large-scale multi-dimensional state vectors, defined over time and (geographic) space.

Simulation studies in SimEnv are supported by introducing standardized, pre-formed experiment types. An experiment type represents a multi-run simulation experiment technique with a sequence of co-ordinated single runs. According to the strategy of a selected experiment type the experiment inputs **t** (so-called targets) are sampled in the target space **{t}.** For each realization from the sample, a single simulation run of the run ensemble is performed. After setting up an experiment by equipping an experiment type with related information about the sample in {**t**} all single runs from the run ensemble are performed independently of each other. Consequently, they can be performed sequentially or in distributed mode on a cluster of networking computers using the generic Message Passing Interface MPI [4].

Preparation of a model for coupling it to SimEnv involves minimal source code manipulations for a set of supported model programming languages. Experiment-specific model output post-processing enables navigation in the combined experiment - model output space **{t**∪**z}** spanned up from the considered targets **t** and the multi-dimensional state vectors **z**. Application of built-in and user-defined post-processing operator sequences enables interactive filtering of model output and of reference data. Visualization of post-processed model output with pre-formed visualization modules forms a major component within the result evaluation component. Fig. 1 shows the general pathway for experimenting within SimEnv.

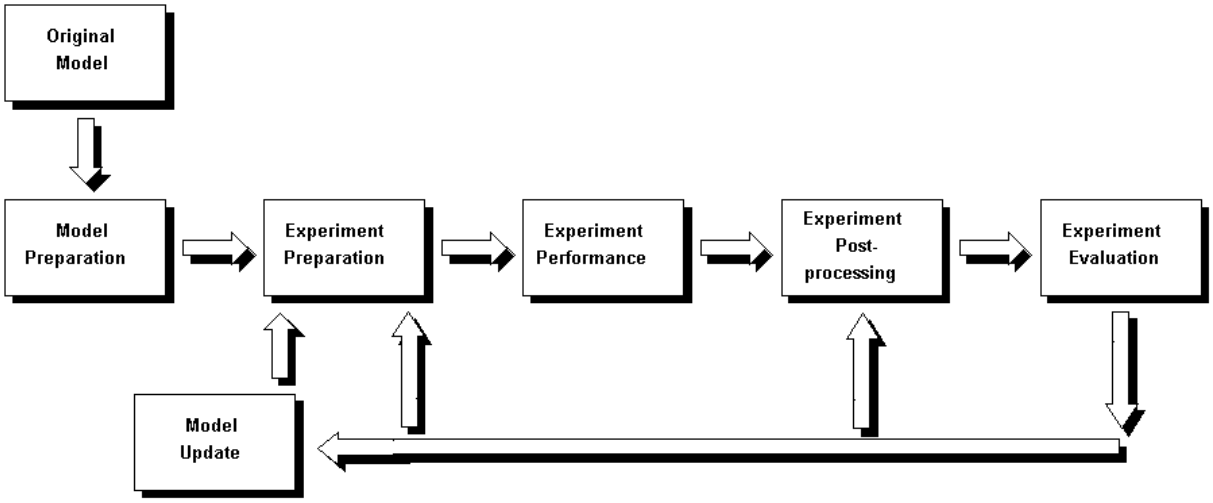


Figure 1. SimEnv System Design.

3. MODEL COUPLING

The SimEnv approach to plug in models to the simulation environment demands the availability of source code for minimal source code adaptations in order

- to map targets **t** with which the modeller wants to experiment and numerical adjustments of these from the simulation environment to the model **M,** and
- to store (n-dimensional) state variables **z** and targets **t** from **M** to SimEnv data structures for later post-processing

for each realization from the total sample on **{t}**.

The coupling interface is available for models implemented in C, Fortran, Python and in the General Algebraic Modeling System GAMS [5] for mathematical programming problems. It supports all numerical data types. Plugging the model into SimEnv requires for the model source code additional implementation of

- one function call simenv_get for each target **t** to re-adjust its value numerically according to the current single run of the experiment and
- one function call simenv_put for each model output variable **z** to store it in SimEnv output files during the current single run for later post-processing.

Additionally, at the UNIX command shell level analogous scripts are available. Among other things, they enable manipulation of model control files or forwarding re-adjusted target

values as arguments to the model before each model run without changing the model at all. SimEnv-related model output storage uses self-describing Network Common Data Form NetCDF format [6] or IEEE compliant binary format.

A model description file specifies in detail the model state variables **z** and the grid on which a state variable is defined. SimEnv supports usage of rectilinear (orthogonal with variable distance) grid definitions. Due to a flexible assignment of model variables to grids, model variables can exist on the same grid or on completely or partially disjointed grids.

4. EXPERIMENT TYPES

SimEnv aims at a well-tailored and co-ordinated simulation approach by performing run ensembles instead of single simulation runs. Co-ordination is achieved by use of pre-defined experiment types representing multi-run simulations. An experiment type scans a multi-dimensional target space **{t}** with a specific sampling strategy. Experiment types implemented so far are

- Behavioural analysis
  Deterministic inspection of the model's behaviour with a flexible sampling strategy in the target space
- Monte-Carlo analysis
  Probabilistic sampling of targets according to pre-defined distributions using different sampling methods
- Local sensitivity analysis
  Deterministic sampling in a local neighbourhood of the control scenario as the numerical nominal (default) target constellation of the model **M**.

Experiments are specified in an experiment description file by selecting an experiment type and defining the target space **{t}** and the sampling strategy.

SimEnv behavioural analysis is a generalization of the one-dimensional case, where the model behaviour is scanned in dependence on deterministic adjustments of one target **t**. The n-dimensional case demands a strategy for scanning multi-dimensional spaces in a flexible manner. On the basis of the SimEnv predecessors [7] and [8] subspaces of **{t}** can be scanned on the subspace diagonal (parallel on a one-dimensional hyperspace) or completely for all dimensions (combinatorial on a grid) and both techniques can be combined. Besides this regular sampling method an irregular, file-based technique is provided.

Fig. 2 describes the regular scanning technique by an example. In the left scheme the two-dimensional target space $\{t\} = \{p_1 , p_2\}$ is scanned in a combinatorial manner, resulting in 4*4 = 16 model runs, while the middle scheme represents a parallel scanning pattern of the two targets at the diagonal by 1+1+1+1 = 4 model runs. The scheme on the right shows a combined scanning strategy of the 3-dimensional target space $\{t\} = \{p_1 , p_2 , p_3\}$ with (1+1+1+1)*3 = 12 model runs. Each filled dot represents a single model run.

In Monte-Carlo analysis pre-defined distributions can be used to generate a sample in the target space. Random and Latin hypercube sampling [9] is supported for uniform, normal, log-normal and exponential distributions. Currently, SimEnv only supports sampling of uncorrelated targets; as a workaround, there is an interface to import external samples.

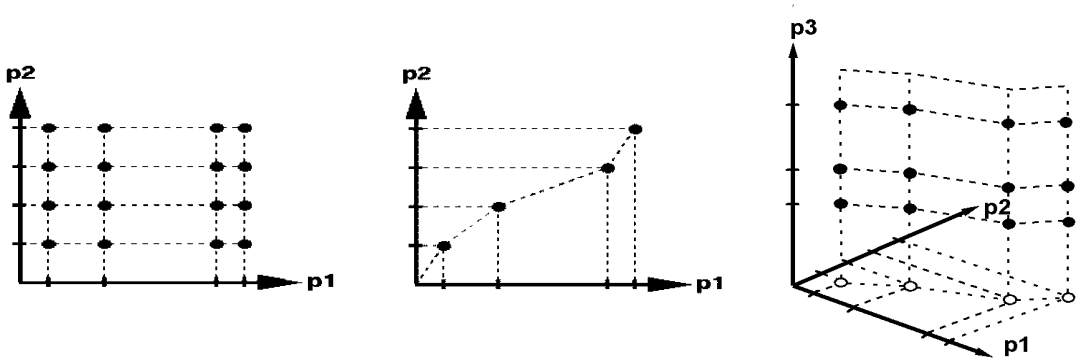


Figure 2. Behavioural analysis: Deterministic sampling of multi-dimensional target spaces.

For local sensitivity analysis the experiment is set up by single model runs in ε-neighbourhoods of the control scenario in the target space **{t}**. For each target $\mathbf{t}_i$ from the control scenario $\mathbf{t} = (t_1, \ldots, t_n)$ and each ε_j from $\varepsilon_j = (\varepsilon_1, \ldots, \varepsilon_m)$ two runs are performed for the both target constellations $(t_1, \ldots, t_{i-1}, t_i \pm \varepsilon_j, t_{i+1}, \ldots, t_n)$.

5. EXPERIMENT POST-PROCESSING AND VISUAL EVALUATION

Interactive post-processing is applied to compute output functions **y** from the model's outputs **z** by state space transformation operators and to derive uncertainty and sensitivity measures from these output functions by experiment type-specific operators. For this purpose, the SimEnv post-processor enables application of operator sequences to both model output and reference data. Currently, about 100 built-in operators are available. An interface enables users easily to declare their own operators and plug them into the environment. Each operator assigns to its result output a unique grid definition, derived from the operator definition and the grids of its operands. SimEnv post-processor output can be stored in NetCDF, IEEE compliant binary or ASCII format.

State space transformation operators cover elemental, selective, analytical, and statistical techniques, among others. The main focus is reduction of and aggregation in the output model state space to cope with its potentially high dimensionality and extent. Selective operators provide methods to access to a selected single run, to external data and other SimEnv experiments and to clip the extents or to reduce the dimensionality of an operand on its assigned grid. Statistical operators supply basic statistical information from operands on the whole grid or on grid layers for single grid dimensions.

Analysis and evaluation of post-processed data derived from large amounts of relevant model output benefit from visualization techniques. Based on metadata information about the post-processed experiment type, the applied operator sequence, and the dimensionality of the post-processor output, pre-formed visualization modules are evaluated by a suitability coefficient to determine how they can map post-processor output in an appropriate manner.

The visualization modules offer a high degree of user support and interactivity to cope with multi-dimensional data structures. Among others, they cover standard techniques such as scatter and parallel coordinate plots (the latter for abstract data visualization), and isolines, isosurfaces, direct volume rendering and 3D difference visualization techniques. Furthermore, approaches to navigate intuitively through large multi-dimensional data sets have been applied, including details on demand, interactive filtering and animation [10]. Using the open source visualization platform OpenDX [11] based on IBM's Data Explorer, extended

OpenDX techniques have been designed and implemented, suited to the context of analysis and evaluation of simulated multi-run output functions.

6. UNCERTAINTY AND SENSITIVITY ANALYSES

The key methodological approach for uncertainty and sensitivity analyses in SimEnv is the combination of experiments from the set of pre-defined experiment types with interactive exploration of the model output variables' set from the run ensemble in experiment post-processing, applying sequences of experiment-specific operators to both state space model output functions and reference data. Derived from the general experiment layout, SimEnv experiment types are associated with uncertainty and sensitivity analyses techniques in the following way:

- Behavioural analysis
  Can be used for uncertainty analysis, factorial screening, general one-factor-at-a-time approach, (fractional) factorial experiments and response surface methodology. All methods benefit from the flexible screening strategy of multi-dimensional target spaces in SimEnv.
- Monte-Carlo analysis
  Can be used for uncertainty analysis and global sensitivity analysis.
- Local sensitivity analysis
  Can be used for local first order sensitivity analysis by investigating finite difference approximations of derivatives.

During post-processing uncertainty and/or sensitivity measures are provided by experiment-specific operators. A general behavioural analysis operator enables the modeller/user to navigate in the target space **{t}** and to derive aggregations and moments in its sub-spaces in a flexible manner. Monte-Carlo analysis operators support (among other things) computation of extremes, moments, quantiles and heuristic probability density functions from targets and output functions as well as regression, correlation, and covariance measures from targets, model output, or both of these together. For local sensitivity analysis a set of sensitivity operators (linear, squared, absolute, relative, symmetric) are available as finite approximations of the classical local sensitivity measure $\partial\mathbf{z}/\partial\mathbf{t}$.

7. EXAMPLE

We show from an ongoing study sensitivity results for CLM, a regional meteorological model CLM [12] in climate mode [13] where parameters controlling both the dynamic forecast part and the parametrization part for subgrid-scale diabatic source and sink processes in their relation to diagnostic and prognostic model output variables have been under investigation.

CLM is used with a horizontal resolution of 0.5° x 0.5° latitude/longitude and with 20 layers in the vertical for a region covering the Baltic Sea and most of Northern and Central Europe. The model time step is 90 seconds and output is stored every six hours. The model is based on the non-hydrostatic, fully compressible primitive equations of the atmospheric motion without scale approximations. The model uses a generalized terrain-following vertical coordinate and rotated geographical coordinates. It is subdivided into a so-called dynamic part, where the basic equations, spatially discretized by use of second-order finite differences, are solved for the prognostic variables wind velocity in x- and y- direction of the orthogonal z-system, perturbation pressure, to the hydrostatic basic stage, temperature, specific humidity,

cloud water content and (optionally) cloud ice content. Sub-grid scale source- and sink-processes have to be parametrized and are computed before the dynamic part. Among others, also soil hydrological and thermal processes are described by such a parametrization.

In our investigations, we consider the hydrological section of the soil parametrization in CLM. One of the components of the near-surface water balance is transpiration by plants from two soil layers with a depth of 10 cm and 90 cm. This process is described by a Biosphere-Atmosphere Transfer Scheme [14]. The basic idea is to apply a resistance concept as in electricity to compute plant transpiration affected by atmospheric and stomatal factors. One of the used transpiration reduction factors accounts for the reduction of transpiration by the stomatal resistance r_s.

$$\frac{1}{r_s} = \frac{1}{crs_{max}} + \left(\frac{1}{crs_{min}} - \frac{1}{crs_{max}} \right) \cdot F_{rad} \cdot F_{wat} \cdot F_{temp} \cdot F_{hum}$$

r_s is described by the two parameters crs_{min} and crs_{max} and various influence functions F. For $crs_{min} = crs_{max}$ transpiration is not reduced by any of the influence functions. In the function

$$F_{temp} = \max\left[0 \, , \, \min\left\{ 1 \, , \, 4 \cdot \frac{(T_s - T_0) \cdot (T_{end} - T_s)}{(T_{end} - T_0)^2} \right\} \right]$$

for the influence of the surface temperature T_s the empirical constant T_{end} describes optimal conditions for plant transpiration. F_{temp} reaches its maximum for $T_s \approx T_{end}/2$.

We apply a behavioural analysis to assess the effect of the empirical parameters crs_{min} and T_{end} on latent and sensible heat fluxes lhf and shf from soil in a deterministic manner for crs_{max} = 1000 s/m. Both fluxes are defined on a grid spanned up from latitude, longitude and time. In Box 1 the experiment description file to scan the 2-dimensional parameter space $\{crs_{min} \, , \, T_{end}\}$ combinatorially is shown. Additionally, in the model source code crs_{min} and T_{end} have to be re-adjusted by a simenv_get-call for each of both parameters.

```
target    crsmin   adjusts   30.(5.)120.          # specifies 19 adjustments for crs_min
target    crsmin   default    60.                 # default model value of crs_min
target    crsmin   type       set                 # do not modify adjustments by default
target    Tend     adjusts   273.15(5.)333.15     # specifies 13 adjustments for T_end
target    Tend     default    313.15
target    Tend     type       set
specific           comb      crsmin*tend         # factorial screening: 19*13+1=248 runs
```

Box 1. Experiment description file for a behavioural analysis.

Post-processed results for a simulated period of seven days are shown in Fig.3. The influence of the variation of crs_{min} and T_{end} on lhf and shf anomalies from the model nominal constellation is shown on the left. To produce during SimEnv post-processing this result from model output the applied operator sequences are

behav(‘ ‘, avg(shf)) - run(‘default’, avg(shf)) and
behav(‘ ‘, avg(lhf)) - run(‘default’, avg(lhf)),

where behav is the general behavioural operator to navigate in the experiment space, the operator run addresses one single run from the whole run ensemble, and the operator avg supplies the total average from a multi-dimensional model output variable. To get area-averaged flux anomalies for each time step time dependent on crs_{min} and for the default value of T_{end} we have to apply in post-processing

behav('sel_t(Tend=313.15)', avg_l('time', shf)) - run('default', avg_l('time', shf))

where avg_l supplies area averages for each time step. Fig. 3 on the right is the corresponding graphical representation.

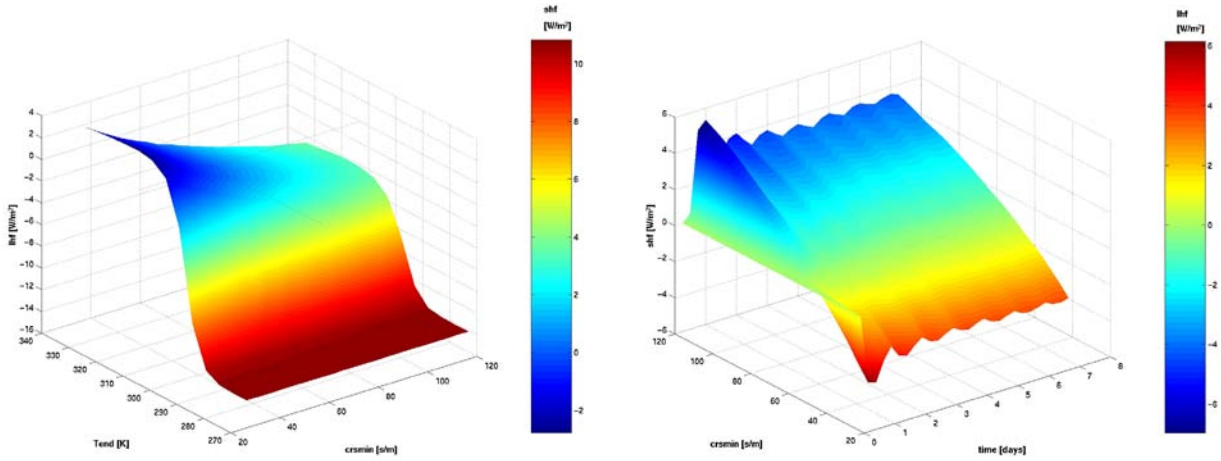


Figure 3. Surface heat flux anomalies from soil. Dynamic was compiled in SimEnv post-processing from the results of the 248 single runs of the experiment.
Left: Area and temporal mean dependent on T_{end} and crs_{min}.
Right: Area mean for each time step dependent on crs_{min}.

First it becomes visible from the right panel of Fig. 3 that both heat fluxes behave inversely to the changes in crs_{min}. As to be expected, the latent heat flux lhf decreases with increased resistance values, whereas the sensible heat flux shf increases to transport heat back from the surface to the atmosphere in this case and to maintain the surface energy balance. Secondly, the reaction of both heat fluxes is rather linear for the entire parameter space. Together with changes in T_{end}, however, the behaviour of the heat fluxes is significantly different: As shown on the left, only for rather high values of T_{end} the heat fluxes change with crs_{min} as for the default of T_{end} on the right. For T_{end} below about 273.15 K, this parameter dominates the reaction of the sensible and latent heat fluxes and nearly no modifications in the results due to crs_{min} can be identified.

The results of a Monte-Carlo study on T_{end} and crs_{min} for a simulated period of seven days are shown in Fig. 4. Both parameters are drawn from a normal distribution with a Latin hypercube sampling technique where the mean is the nominal parameter value and variance is set to 20. Sample size is 150 runs. For the left panel of Fig. 4 the applied operator sequence is

hgr_e(15, avg_l('time', shf) – run('default', avg_l('time', shf))),

where the operator hgr_e supplies for each element of its second argument a heuristic probability density function over the whole run ensemble with 15 bins.

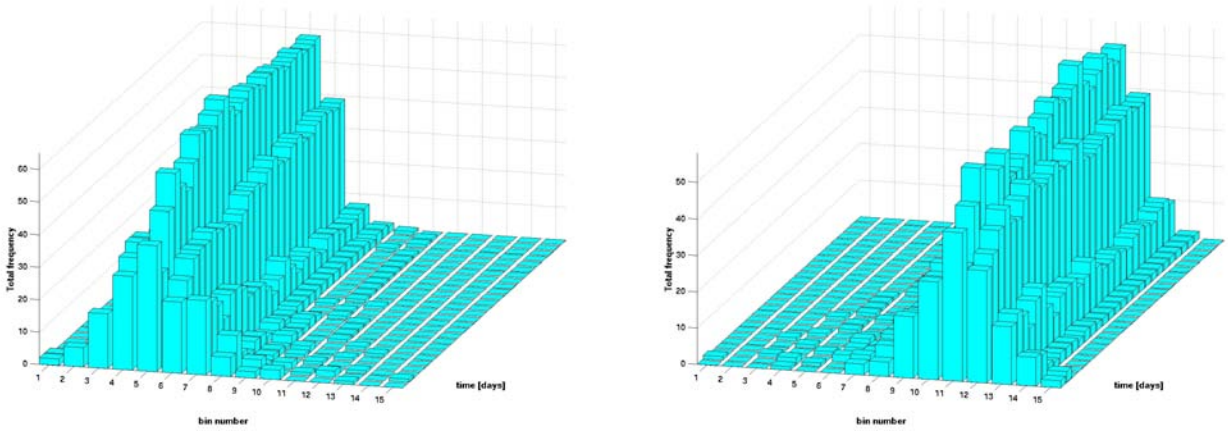


Figure 4. Probability density functions of surface heat flux anomalies from soil for time steps 2 - 28. Left: lhf, right: shf. lhf anomaly values range from -5.02 (bin # 1) to 2.03 (bin # 15), shf anomaly values from -1.51 to 3.87.

8. RESULTS AND CONCLUSIONS

The methodology presented here and its implementation have been proven to support the process of model evaluation from various perspectives of both model developers and users. In contrast to other simulation environments (e.g., SimLab [15], SCIRun [16], and Pingo [17]) that also focus on uncertainty and sensitivity matters, with SimEnv we try to support all steps in experimenting with models from easy-to-use model coupling to the system via experiment design, experiment load distribution, and model output post-processing to visual evaluation. The supported languages cover most of the model sources codes used in global change research. The concept of pre-defined experiment types seems to be an appropriate way to guide model developers and/or users in the process of experimenting with models and frees them from expensive workload. The plug-in interface for user-defined operators opens the post-processor to permit coupling to special-purpose applications or libraries on user demand. Additionally, output formats from the post-processor can be used to export model results to other applications, e.g. as statistical diagnosis and analysis tools, for in-depth investigations of specific research goals. One of the strengths of SimEnv is its support of multi-dimensional model output data on rectilinear grids in a persistent manner for model plug-in, post-processing, and visualization.

On the other hand, this holistic approach is at the same time one of the weaknesses of SimEnv. With SimEnv, we provide a general simulation environment for a broad spectrum of tasks without supporting special features in detail. For example, sampling strategies and built-in operators especially for uncertainty and sensitivity analyses techniques are limited.

9. PROSPECTS

The following work packages are planned for further development of SimEnv:

- Special-purpose sampling designs: Support of special uncertainty and sensitivity experiments, e.g., the Fourier amplitude sensitivity test FAST and/or the method of Sobol [18] and implementation of corresponding post-processing operators and visualization techniques.
- Simulation-based optimization: Application of gradient-free methods for (mono- and) multi-criterial optimization of cost functions $f_i(z)$ in the target space **{t}.**

- Support of distributed models across computer networks or the Internet: Setting up a SimEnv experiment server to handle target dissemination and model output collection.

ACKNOWLEDGEMENT

The authors thank Martin Kücken and Detlef Hauffe for performing the corresponding experiments with the CLM model.

REFERENCES

1. M. Webster. Communicating Climate Change Uncertainty to Policy-Makers and the Public. *Climatic Change*, 61:1-8, 2003.
2. M. Webster, C. Forst, J. Reilly, M. Babiker, D. Kickligther, M. Mayer, R. Prinn, M. Sarofim, A. Sokolov, P. Stone, and C. Wang. Uncertainty Analysis of Climate Change and Policy Response. *Climatic Change*, 61:295-320, 2003.
3. M. Flechsig, U. Böhm, T. Nocke, and C. Rachimow. *The Multi-Run Simulation Environment SimEnv: User's Guide.* Potsdam Institute for Climate Impact Research, Potsdam. http://www.pik-potsdam.de/topik/pikuliar/simenv/home/simenv.pdf, 2004.
4. MPI. http://mpi-forum.org
5. GAMS: http://www.gams.com
6. NetCDF. http://www.unidata.ucar.edu/packages/netcdf/
7. V. Wenzel, E. Matthäus, and M. Flechsig. One Decade of SONCHES. *Syst. Anal. Mod. & Sim.*, 7:411-428, 1990.
8. M. Flechsig: *SPRINT-S: A Parallelization Tool for Experiments with Simulation Models.* PIK-Report No. 47. Potsdam Institute for Climate Impact Research, Potsdam, 1998 http://www.pik-potsdam.de/reports/pr-47/pr47.pdf
9. R.L. Imam and J.C. Helton. An Investigation of Uncertainty and Sensitivity Analysis Techniques for Computer Models. *Risk Anal.*, 8(1):71-90, 1998.
10. T. Nocke, H. Schumann, U. Böhm, and M. Flechsig. Information Visualization Supporting Modelling and Evaluation Tasks for Climate Models. In: S. Chick, P.J. Sanchez, D. Ferrin, D.J. Morrice, editors, *Proceedings of the 2003 Winter Simulation Conference*, December 2003, New Orleans, 2003.
11. OpenDX. http://www.opendx.org
12. G. Doms and U. Schättler. *The Nonhydrostatic Limited-Area Model LM of DWD – Part I: Scientific Documentation.* Technical Report, German Meteorological Office DWD, Offenbach/M., 1997.
13. M. Kücken and D. Hauffe. *The Nonhydrostatic Limited-Area Model LM of DWD with PIK Extensions. Part III: Extensions User Guide.*2003. http://w3.gkss.de/CLM/clm_home.html
14. R. Dickinson. Modeling Evapotranspiration for Three-dimensional Global Climate Models. In *Climate Processes and Climate Sensitivity.* Geophysical Monograph 29, Maurice Ewing Volume 5, American Geophysical Union, Washington, D.C., 58-72, 1984.
15. SimLab. *Software for Uncertainty and Sensitivity Analysis. POLIS-JRC-ISIS* http://sensitivity-analysis.jrc.cec.eu.int/default2.asp?page=forum
16. SCIRun Scientific Computing and Imaging http://software/sci/utah.edu
17. J. Waszkewitz, P. Lenzen, and N. Gillet (2001) *The PINGO Package* http://www.mad.zmaw.de/Pingo/pingohome.html, 2001.
18. K. Chan, K., S. Tarantola, A. Saltelli, and I.M. Sobol. Variance-Based Methods. In: Saltelli, A., Chan, K., Scott, E.M., editors, *Sensitivity Analysis.* J. Wiley & Sons, Chichester, 2000.

GLOBAL SENSITIVITY ANALYSIS FOR CALCULATING THE CONTRIBUTION OF GENETIC PARAMETERS TO THE VARIANCE OF CROP MODEL PREDICTION

David Makowski[a], Cédric Naud[b], Hervé Monod[b], Marie-Hélène Jeuffroy[a], Aude Barbottin[a]

[a] UMR Agronomie INRA/INA-PG, B.P. 01 78850 Thiverval-Grignon, France.
[b] Unité de biométrie INRA, Domaine de Vilvert, F78352 Jouy-en-Josas Cedex, France.
Email: makowski@grignon.inra.fr, monod@banian.jouy.inra.fr

Abstract: Dynamic models are often used to predict the effects of farmers' practices on crop yield, crop quality and environment. These models usually include many parameters that must be estimated from experimental data before practical use. Some of the parameters may vary across genotypes. Such genetic parameters may be estimated from plant breeding experiments but this is very costly and requires a lot of experimental work. Moreover, some of the genetic parameters may account for only a very small part of the output variance and, so, do not deserve an accurate determination. This papers shows how methods of global sensitivity analysis can be used to evaluate the contributions of the genetic parameters to the variance of model prediction. Two methods are applied to a complex nonlinear dynamic crop model for estimating the sensitivity indices associated to 13 genetic parameters. The results show that only 5 genetic parameters have a significant effect on yield and grain quality.

Keywords: Crop model, fast, genetic parameter, global sensitivity analysis, winding stairs

1. INTRODUCTION

Crop models are complex nonlinear dynamic models simulating several output variables related to crop yield, crop quality, farmer's income and environment. These models are valuable tools for crop management because they can be used to predict the effects of farmers' practices in function of soil type, climate, and crop characteristics.

Crop models can include up to 200 parameters whose values must be estimated from past experiments. The estimation of these parameters is an important problem because crop model performances depend for a large part on the accuracy of the parameter estimates. Results obtained with crop models are not reliable when inaccurate parameter values are used. A large amount of data is always required for estimating accurately crop model parameters, in particular when the model includes genetic parameters. As genetic parameters vary across genotypes, the estimation of these parameters must be based on specific measurements collected for each genotype. Such measurements can be performed in plant breeding experiments but this is very costly and requires a lot of experimental work. Moreover, recent studies have shown that crop model predictions are not always improved when genotypic parameters are estimated genotype per genotype. This may be due to the small contribution of some of the genetic parameters to the total model output variance.

In this study, we investigate how methods of sensitivity analysis can be used to reduce the quantity of field experiments performed for estimating genetic parameters. The basic principle consists in evaluating the contributions of the genetic parameters to the variance of the model prediction and in estimating genotype per genotype only the key

parameters whose uncertainty affects most the outputs. This approach is illustrated below with the AZODYN crop model [1] developed for simulating winter wheat crops. Two methods of global sensitivity analysis are applied to this model in order to evaluate the contribution of 13 genetic parameters to the variances of several output variables of agronomic interests.

2. METHODOLOGY

2.1. The AZODYN model

The AZODYN crop model [1] is a nonlinear dynamic model simulating winter wheat crop in function of numerous input variables describing the characteristics of the crop at the end-of-winter (initial biomass and nitrogen content), soil characteristics (soil texture, organic matter, soil mineral nitrogen), climate (daily radiation and temperature), and nitrogen fertilization (dates and rates of fertilizer applications). In this paper, the input variable are set equal to values obtained in a field located in the north of France (Grignon) and harvested in 2001.

AZODYN is a useful tool for studying the effects of nitrogen management on crop yield, grain quality and risk of pollution by nitrate [2]. The model includes several state variables that are simulated at a daily time step. One of the state variable, namely the nitrogen nutrition index (*NNI*), indicates if the nitrogen content of the crop is at its optimal level or not. AZODYN can be used to predict the characteristics of the crop at harvest, notably grain yield and grain protein content at harvest. Grain yield is an important variable because it determines the farmer's income. Grain protein content is a major grain quality criterion for agro-industries.

Table 1. The 13 genetic parameters of AZODYN and their ranges of variation.

Parameter	Definition	Range	Unit
RDTMAXVAR	Maximal yield	100-137	$q.ha^{-1}$
Ebmax	Radiation use efficiency	2.7-3.3	$g.MJ^{-1}$
D	Ratio of leaf area index to critical nitrogen	0.02-0.045	-
REM2	Fraction of remobilized nitrogen	0.5-0.9	-
K	Extinction coefficient	0.6-0.8	-
Eimax	Ratio of intercepted to incident radiation	0.9-0.99	
Tep.flo	Duration between earing and flowering	100-200	°C.day
R	Ratio of total to above ground nitrogen	1.0-1.5	-
P1GMAXVAR	Maximal weight of one grain	47-65	Mg
Lambda	Parameter for calculating nitrogen use efficiency	25-45	-
Mu	Parameter for calculating nitrogen use efficiency	0.6-0.9	-
DJPF	Temperature threshold	150-250	°C.day
NGM2MAXVAR	Maximal grain number	107.95-146.05	-

The AZODYN crop model includes 69 parameters whose values must be estimated before practical use. Among all the parameters, 13 parameters were found to vary across genotypes in past studies. Their values are not perfectly known. The genetic parameters and their ranges of variation are described in table 1. The purpose of this paper is to evaluate the contributions of the 13 genetic parameters to the variances of three model outputs, namely yield, grain protein content and nitrogen nutrition index (*NNI*).

2.2. Comparing different sensitivity analysis methods on AZODYN

2.2.1. Sensitivity indices

We note further Y the output variables of AZODYN. Y will represent in turn yield, grain protein content, and the daily values of *NNI*. Yield and grain protein content are calculated only at harvest, whereas *NNI* is calculated each day between the end-of-winter and harvest. Our approach consists in partitioning the total variance of $V(Y)$ as follows [3]:

$$V(Y) = \sum_{i=1}^{13} V_i + \sum_{i<j} V_{ij} + \sum_{i<j<m} V_{ijm} + V_{1,2...,13}, \quad (1)$$

where $V(Y)$ is the total variance of the output variable Y induced by the 13 genetic parameters, $V_i = V\left[E\left(Y|x_i\right)\right]$ measures the main effect of the parameter x_i, i=1, …, 13, and the other terms measure the interaction effects. The decomposition (1) is used to derive two types of sensitivity indices defined by

$$S_i = \frac{V_i}{V(Y)}, \quad (2)$$

$$S_{Ti} = \frac{V(Y) - V_{-i}}{V(Y)}, \quad (3)$$

where V_{-i} is the sum of all the variance terms that do not include the index i. S_i is the *first-order sensitivity index* for the i^{th} parameter. This index measures the main effect of parameter x_i on the output variable Y. S_{Ti} is the *total sensitivity index* for the i^{th} parameter and is the sum of all effects (first- and higher-order) involving the parameter x_i. S_{Ti} takes into account the interactions between the i^{th} parameter and the other 12 parameters. The total sensitivity index can be though as the expected fraction of variance that would be left if only the parameter x_i were to stay undetermined. S_i and S_{Ti} are both in the range (0, 1). The sensitivity indices S_i and S_{Ti} do not differ much from zero when the parameter x_i has a small effect on the output variable Y. On the contrary, if the parameter i has a strong effect on Y, the indices take values near from one. The two sensitivity indices S_i and S_{Ti} are equal if the effect of the i^{th} parameter on the model output is independent from the effects of the other parameters.

In the next two sections, we present two methods for estimating the indices (2) and (3) for each parameter and each output variable.

2.2.2. Winding stairs

The calculation of the indices (2) and (3) requires the knowledge of $V(Y)$, V_i, and V_{-i}. The computations can be performed by using a Monte Carlo method [3, 4]. The principle is to generate randomly samples of parameters and to estimate $V(Y)$, V_i, and V_{-i} as follows:

$$\hat{V}(Y) = \frac{1}{N}\sum_{m=1}^{N}\left[f(X_m)\right]^2 - \hat{f}_0^2, \tag{4}$$

where $X_m = \left(x_{1m},...,x_{im},...,x_{13m}\right)$ is a sample of 13 parameter values drawn in the ranges of variation displayed in table 1, $f(X_m)$ is the simulated value of the output variable, and $\hat{f}_0 = \frac{1}{N}\sum_{m=1}^{N} f(X_m)$.

$$\hat{V}_i = \frac{1}{N}\sum_{m=1}^{N}\left[f\left(X^{(1)}_{(-i)m}, x_{im}\right) f\left(X^{(2)}_{(-i)m}, x_{im}\right)\right] - \hat{f}_0^2, \tag{5}$$

where x_{im} is the m^{th} value of the i^{th} parameter, and $X^{(1)}_{(-i)m}$ and $X^{(2)}_{(-i)m}$ are two different vectors including values of the 12 other parameters defined by $X^{(1)}_{(-i)m} = \left(x^{(1)}_{1m},...,x^{(1)}_{(i-1)m},x^{(1)}_{(i+1)m}...,x^{(1)}_{13m}\right)$ and $X^{(2)}_{(-i)m} = \left(x^{(2)}_{1m},...,x^{(2)}_{(i-1)m},x^{(2)}_{(i+1)m}...,x^{(2)}_{13m}\right)$.

$$\hat{V}_{-i} = \frac{1}{N}\sum_{m=1}^{N}\left[f\left(X_{(-i)m}, x^{(1)}_{im}\right) f\left(X_{(-i)m}, x^{(2)}_{im}\right)\right] - \hat{f}_0^2 \tag{6}$$

where $x^{(1)}_{im}$ and $x^{(2)}_{im}$ are two different values of the i^{th} parameter and $X_{(-i)m}$ is a vector including the values of the 12 other parameters.

Different sampling methods can be used to generate the parameter values and organize the computations. Here, we apply the winding stairs sampling scheme [5]. This method was designed to make multiple use of model evaluations. With a single series of *N* model evaluations, it can compute both the first-order and the total sensitivity indices. The winding stairs method consists in computing the model outputs after each drawing of a new value for an individual parameter. Various procedures can be used to generate the parameter values. Here, the parameter values are generated by Latin hypercube sampling.

The sequence of model outputs generated by the Winding stairs method is shown is table 2. The model outputs are grouped by pairs and are used to compute all the sensitivity indices. For example, the variance (6) is estimated for the first parameter by using the following pairs of model outputs: {1, 2}, {13+1, 13+2}, …, {(*N*-1)*13+1, (*N-1)*13+2}. The variance (5) is estimated for the first parameter by using {2, 13+1}, {13+2, 2*13+1}, …{(*N*-2)*13+2, (*N*-1)*13+1}.

The number of model evaluations required for calculating the indices associated to the 13 genetic parameters is equal to 13*N* (table 2). The winding stairs sampling scheme is applied here with two values of *N*, specifically *N*=5000 and *N*=10000. The Winding stairs method is run 15 times in order to obtain 15 estimates of the first-order and total sensitivity indices for each *N* value and each parameter. The 15 estimates are averaged and their accuracy is evaluated by computing standard deviation and 95% confidence interval.

Table 2. Sequence of model outputs generated by the Winding stairs method.

Simulation number	Output
1	$f\left(x_{1,1}, x_{2,1}, x_{3,1}, ..., x_{13,1}\right)$
2	$f\left(x_{1,2}, x_{2,1}, x_{3,1}, ..., x_{13,1}\right)$
3	$f\left(x_{1,2}, x_{2,2}, x_{3,1}, ..., x_{13,1}\right)$
4	$f\left(x_{1,2}, x_{2,2}, x_{3,2}, ..., x_{13,1}\right)$
…	
13+1	$f\left(x_{1,2}, x_{2,2}, x_{3,2}, ..., x_{13,2}\right)$
13+2	$f\left(x_{1,3}, x_{2,2}, x_{3,2}, ..., x_{13,2}\right)$
…	
(N-1)*13+1	$f\left(x_{1,N}, x_{2,N}, x_{3,N}, ..., x_{13,N}\right)$
(N-1)*13+2	$f\left(x_{1,1}, x_{2,N}, x_{3,N}, ..., x_{13,N}\right)$
(N-1)*13+3	$f\left(x_{1,1}, x_{2,1}, x_{3,N}, ..., x_{13,N}\right)$
…	
N*13	$f\left(x_{1,1}, x_{2,1}, x_{2,1}, ..., x_{12,1}, x_{13,N}\right)$

2.2.3. Extended FAST

The sensitivity indices are estimated by using a second method named extended FAST [3, 6]. In extended FAST, the sensitivity indices are evaluated by a search curve that scans the space of the 13 parameters, in such a way that each parameter is explored with a selected integer frequency. The basic idea of the method is to convert the 13-dimensional integral in the parameters into a one-dimensional integral by using the transformation function G_i for i=1, …, 13 defined by

$$x_i = G_i\left(\sin \omega_i s\right) \tag{7}$$

where $s \in \left]-\pi, \pi\right[$ and $\left\{\omega_i; i = 1, ..., 13\right\}$ is a set of integer angular frequencies. The function (7) allows each parameter to be explored globally across its range of variation, as the parameter s is varied over $\left]-\pi, \pi\right[$. The implementation of the FAST and extended FAST methods is described in detail in [6].

The method is applied here by using the transformation function $x_i = \frac{1}{2} + \frac{1}{\pi}\arcsin\left[\sin\left(\omega_i s + \varphi_i\right)\right]$ where φ_i is a random phase-shift parameter drawn in

$[0, 2\pi]$. The frequencies $\{\omega_i; i = 1,...,13\}$ are chosen according to the method described by Saltelli [6]. The method requires the computation of several integrals by using a set of model simulations. Here, the indices are calculated from 5000 simulations for each parameter. So, the total number of model simulations is equal to 13*5000. The extended FAST method is run 15 times with different values for the shift parameters in order to derive 15 different estimates of first-order and total sensitivity indices. Like with the Winding stairs method, the 15 estimates are averaged and their accuracy is evaluated by computing standard deviation and 95% confidence interval.

3. RESULT

3.1. Sensitivity indices for yield and grain protein content

Figure 1 shows the total sensitivity indices calculated for the 13 genetic parameters with the Winding stairs method and *N*=10000. For yield, the parameter *RDTMAXVAR* has the highest total sensitivity index. Its value is equal to 0.77. It means that about 77% of yield variance would be left if only the parameter *RDTMAXVAR* were to stay undetermined. The strong influence of *RDTMAXVAR* is logical because this parameter determined the maximal yield values in the model equations. Two other parameters have a significant influence on yield, namely *Ebmax* and *D*. The total sensitivity indices calculated with the Winding stairs method are equal to 0.18 and 0.17 for these two parameters. The indices of the 10 other parameters are lower than 0.07.

Figure 1 shows that, for grain protein, the parameter with the highest total sensitivity index is *REM2* (index=0.49). This parameter is used by the model to calculate the fraction of the total plant nitrogen that can be allocated to the grains after flowering, and this fraction determines for an important part the value of the grain protein content at harvest. Two other parameters have a total sensitivity index higher than 0.1, namely *R* and *RDTMAXVAR*.

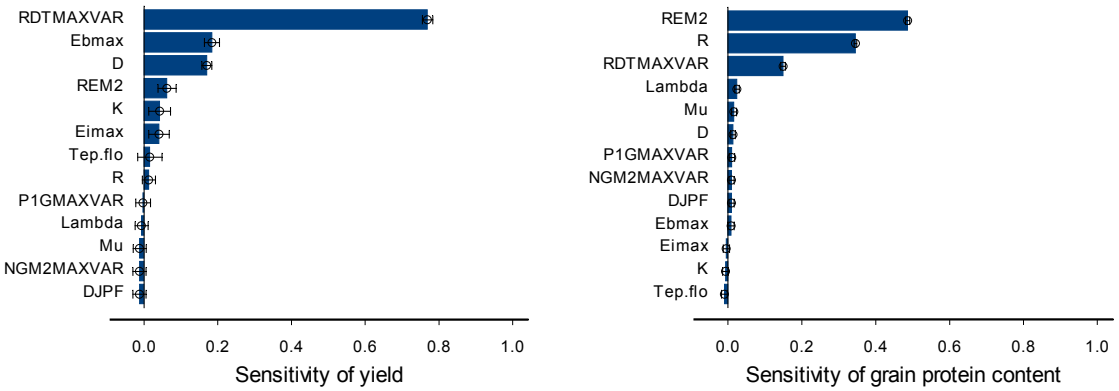


Figure 1. Total sensitivity indices for the output variables yield and grain protein content obtained with the methods Winding stairs (*N*=10000). The horizontal bars show the average values of the sensitivity indices calculated from 15 estimates for the 13 genetic parameters. Error bars indicate the 95% confidence interval.

Table 3. Sensitivity indices (first order and total effect) for yield and for the 13 genotypic parameters. Mean and standard deviation are calculated from 15 estimates.

	Winding stairs *N*=10000				Winding stairs *N*=5000				Extended Fast			
	Total sensitivity		First order sensitivity		Total sensitivity		First order sensitivity		Total sensitivity		First order sensitivity	
Parameter	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
RDTMAXVAR	0.769	*0.033*	0.622	*0.036*	0.753	*0.049*	0.627	*0.052*	0.766	*1.2E-02*	0.627	*8.6E-03*
Ebmax	0.184	*0.048*	0.061	*0.045*	0.092	*0.095*	0.155	*0.090*	0.179	*2.6E-03*	0.074	*1.5E-03*
D	0.170	*0.032*	0.055	*0.029*	0.161	*0.053*	0.071	*0.047*	0.159	*8.8E-03*	0.067	*2.8E-03*
REM2	0.062	*0.059*	-0.010	*0.057*	-0.032	*0.122*	0.079	*0.119*	0.043	*1.3E-03*	0.014	*4.7E-04*
K	0.042	*0.069*	-0.010	*0.070*	-0.007	*0.103*	0.050	*0.090*	0.036	*1.5E-03*	0.008	*4.7E-04*
Eimax	0.041	*0.066*	0.010	*0.066*	-0.015	*0.110*	0.075	*0.097*	0.046	*1.5E-03*	0.015	*6.8E-04*
Tep.flo	0.016	*0.078*	-0.015	*0.081*	-0.039	*0.108*	0.050	*0.098*	8.25E-03	*7.3E-04*	1.08E-03	*1.1E-04*
R	0.013	*0.042*	-0.003	*0.039*	0.064	*0.054*	-0.046	*0.052*	1.35E-02	*3.0E-03*	7.14E-04	*1.1E-04*
P1GMAXVAR	-0.003	*0.047*	0.007	*0.041*	0.055	*0.070*	-0.042	*0.070*	1.05E-02	*3.9E-03*	4.17E-04	*1.6E-04*
Lambda	-0.007	*0.042*	0.009	*0.038*	0.057	*0.067*	-0.045	*0.066*	9.58E-03	*2.3E-03*	3.85E-04	*1.6E-04*
Mu	-0.012	*0.042*	0.010	*0.038*	0.052	*0.071*	-0.045	*0.071*	5.50E-03	*1.1E-03*	5.77E-05	*3.5E-05*
DJPF	-0.012	*0.041*	0.008	*0.037*	0.051	*0.067*	-0.046	*0.069*	4.03E-03	*4.6E-04*	4.37E-06	*2.2E-06*
NGM2MAXVAR	-0.012	*0.041*	0.008	*0.037*	0.051	*0.067*	-0.046	*0.069*	4.03E-03	*4.6E-04*	4.37E-06	*2.2E-06*

Table 4. Sensitivity indices (first order and total effect) for grain protein content and for the 13 genotypic parameters. Mean and standard deviation are calculated from 15 estimates.

	Winding stairs *N*=10000				Winding stairs *N*=5000				Extended Fast			
	Total sensitivity		First order sensitivity		Total sensitivity		First order sensitivity		Total sensitivity		First order sensitivity	
Parameter	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
REM2	0.486	*0.013*	0.480	*0.013*	0.495	*0.024*	0.473	*0.021*	0.493	*5.0E-03*	0.475	*5.1E-03*
R	0.345	*0.009*	0.341	*0.007*	0.339	*0.017*	0.351	*0.013*	0.349	*2.3E-03*	0.337	*2.4E-03*
RDTMAXVAR	0.149	*0.015*	0.111	*0.009*	0.143	*0.016*	0.111	*0.013*	0.144	*2.1E-03*	0.114	*1.6E-03*
Lambda	0.024	*0.014*	0.009	*0.011*	0.017	*0.015*	0.017	*0.012*	0.018	*2.4E-04*	0.012	*4.3E-04*
Mu	0.016	*0.016*	-0.002	*0.011*	0.008	*0.015*	0.007	*0.010*	0.008	*2.5E-04*	0.004	*1.2E-04*
D	0.014	*0.015*	0.013	*0.011*	0.013	*0.013*	0.011	*0.017*	0.024	*7.2E-04*	0.006	*2.8E-04*
P1GMAXVAR	0.011	*0.018*	-0.005	*0.010*	0.004	*0.018*	0.002	*0.011*	2.8E-03	*5.6E-04*	5.4E-05	*2.2E-05*
DJPF	0.010	*0.017*	-0.005	*0.010*	0.003	*0.017*	0.002	*0.011*	2.0E-03	*8.2E-05*	3.7E-07	*1.8E-07*
NGM2MAXVAR	0.010	*0.017*	-0.005	*0.010*	0.003	*0.017*	0.002	*0.011*	2.0E-03	*8.2E-05*	3.7E-07	*1.8E-07*
Ebmax	0.008	*0.020*	0.009	*0.014*	0.021	*0.020*	-0.003	*0.023*	1.9E-02	*4.1E-04*	2.9E-03	*1.2E-04*
Eimax	-0.005	*0.020*	0.006	*0.015*	0.005	*0.023*	-0.002	*0.024*	6.5E-03	*2.1E-04*	5.5E-04	*3.9E-05*
K	-0.007	*0.019*	0.008	*0.015*	0.004	*0.020*	-0.001	*0.019*	6.0E-03	*1.6E-04*	4.1E-04	*3.7E-05*
Tep.flo	-0.010	*0.018*	0.008	*0.016*	0.002	*0.020*	-0.003	*0.018*	2.7E-03	*9.4E-05*	1.7E-04	*1.0E-05*

Tables 3 and 4 show the average values and standard deviations of the first-order and total sensitivity indices obtained for yield and grain protein content with Winding stairs (*N*=1000 and *N*=5000) and extended FAST. The average sensitivity indices obtained with the different methods confirm the results shown in figure 1. For yield, the three parameters with the highest sensitivity indices are *RDTMAXVAR*, *Ebmax* and *D*. For grain protein content, the parameters *REM2*, *R* and *RDTMAXVAR* have the highest indices.

The average values of the indices obtained with the different methods are quite similar in most cases (tables 3 and 4). For example, the total sensitivity index obtained for grain protein content and parameter *REM2* is equal to 0.486 with Winding stairs *N*=10000, to 0.495 with Winding stairs *N*=5000, and to 0.493 with extended FAST. There are few discrepancies between methods. For example, the parameter with the second highest total sensitivity index is *Ebmax* with Winding stairs *N*=10000 and extended FAST, but is *D* with Winding stairs *N*=5000.

The standard deviations of the indices are very different among the methods. With Winding stairs, the standard deviation is lower when the computations are performed from *N*=10000 simulations than from *N*=5000. For example, for *REM2* and grain protein content, the standard deviation of the total sensitivity index is equal to 0.013 when *N*=10000 but is equal to 0.024 when *N*=5000 (table 4). The use of only 5000 simulations can lead to inaccurate estimations of sensitivity indices. Consequently, with the Winding stairs method, it seems necessary to use at least 10000 simulations to obtain accurate results. Tables 3 and 4 also show that the standard deviations of the indices are much lower with extended FAST than with Winding stairs. For instance, for REM2 and grain protein content, the standard deviation of the estimated values of the total sensitivity index is only equal to 0.005 with extended FAST (table 4). With extended FAST, the variability of the estimated values is due to the variability of the random phase-shift parameter. Note that the results of extended FAST are based only on 13*5000 simulations. Consequently, in terms of efficiency, the extended FAST method seems to perform better than Winding stairs.

Another advantage of the extended FAST method is that it does not give negative estimated values. Negative values are obtained with the Winding stairs method for the parameters that have a very small influence on the model outputs. Of course, such values are unrealistic. For example, for yield and parameter *P1GMAXVAR*, the total sensitivity index is equal to –0.003 with the Winding stairs method and *N*=10000 (table 3). The value of the same index is equal to 0.001 when computed from the extended FAST method. Negative values could be avoided with the Winding stairs method by calculating the correction term suggested by Homma and Saltelli [4]. But the calculation of this correction term requires additional model evaluations.

The comparison between the first-order and total sensitivity indices allows us to study the contribution of the main effect of the genetic parameters to the total output variances (tables 3 and 4). In most cases, the first-order indices represent an important fraction of the total indices, notably for grain protein content. For example, for *REM2* and grain protein content, the estimated value of the first order index is in the range 0.47-0.48 depending on the method. These values are very near from the average of the 15 estimated values of the total sensitivity index (~0.49).

3.2. Sensitivity indices for *NNI*

The Winding stairs and extended FAST methods were also applied to compute sensitivity indices for the state variable *NNI* (nitrogen nutrition index). This variable is calculated each day between end-of-winter and harvest by the model. Sensitivity indices were computed for each daily value of *NNI*. The results obtained with Winding stairs and extended FAST for parameter *D* (ratio of leaf area to critical nitrogen) are shown in figure 2. At the beginning of the growing period the total sensitivity indices are in the range 0.6-0.8. After day 60, the index decreases sharply and is almost equal to zero after day 100. This result is easily explained by studying the model equations. With AZODYN, *NNI* is calculated in function of the crop biomass and the biomass depends on the leaf area. The leaf area is calculated in function of parameter *D* only at the beginning of the growing period. The leaf area reaches its maximal value after few month of growing. After this stage, the leaf area does not depend on *D* anymore.

As for yield and grain protein content, the results obtained with Winding stairs for *NNI* are inaccurate when *N*=5000; the confidence intervals are larger with *N*=5000 than with *N*=10000. Also, with *N*=5000, the first-order sensitivity indices are higher than the total sensitivity indices at the beginning of the growing period. This is an unrealistic result.

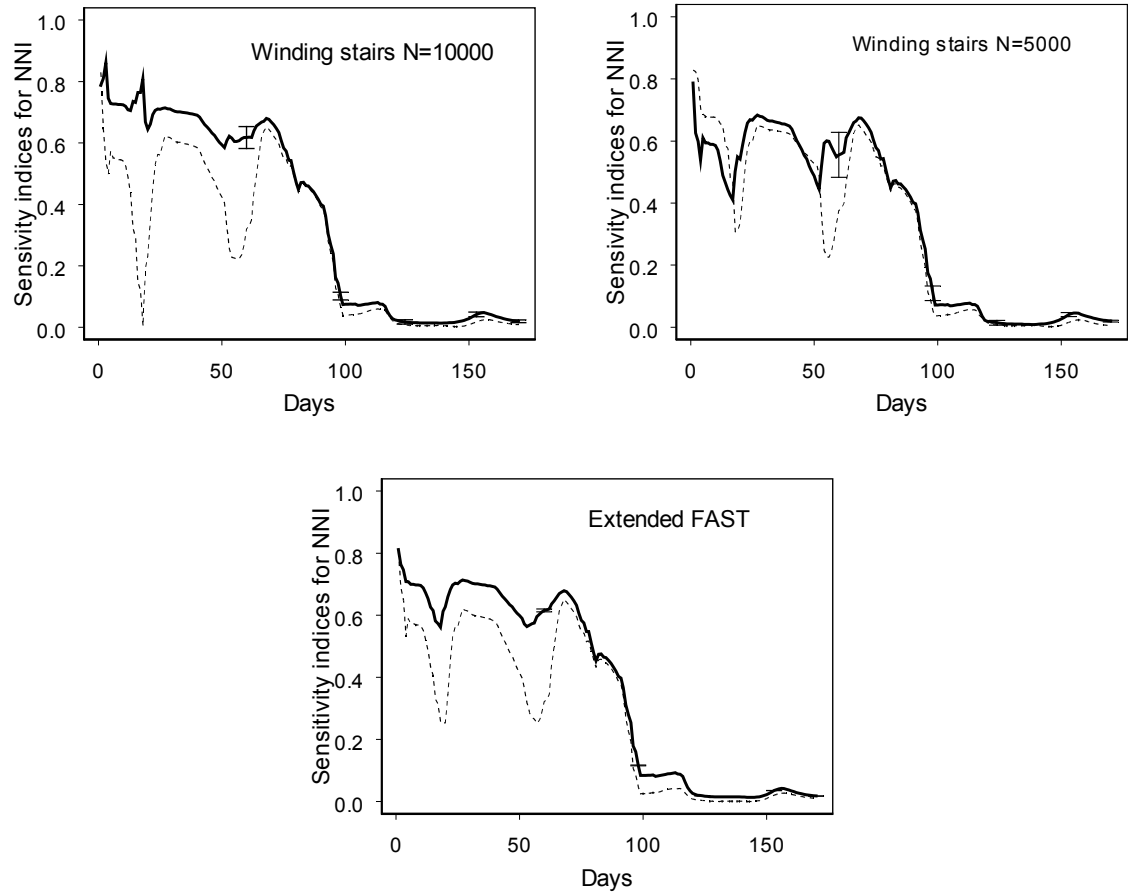


Figure 2. Total sensitivity indices (continuous line) and first-order sensitivity indices (dashed line) for the output variable *NNI* (nitrogen nutrition index) and for parameter *D* (ratio of leaf area to critical nitrogen) obtained with the methods Winding stairs (*N*=10000 and *N*=5000), and Extended FAST. The curves indicate the average values of the sensitivity indices calculated from 15 estimates between the end-of-winter and harvest. The vertical bars indicate the 95% confidence intervals at four dates.

4. CONCLUSION

Our study demonstrates that global sensitivity analysis can be used to identify the genetic parameters that must be estimated from plant breeding experiments. The methods considered in this study allow agronomists to determine which subset of parameters accounts for most of the output variance. These methods are useful and easy to interpret. Those factors with a small contribution can be set equal to any value within their range. This contributes to a model simplification and a reduction of the number of experiments performed for estimating crop model parameters.

Our application shows that only 5 parameters have a significant influence on the yield and grain protein content values simulated by the AZODYN crop model. Among these parameters, some can be easily estimated from plant breeding experiments like, for instance, the parameter *RDTMAXVAR* that represents the maximal yield value of a wheat genotype. Others are much more difficult to estimate like the parameter *R* (ratio of total to above ground nitrogen). Note that these results were obtained by running the crop model for a single field and a single year. It would be useful to repeat the analysis for other fields and several years.

The results obtained with the Winding stairs and extended FAST methods are quite similar but the extended FAST method seems to be more efficient. With the Winding stairs method, it is necessary to use at least 10000 model evaluations per parameter for estimating accurately the first-order and total sensitivity indices. With 5000 model evaluations, the Winding stairs method gives inaccurate estimates of sensitivity indices for the parameters that have a small influence on the model outputs.

REFERENCES

1. M-H. Jeuffroy, and S. Recous. Azodyn : a simple model for simulating the date of nitrogen deficiency for decision support in nitrogen fertilization. *European Journal of Agronomy*, 10:129-144, 1999.
2. J-M. Meynard, M. Cerf, L. Guichard, M-H. Jeuffroy, and D. Makowski. Which decision support tools for the environmental management of nitrogen ? *Agronomie*, 22:817-829, 2002.
3. A. Saltelli, S. Tarantola, and F. Campolongo. Sensitivity analysis as an ingredient of modeling. *Statistical Science*, 45:377-395, 2000.
4. T. Homma, and A. Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety*, 52:1-17, 1996.
5. M.J.W. Jansen, W.A.H. Rossing, and R. Daamen. Monte Carlo estimation of uncertainty contributions from several independent multivariable sources. In J. Gasman and G. van Straten, editors, *Predictability and Nonlinear Modelling in Natural Sciences and Economics*, pages 334-343. Kluwer Academic Publishers, Dordrecht, 1994.
6. K. Chan, S. Tarantola, A. Saltelli, and I. M. Sobol'. Variance-based methods. In Saltelli et al., editors, *Sensitivity Analysis*, pages 167-197. Wiley, New York, 2001.

Sensitivity Analysis in presence of model uncertainty and correlated inputs

Julien JACQUES[a], *Christian LAVERGNE*[b] *and Nicolas DEVICTOR*[c]

[a] INRIA Rhône-Alpes, 655, avenue de l'Europe - Montbonnot Saint-Martin,
38334 Saint-Ismier Cedex, FRANCE;
[b] Unité Mixte de Recherche CNRS / Université Montpellier II,
Institut de Mathématiques et de Modélisation de Montpellier, Place Eugène Bataillon,
34095 Montpellier Cedex 5, FRANCE;
[c] CEA Cadarache, DEN/DER/SESI/LCFR Bat 212,
13108 St Paul lez Durance Cedex, FRANCE
E-mail: julien.jacques@inrialpes.fr, christian.lavergne@inria.fr, nicolas.devictor@cea.fr

Abstract: First motivation of this work is to take into account model uncertainty in sensitivity analysis. So, we present in a first part, with some cases, an outline of the methodology used to treat uncertainty due to a mutation of the studied model. Development of this methodology have highlighted an important problem, frequently encoutered in sensitivity analysis: how to interpret sensitivity indices when model random inputs are non-independent? Also, we present a method to solve this problem, which introduce multidimensional sensitivity indices. Practical and theoretical applications will illustrate interest of this method.

1. INTRODUCTION

In many fields like reliability of mechanical structures, behavior of thermohydraulic systems, or nuclear safety, mathematical models are used, for simulation, when experiments are too expensive or even impracticable (nuclear accident), and for prediction.
In this context, sensitivity analysis is often used for model calibration or model validation, and to find which variables mostly contribute to output variability. In this paper, we consider global sensitivity analysis, like named in [3], based on the study of the variances of model variables. Those methods consist in the computation of sensitivity indices, which apportion the sensitivity of model output variance to model inputs. For a model

$$Y = f(X_1, ..., X_p),$$

first order sensitivity indices are defined by

$$S_i = \frac{V(E[Y|X_i])}{V(Y)}, \tag{1}$$

and express the part of variance of model output Y due to model input X_i. Higher order indices are also defined, to express effect of input interactions and total indices for total effect of one input. An important property, which enables us to easily interpret sensitivity indices values, is that the sum of all these indice is equal to 1, when inputs are independent (for more details on this property, the reader is referred to [5]). Methods of estimation of those indices are introduced by Cukier (FAST [1], [4]), Sobol [5], McKay [2], among others. We will use Sobol

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

method for numerical experiments.
The purpose of our works is to take into account a particular characterization of model uncertainty in sensitivity analysis. First of all, let us present this problem, often encountered in practice: consider that a model, on which sensitivity analyse have been made, undergoes a transformation, or, in other words, a mutation. In this case, is it possible to obtain information about sensitivity analysis of the mutated model, without doing a new complete analysis, but by using sensitivity results on the original model? In the first part, we will present an outline of the methodology which we used to answer to this question. For some possible mutations, we will mathematically relate sensitivity indices of original model with those of mutated model. Following nature of the mutation, some assumptions are necessary, and which one is most often met, is independence of the model inputs. As this last assumption is sometimes difficult to justify in practice, and as usual sensitivity indices (1) aren't meaningful when inputs are non-independent, we will present in a second part a new method of sensitivity analysis for those models.

2. IMPACT OF MODEL UNCERTAINTY ON SENSITIVITY ANALYSIS

Assume that a sensitivity analysis have been made on a model $M : Y = f(X_1, ..., X_p)$, where the n inputs variables X_i are independent. Let us suppose that new informations about the model, new measurements, or even changes in the modelled process, oblige us to consider a new model M_{mut}, that is also a mutation of the original model M. Rather than to make an exhaustive list of all possible mutations, let us present only some usefull mutations, for which interesting results have been obtained.
Firstly, consider a model $M : Y = f_1(X_1) + f_2(X_2, ..., X_p)$, where $(X_1, ..., X_p)$ are independent random variables, and suppose that M undergoes a mutation, and is also transformed in a new model M_{mut} where X_1 is fixed to its mean $\mu_1 = E[X_1]$. Thus, this new model is $Y^m = f_1(\mu_1) + f_2(X_2, ..., X_p)$. Writing definition of sensitivity indices, we show that M_{mut} sensitivity indices (S^m) can be express from sensitivity indices (S) of M by:

$$S^m = S \times \frac{V(Y)}{V(Y^m)} \quad \text{for first and higher order sensitivity indices.}$$

and by:

$$S_T^m = 1 - (1 - S_T) \times \frac{V(Y)}{V(Y^m)} \quad \text{for all total sensitivity indices.}$$

Of course, all indices relating to variable X_1 disappear.
Let us consider now inverse case, which can be view as introduction of noise in the model, and which consist to consider a deterministic parameter like a random variable. So the model $M : Y = f_1(\mu) + f_2(X_1, ..., X_p)$ is mutated in a model $M_{mut} : Y^m = f_1(X_{p+1}) + f_2(X_1, ..., X_p)$. In this case, sensitivity indices of M_{mut}, are given by those of M multiplied by $V(Y)$ and divided by $V(f_1(X_{p+1})) + V(Y)$. For the new variable, only first order indice are non zero, and is given by

$$\frac{V(f_1(X_{p+1}))}{V(f_1(X_{p+1})) + V(Y)}.$$

For the same mutation carried out on the model $M : Y = f_1(\mu) \times f_2(X_1, ..., X_p)$, sensitivity indices of M_{mut} can be obtain multiplying indices of M by

$$\frac{V(Y)}{V(Y^m)} \times \left(\frac{E[f_1(X_{p+1})]}{f_1(\mu)} \right)^2 .$$

Now, if we consider the new variable X_{p+1} as dependent from the others variables, we are again confronted with the problem of sensitivity analysis for model with dependent inputs previously evoked. Also, we don't know to deduce sensitivity indices of the mutated model from the knowledge of the M model.
Let us finally present an other type of mutation. Assume that two analysis have been made on two models $M_1 : Y_1 = f_1(X_1, ..., X_p)$ and $M_2 : Y_2 = f_2(X_{p+1}, ..., X_{p+q})$, and also that sensitivity indices S^1 for M_1 and S^2 for M_2 have been computed. We suppose that inputs variables of the two models are different and independent. Let us create a new model $M_{mut} : Y^m = Y_1 + Y_2$. Sensitivity indices of M_{mut} are obtained by multiplying

$$\text{those of } M_1 \text{ by } \frac{V(Y_1)}{V(Y_1) + V(Y_2)} \text{ and those of } M_2 \text{ by } \frac{V(Y_2)}{V(Y_1) + V(Y_2)}.$$

All sensitivity indices, relative to interaction between M_1 variables and M_2 ones are equal to zero. If we suppose that there are dependences between variables of the two models, we are afresh confronted with the same problem of sensitivity analysis for dependant or correlated inputs.
To conclude, if an original model, on which sensitivity analysis have been made, is transformed, it's possible to deduce sensitivity indices of the mutated model, without starting again heavy calculation of Monte Carlo, in a given number of cases. Those cases are principally deletion of variables or introduction of new independent variables. On the other hand, introduction of dependent variables, or even of existing variables poses the problem of sensitivity analysis with dependent inputs, for which we propose a new method.

3. SENSITIVITY ANALYSIS FOR MODEL WITH DEPENDENT OR CORRELATED INPUTS

Highlighted in previous section, the problem of sensitivity analysis for model with dependent inputs is a real one, because naturally frequently met in practice.
This problem concern the interpretation of sensitivity indices values. When inputs are independent, I.M.Sobol demonstrates that the sum of all sensitivity indices is equal to 1. Effectively, in Sobol's decomposition of model function, all term are mutually orthogonal if inputs are independent, and so we can obtain a variance decomposition of model output. Dividing this decomposition by output variance, we obtain exactly that the sum of all order indices is equal to 1. If we don't assume that the inputs are independent, the terms of model function decomposition are not orthogonal, and so it appears a new term in the variance decomposition. That's this term which implies that the sum of all order sensitivity indices is not equal to 1. We can give the following interpretation to this : when we study sensitivity of one input, which is correlated with another one, we study too sensitivity of this last. Effectively, variabilities of two correlated variables are link, and so when we quantify sensitivity to one of this two variables, we quantify

too a part of sensitivity to the other variable. And so, in sensitivity indices of the two variables, the same information is taken into account several times, and sum of all indices is thus greatest than 1.
Natural idea is also coming: to define multidimensional sensitivity indices for groups of correlated variables.

3.1. Multidimensional sensitivity analysis

Consider the model

$$Y = f(X_1, ..., X_p),$$

where

$$(X_1, ..., X_p) = (\underbrace{X_1}_{\mathbb{X}_1}, ..., \underbrace{X_i}_{\mathbb{X}_i}, \underbrace{X_{i+1}, ..., X_{i+k_1}}_{\mathbb{X}_{i+1}}, \underbrace{X_{i+k_1+1}, ..., X_{i+k_2}}_{\mathbb{X}_{i+2}}, ..., \underbrace{X_{i+k_{l-1}+1}, ..., X_p}_{\mathbb{X}_{i+l}})$$

$(X_1, ..., X_i) = (\mathbb{X}_1, ..., \mathbb{X}_i)$ are independent inputs, and $(\mathbb{X}_{i+1}, ..., \mathbb{X}_{i+l})$ are l groups of intra-dependent or intra-correlated inputs ($\mathbb{X}_i$ are independent of $\mathbb{X}_j$, for all $1 \leq i, j \leq l$).
We wrote monodimensional non independent variables $(X_1, ..., X_p)$ like multidimensional independent variables $(\mathbb{X}_1, ..., \mathbb{X}_{i+l})$.
Thus, we define first order sensitivity indices

$$S_j = \frac{V(E[Y|\mathbb{X}_j])}{V(Y)} \quad \forall j \in [1, i+l]$$

To connect this to monodimensional variables, if $j \in [1, ..., i]$, we have well define the same indice:

$$S_j = \frac{V(E[Y|\mathbb{X}_j])}{V(Y)} = \frac{V(E[Y|X_j])}{V(Y)} \tag{2}$$

and if $j \in [i+1, ..., i+l]$, for example $j = i+2$:

$$S_j = S_{\{i+k_1+1,...,i+k_2\}} = \frac{V(E[Y|X_{i+k_1+1}, ..., X_{i+k_2}])}{V(Y)} \tag{3}$$

Now, like in classical analysis, we can also define higher order indices and total sensitivity indices. Second order indices are given by

$$S_{jk} = \frac{V(E[Y|\mathbb{X}_j, \mathbb{X}_k] - E[Y|\mathbb{X}_j] - E[Y|\mathbb{X}_k])}{V(Y)},$$

and so on for higher order indices. And finally, total order indices are defined by :

$$S_{T_j} = \sum_{k \# j} S_k,$$

where $\#j$ represent all subsets of $\{1, ..., i+l\}$ which include j.
It's very important to note that if all input variables are independent, those sensitivity indices are clearly the same than (1). And so, multidimensional sensitivity indices can well be interpreted like a generalization of usual sensitivity indices (1).

3.2. Numerical estimation

Like in classical analysis (Sobol), Monte-Carlo estimations are possible.
We estimate mean and variance of Y by :

$$\hat{f}_0 = \frac{1}{N}\sum_{k=1}^{N} f(\mathbf{x}_1^k, ..., \mathbf{x}_{i+l}^k) \qquad \hat{D} = -\hat{f}_0^{\,2} + \frac{1}{N}\sum_{k=1}^{N} f^2(\mathbf{x}_1^k, ..., \mathbf{x}_{i+l}^k),$$

and first order indice by $\hat{S}_j = \frac{\hat{D}_j}{\hat{D}}$ with :

$$\hat{D}_j = \frac{1}{N}\sum_{k=1}^{N} f(\mathbf{x}_1^k, ..., \mathbf{x}_{j-1}^k, \mathbf{x}_j^k, \mathbf{x}_{j+1}^k, ..., \mathbf{x}_{i+l}^k) f(\mathbf{x}_1^k, ..., \mathbf{x}_{j-1}^k, \underline{\mathbf{x}}_j^k, \mathbf{x}_{j+1}^k, ..., \mathbf{x}_{i+l}^k) - \hat{f}_0^2,$$

where $(\mathbf{x}_1^k, ..., \mathbf{x}_{i+l}^k)_{k=1,N}$ and $(\underline{\mathbf{x}}_1^k, ..., \underline{\mathbf{x}}_{i+l}^k)_{k=1,N}$ are two independent sets of N (multidimensional) inputs simulations. Equivalent estimations for higher order and total indices exist.

3.3. Application in nuclear field - epithermal indice

Study presented here is a sensitivity analysis of a model, which compute an epithermal indice for a given nuclear reactor. The epithermal indice is defined by the value of the neutron epithermal flow divided by the neutron thermal flow. This indice is useful in studies of nuclear reactor vessel dosimetry.
This model is made of 4 inputs, of which two are correlated:

$$\begin{aligned}
\text{resonance integral of Co59} &: X_1 \sim \mathcal{N}(72, 7.2^2) \\
\text{factor Fcd} &: X_2 \sim \mathcal{N}(\log(1.01989), 0.0147051^2) \\
\text{activity of the dosimeter Co59 "nu"} &: X_3 \sim \mathcal{N}(4.703 \times 10^7, 1147732^2) \\
\text{activity of the dosimeter Co59 under Cadmium} &: X_4 \sim \mathcal{N}(2.522 \times 10^7, 615368^2) \\
\text{with correlation coefficient} & \quad \rho_{X_3X_4} = 0.85,
\end{aligned}$$

one output Y : epithermal indice, and one function which links inputs and output:

$$Y = \frac{\exp X_2(1.008843 - 0.02114316X_1 + 9.858080.10^{-5}X_1^2 + 1.931988.10^{-8}X_4)}{(1 - \exp X_2\frac{X_4}{X_3})(-0.00575077 + 3.73935.10^{-8}X_3)}.$$

Like explicited previously, as two inputs are correlated, it's useless to compute usual sensitivity indices, because results will not be meaningful. We thus carried out a multidimensionnal sensitivity analysis. Numerical experiments have been made repeating all indices computations 20 times, with $N = 20000$ Monte-Carlo iterations. Mean of these 20 estimations, represented on figure 1, are the following:

$$\begin{gathered}
S_1 \simeq 0.85 \quad S_2 \simeq 0.09 \quad S_{\{3,4\}} \simeq 0.07 \\
S_{12} \simeq S_{1\{3,4\}} \simeq S_{2\{3,4\}} \simeq S_{12\{3,4\}} \simeq 0
\end{gathered}$$

where $S_{\{3,4\}}$ are the first order sensitivity indices of the multidimensional variable $\{X_3, X_4\}$.
Multidimensional analysis allows us to conclude that this model is sensitive essentially to input

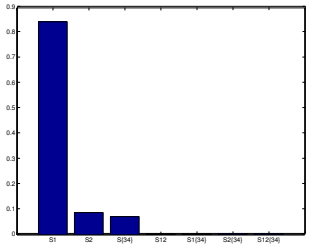


Figure 1. Sensitivity indices of epithermal indice model

X_1 (*resonance integral of Co59*), and that others variables are less significant. But in this application, the interest of our method is not very well exhibit. Effectively, as X_1 and X_2 are independent from the other variables, we can apply classical sensitivity analysis and find the same value for S_1 and S_2. And also, as the sum of this two indices are equal to 0.94, we can deduce that the other variables and all the interaction with them, have only small importance. We will present a theoretical application, which emphasizes more multidimensional sensitivity analysis.

3.4. Theoretical application

Consider the model

$$Y = aX_1X_2 + bX_3X_4 + cX_5X_6,$$

where $X_i \sim \mathcal{N}(0,1)$, for $i = 1$ to 6, and where X_3 and X_4 are correlated ($\rho_{X_3,X_4} = \rho_1$), like X_5 and X_6 ($\rho_{X_5,X_6} = \rho_2$). Sensitivity indices are the following:

$$\begin{aligned} S_{12} &= \frac{a^2}{a^2 + b^2(1+\rho_1)^2 + c^2(1+\rho_2)^2} \\ S_{\{3,4\}} &= \frac{b^2(1+\rho_1)^2}{a^2 + b^2(1+\rho_1)^2 + c^2(1+\rho_2)^2} \\ S_{\{5,6\}} &= \frac{c^2(1+\rho_2)^2}{a^2 + b^2(1+\rho_1)^2 + c^2(1+\rho_2)^2} \end{aligned}$$

and all the other indices are equal to 0. We constate that the value of the numerator of the interaction sensitivity indice S_{12} is a function of the coefficient a. The values of numerators of the non zero sensitivity indices $S_{\{3,4\}}$ and $S_{\{5,6\}}$ are function of the model coefficients b and c, but too of the correlation coefficient ρ_1 or ρ_2. To illustrate this, let us present some numerical values of those indices, for different values of the coefficients of the model (a, b and c) and the correlation coefficients.

situation	a	b	c	ρ_1	ρ_2	S_{12}	$S_{\{3,4\}}$	$S_{\{5,6\}}$
(i)	1	1	1	0.8	0.8	0.2336	0.3832	0.3832
(ii)	3	1	1	0.8	0.8	0.7329	0.1336	0.1336
(iii)	1	1	3	0.8	0.8	0.0575	0.0943	0.8483
(iv)	1	1	1	0.8	0.3	0.2881	0.4397	0.2922
(v)	1	1	3	0.8	0.3	0.0803	0.1317	0.7880
(vi)	1	1	3	0.3	0.8	0.0593	0.0647	0.8760

First of all, let us underline that as X_1 and X_2 are independent variables, indices S_1, S_2, and S_{12} are usual sensitivity indices, and can also be computed without our multidimensional method.
In the situation (ii), as X_1 and X_2 are independent variables, usual sensitivity indices allows us to conclude that variance of Y is essentially (73%) due to interaction between X_1 and X_2. But in the others situations, when X_1 and X_2 are less important, we need multidimensional sensitivity indices to apportions effect to the two couple (X_3, X_4) and (X_5, X_6). These multidimensionnal indices allow us to know that couple (X_3, X_4) and (X_5, X_6) have the same importance in the situation (i), and that (X_5, X_6) is the most important in situation (iii). Effectively, in situation (i) couples (X_3, X_4) and (X_5, X_6) are symmetric in the model, and so they have same importance. In (iii) a coefficient equal to 3 is multiplying the product X_5X_6, that's why the couple (X_5, X_6) is most important than (X_3, X_4).
Situations (iv), (v) and (vi) illustrate that indices $S_{\{3,4\}}$ and $S_{\{5,6\}}$ are function to the correlation (S_{12} is too function to the correlation, but it's due to its denominator, which is the variance of Y). As couples (X_3, X_4) and (X_5, X_6) are in the model in a product form: X_3X_4 and X_5X_6, greater is the correlation, greater is the importance of the couple, and so greater is the value of the sensitivity indices. In (iv) the correlation of (X_3, X_4) is greater than correlation of (X_5, X_6), and so $S_{\{3,4\}}$ is greater than $S_{\{5,6\}}$. In situations (v) and (vi), we can see the same behaviour.

4. CONCLUSION AND FUTURE WORK

We have presented in this paper two works : the first concern integration of a view point of model uncertainty in sensitivity analysis, which we interpret like a model mutation. We drew up an outline of the employed methodology, which consists in a listing of possible mutations, for each one which we examine the impact on the computing of sensitivity indices. Second work introduces a new method which allows to compute useful and comprehensible sensitivity indices for a model with non-independent inputs. Practical and theoretical illustrations of interest of this method have been presented.
Further applications and developments are envisaged, in particular when there are many model inputs.

REFERENCES

1. R.I.Cukier, H.B.Levine, K.E.Shuler. *Nonlinear Sensitivity Analysis of Multiparameter Model Systems*. Journal of Computational Physics, 1978.
2. M.D.McKay. *Evaluating Prediction uncertainty. Technical Report NUREG/CR-6311*. US Nuclear Regulatory Commission and Los Alamos National Laboratory, 1995.
3. A.Saltelli, K.Chan, E.M.Scott. *Sensitivity Analysis*. Wiley, 2000.
4. A.Saltelli, S.Tarantola, K.Chan. *A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output*. Technometrics, Vol.41, No.1, 1999.
5. I.M.Sobol. *Sensitivity Estimates for Nonlinear Mathematical Models*. Mathematical Modelling and Computational Experiments, 1993.

Varying Correlation Coefficients Cannot Account for Uncertainty About Dependence, But There Are Comprehensive Methods to Do So

S. Ferson

Applied Biomathematics, 100 North Country Road, Setauket, New York 11733, USA
E-mail: scott@ramas.com

Abstract: In accounting for the dependencies among variables in probabilistic (convolution) models, a sensitivity study that varies a correlation between plausible values, even the extremes of +1 and −1, cannot characterize the possible range of results that could be entailed by nonlinear dependencies. Because a functional modeling strategy that seeks to model mechanistically the underlying sources of the dependencies will often be untenable, a phenomenological approach will often be needed to handle dependencies. We summarize recent algorithmic advances that allow the calculation of results under particular bivariate dependence functions, under only partially specified dependence functions, or even without any assumption whatever about dependence.

Keywords: dependence, correlation, copula, comonotonicity, functional modeling

1. INTRODUCTION

Although risk analyses often still assume independence among input variables as a matter of mathematical convenience, most analysts recognize that intervariable dependencies can sometimes have a substantial impact on computational results. In the face of epistemic uncertainty about dependencies, analysts occasionally employ a sensitivity study in which the correlation coefficient is varied between plausible values. This strategy is insufficient to explore the possible range of results however, as can be shown by simple examples. Fortunately, comprehensive bounds on convolutions of probability distributions (or even bounds thereon) can be obtained using simple formulas that are computationally cheaper than Monte Carlo methods. We review the use of these formulas in the cases of variously restricting assumptions about dependence, from no assumption at all, to specified sign of the dependence, to a particular dependence function.

2. FUNCTIONAL MODELING

Some analysts argue that it is best to reduce any problem involving dependent variables into one with only independent variables. This changes the problem of statistically representing dependent variables into a modeling problem of reproducing the functional or mechanistic relationships that induce the dependence. It is not sufficient to transform the model into one in terms of uncorrelated variables; they must be statistically *independent* variables. Of course, this functional modeling approach could entail considerable effort far beyond the scope of the immediate assessment. The extra modeling effort required by this strategy may not be workable in many situations. For instance, a dam safety engineer worried about computing risks to a water control structure from hydrological factors influenced by weather patterns would need to model various meteorological and even climatological phenomena. At some

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

point, the analytical demands of a functional modeling approach will likely become prohibitive.

3. STOCHASTIC MODELS OF DEPENDENCE: THE COPULA

There are three other approaches to the problem of accounting for dependence among variables: (*i*) assume a particular dependence function, (*ii*) make no assumptions about dependence, and (*iii*) relax assumptions to a partial specification of dependence. In the first approach, one must assume a particular dependence function among the variables. Assuming independence is of course a special case of this approach. Another special case is assuming perfect dependence among variables such that each variable is almost surely a monotonically increasing function of the other. In general, the dependence function is specified as some copula [1]. A copula is the function that characterizes how the marginal distributions are knitted together to form the joint distribution. In the two-dimensional case, a copula is just a bivariate distribution function from the unit square onto the unit interval that has uniform marginals. A bivariate distribution function $F(x, y)$ can be expressed in terms of the copula C as $C(F(x), G(y))$ where $F(x)$ and $G(y)$ are its marginal distribution functions. The dependence function could be specified by selecting a copula from a parameterized family of copulas such as the Frank, Mardia, normal or Clayton families [1, 2]. It could also be specified with an empirical copula, which is an analog for dependence of an empirical distribution function.

4. COMPUTING WITH A SPECIFIED COPULA

In risk analyses, distributions characterizing random variables are convolved together to estimate arithmetic functions (such as sums, products, differences, quotients, etc.) of the random variables. For instance, if X and Y are random variables with distributions F and G respectively, the distribution of the sum $Z = X+Y$ can be obtained with the Lesbesgue-Steiltjes integral

$$\sigma_{+,C}(F,G)(z) = \int_{x+y<z} dC\big(F(x),G(y)\big)$$

which always exists. This formulation includes the independence case where $C(u,v) = uv$. Similar formulas are available to compute distributions of products, differences, quotients, etc. We describe a straightforward numerical procedure to compute σ given discretizations for the marginal distributions F and G and an arbitrary copula C. The numerical methods extend easily to other arithmetic operations. Note that this approach can handle arbitrarily complicated dependence between the input variables. This makes the approach significantly more general than methods implemented in common risk analysis software packages which model correlations but not dependencies in general.

5. COMPUTING WITHOUT ANY ASSUMPTION ABOUT DEPENDENCE AT ALL

The second approach to accounting for dependence is to *make no assumptions whatever* about the dependence between variables. In this approach, bounds on the distribution of an arithmetic function can be computed directly using infimal and supremal convolution of the marginal distributions of the addends. For example, if X and Y are random variables with marginal distributions F and G respectively, then the bounds on the distribution of $Z = X+Y$ are

$$\left[\sup_{z=x+y} \max\big(F(x)+G(y)-1,\,0\big),\ \inf_{z=x+y} \min\big(F(x)+G(y),\,0\big) \right].$$

where the supremum gives the left bound on the distribution (i.e., the upper bound on the cumulative probability associated with any value of the sum z), and the infimum gives the right bound on the distribution function (the lower bound on the value of the cumulative probability). These bounds satisfy a problem originally posed by Kolmogorov of finding bounds on the distribution of a sum given only distributions of the addends. Kolmogorov's problem was solved by Makarov [3] and Frank et al. [4]. Analogous bounds on distributions of products, differences, quotients, etc., can likewise be obtained from similar supremal and infimal convolutions of the marginal distributions of the factors, etc. Williamson and Downs [5] described convenient numerical algorithms to compute these bounds in a way that accounts for discretization error introduced by encoding the marginal distribution with a finite computer representation. With their algorithms, the bounding convolutions are generally much less expensive than ordinary convolution computed via Monte Carlo simulation. The bounds obtained by the supremal and infimal convolutions are guaranteed to rigorously enclose all distributions that could arise for the sum (or product, etc.), no matter what dependence there may be between the addends (or factors, etc.). Furthermore, these bounds are also best possible, that is, they are as tight as can be justified without any knowledge about the dependence. The breadth between the bounds characterizes the specificity of the answer under the relaxed dependence assumption. It is interesting that these bounds *cannot* be obtained with the standard σ-convolution described above such as by varying the correlation between +1 and –1. Even varying the dependence function C between perfect dependence (maximal correlation and comonotonicity) and opposite dependence (minimal correlation and countermonotonicity) will generally underestimate the breadth of the bounds. The difference is due to nonlinear dependencies which are ignored by merely varying correlations between extreme values. This approach can be combined with independence assumptions, so that some variables are assumed to be independent and no assumptions are made about the dependence between other variables.

6. COMPUTING WITH ONLY PARTIALLY SPECIFIED DEPENDENCE

The third approach to account for dependence in risk assessments is to make some qualitative or quantitative assumptions about the dependence function that partially specify the copula. For instance, a promising approach to tighten risk calculations is to make use of information about the *sign* of the dependence between the variables. The most common notion of sign dependence is positive quadrant dependence. Random variables X and Y with distribution functions F and G whose joint distribution is H are positively quadrant dependent if $H(x, y) \geq F(x)\ G(y)$ for all x and y, so that if the probability that the random variables are both small (or large) is at least as great as if they were independent. There are several conditions that imply variables will be positively quadrant dependent, including when each is a stochastically increasing function of the other, i.e., $P(Y > y \mid X=x)$ is a non-decreasing function of x for all y, and $P(X > x \mid Y=y)$ is a non-decreasing function of y for all x. Positive quadrant dependence implies non-negative Pearson, Spearman and Kendall correlations, although the mere observation that a correlation is positive does not imply the variables are positively quadrant dependent. This idea has been used in many statistical and engineering

settings, and seems to capture one sense analysts have in mind when they use the phrase 'positively depends'.

Risk assessments can make use of assumptions about the sign of the dependence among variables with easy-to-compute convolutions. For example, bounds for a sum of positively quadrant dependent variables whose marginals are F and G are

$$\left[\sup_{z=x+y} \left(F(x)G(y)\right), \inf_{z=x+y} \left(1-\left(1-F(x)\right)\left(1-G(y)\right)\right) \right].$$

These bounds are similar to the supremal and infimal convolutions in the sense that they are guaranteed to bound the distribution function of the sum and are the tightest possible such bounds given only the marginal distributions F and G and the positivity of their dependence. Note that these formulas give bounds that are *not* the same as an envelope of the perfect and independent convolutions (which would be narrower). There are similar formulas for the other arithmetic operations, as well as complementary formulas that assume negative quadrant dependence (variables X and Y are negatively quadrant dependent if X and $-Y$ are positively quadrant dependent). The intersection of the convolution bounds for positive and negative dependencies is *not* the same as the distribution(s) obtained under independence.

One could also make a *quantitative* assumption about dependence such as that the correlation coefficient has a particular magnitude. In such cases, convolutions between distributions can be computed using mathematical programming, although it turns out that specifying only the correlation often provides very little improvement in the specificity of the result. For example, assuming that random variables X and Y are uncorrelated (that is, have Pearson correlation coefficient equal to zero) produces almost no improvement over the bounds obtained by the supremal and infimal convolutions.

7. CONCLUSIONS

The three approaches described above give analysts considerable flexibility to account for knowledge and uncertainty about correlations and dependencies. By making more assumptions, one can increase specificity of the answers that can be obtained. In a sensitivity analysis, of course, an analyst often desires to relax his assumptions and explore how the results might vary in consequence. It is possible to mix strategies so that one could posit independence among some variables, assume particular copulas for some variables, and make limited or no assumptions about the dependence among other variables. This allows an analyst to obtain a sensitivity analysis that reflects what is well known about dependencies and what is in contention about them.

ACKNOWLEDGMENTS

This work was supported by contract 19094 with Sandia National Laboratories, under the Epistemic Uncertainty project, directed by William Oberkampf. It benefited from discussions with Roger Nelsen of Lewis and Clark University, Daniel Berleant and Jianzhong Zhang of Iowa State University, and Roger Cooke of Delft University of Technology. The opinions expressed in the paper are solely those of the author.

REFERENCES

1. Nelsen, R.B. 1999. *An Introduction to Copulas*. Springer-Verlag, New York.
2. Joe, H. 1997. *Multivariate Models and Dependence Concepts*. Chapman & Hall/CRC, Boca Raton.
3. Makarov, G.D. 1981. Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability and its Applications* 26.
4. Frank, M.J., R.B. Nelsen, and B. Schweizer. 1987. Best-possible bounds for the distribution of a sum—a problem of Kolmogorov. *Probability Theory and Related Fields* 74: 199-211.
5. Williamson, R.C. and T. Downs. 1990. Probabilistic arithmetic I: Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* 4: 89-158.

Application of a Markov Chain Monte Carlo calibration and uncertainty framework to a process-based integrated nitrogen model (INCA)

B. M. Jackson[a], *H. S. Wheater*[a], *N. McIntyre*[a] *and P. Whitehead*[b]

[a]Department of Civil and Environmental Engineering, Imperial College London, London SW7 2AZ, U.K.;
[b]Department of Geography, The University of Reading, Reading RG6 6AB, U.K.
E-mail: b.m.jackson@imperial.ac.uk

Abstract:

As part of the production of an integrated modelling system for lowland permeable catchments, a stochastic framework is being developed to allow quantification of uncertainty in the representation of catchment response and impacts of management scenarios, and to investigate how capable various conceptual models are of adequately characterising water flow, nitrate and phosphorous transport given a reasonably calibrated and, where applicable, physically realistic parameter set. To address these considerations, a collection of stochastic routines, including Markov chain Monte Carlo capabilities, have been integrated with a semi-distributed nitrogen model. This Integrated Nitrogen in Catchments model (INCA) simulates flow, nitrate and ammonium and tracks the temporal variations in flow and nitrogen mass operating in both land and river phases. This paper discusses some of the issues and initial results arising from the first application of Markov chain Monte Carlo (MCMC) to scenarios utilising the INCA model. Performance is illustrated with data from the Kennet catchment in southern England. The results demonstrate the power of Markov chain Monte Carlo methods to quantitatively examine the interrelationship between model structure, parameter identifiability and data support, but also the reliance of MCMC and other heuristic methodologies on objective function choices and model robustness.

Keywords: Markov chain Monte Carlo, calibration, catchment modelling, nitrate

1. INTRODUCTION

An integrated nitrogen model has been developed to investigate the fate and distribution of nitrogen in the aquatic and terrestrial environment. This Integrated Nitrogen in CAtchments model (INCA) simulates flow, nitrate and ammonium over the catchment scale, coupling land processes and in-river processes. Dilution, natural decay and biochemical transformation processes are included in the model as well as interactions with plant biomass. It is semi-distributed to account for spatial variations in land use, human impacts, effluent discharges and varying deposition levels, and produces daily estimates of the stream water flow and nitrate and ammonium concentrations, in addition to estimates of annual, land-use specific, N fluxes. The original model was described in Whitehead et al. (1998), while more recent additions to the model structure are contained in [1]. The

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

model has been successfully applied to a range of catchments in the U.K. and Europe, and is still being refined to extend its applicability to a variety of catchment management needs. However, little work on parameter sensitivity and identifiability has been carried out on it to date.

A detailed study of these issues should aid the successful calibration of further catchment applications by highlighting the most significant parameters and allowing informed decisions as to the areas in which experimental resources and measurements should be allocated. For catchment management purposes, there is also a need to provide measures of the uncertainty present due to measurement errors in the inputs, parametric uncertainty, and issues related to model conceptualisation. To address these considerations, methods for propagating uncertainty, analysing parameter sensitivity and model structure, and optimisation are placed within a subjective probability framework, along with a collection of appropriate "objective functions" to specify criteria for successful calibration. The more efficient methodologies utilise heuristic guidance to explore parametric spaces and model output distributions through an automatic semi-random exploration of the parameter space. Included in these routines are Markov chain Monte Carlo methods (using Metropolis Hastings formulae), used to sample parametric and uncertain quantities. The framework permits both parametric and model structural uncertainty to be interrogated, and allows effective calibration and confidence predictions through optimisation of model inputs to fit observations or other criteria, with explicit consideration of effects of data uncertainty.

MCMC methods possess the general virtue of simulation methods, with information regarding parametric probability distributions easily collected along with optimal parameter sets. However, other sampling methods generally fail when the posterior involves many variables or is otherwise intractable. Markov chain methods are capable of sampling from posterior distributions of arbitrary complexity, through the Metropolis Hastings algorithm, which provides simple conditions under which the chain will equilibrate to the required distribution. Since such methods sit naturally within a subjective probability framework, they are also capable of quantifying distortions produced on the outputs by noise. Such a capability is indispensable for rigourous analysis of an environmental model such as INCA, as the input is subject to extreme uncertainty.

The performance of the modelling framework is illustrated with data from the Kennet catchment in southern England. To understand the characteristics of both overall uncertainty and particular parametric sensitivities in INCA, the effect of changes in the parameters and inputs are examined using the Markov chain sampling described above. Response surfaces, in this case distributions of input parameters against single-valued measures of performance (derived from the output parameters and optimality criteria), are examined, and the biases caused by differing optima considered. The influence of such biases on subsequent decisions regarding parameter sensitivities and "optimal" parameter sets is examined. The efficiencies of differing Metropolis proposal functions applied to sample both the "minima" and entirety of a given response surface are also being investigated.

The results demonstrate the power of Markov chain Monte Carlo methods to quantitatively examine the inter-relationship between model structure, parameter identifiability

and data support, and also provide an efficient means of addressing the problem of calibration given large parameter sets and the presence of measurement error and other uncertainties. In the context of model development, however, the need for alternate, non-heuristically guided methodologies to be included in such stochastic tools is also demonstrated.

2. THE STOCHASTIC FRAMEWORK

Any model describing nutrient transport within a catchment, the result of complicated environmental processes with dependencies on both space and time, is necessarily a simplified representation of the phenomena being studied. This imposes a limit upon one's confidence in its responses or outputs, regardless of the accuracy of any input information. The input itself is subject to many sources of uncertainty, including measurement errors, absence of information, temporal and spatial variability, and incomplete understanding of underlying driving forces and mechanisms. Adequate spatial representation is particularly difficult, due to the intrinsic variability present within the environment, such as the continuous variation in soil properties and nitrogen inputs over space, and the difficulty of characterising properties in the subsurface.

To give a measure of confidence in scenario predictions, a reliable catchment modelling tool should provide measures of the uncertainty present due to measurement errors in the inputs, parametric uncertainty, and issues related to model conceptualisation, and be able to translate these measures into prediction confidence limits for management purposes [2]. Where models are still in development, stochastic analysis can also aid in identifying the components of model structures that are most significant in the simulation of nutrient dynamics in river systems, aspects that appear redundant, and the inter-relationship between model structure, parameter identifiability and data support. This allows informed decisions as to the areas in which experimental resources and measurements should be allocated. The relative importance of differing measurands over space is also important.

To address all these concerns within one framework, methods suitable for prediction uncertainty, model sensitivity to parameters and data error and calibration must be associated with appropriate catchment scale models. For successful calibration, collections of appropriate "objective functions" to specify optimal criteria, data processing capabilities for handling and analysing errors and guidelines for choosing calibration criteria and parameter distributions given specific modelling tasks must also be included.

3. UNCERTAINTY AND SENSITIVITY METHODOLOGIES

To address prediction uncertainty and model sensitivity, three possible Monte Carlo methodologies have been implemented: basic Monte-Carlo simulation, Latin hypercube sampling and Markov-chain Monte-Carlo techniques (Metropolis and Metropolis Hastings). The classic Monte Carlo method samples the input parameter space using the exact probability distribution assigned to it, which, given an exact mathematical description of the model, must converge eventually to the precise output distribution. The simulation's main impediment is its high computational cost, with the run numbers necessary for a successful analysis of a model's outputs typically running into the thousands [2]. A degree

of computational efficiency can be accomplished through the use of efficient input sampling methods, which may include heuristic search procedures (purposeful or partially informed searches using heuristic functions for guidance), or less informed approaches where segments of the probability distributions are split or stratified, and systematically explored. The latter approach is contained within the framework through inclusion of the Latin Hypercube method [3].

To formally include subjective probability to be used within the framework, Markov chain Monte Carlo methods (using Metropolis Hastings formulae) are used to sample parametric and uncertain quantities. These methods sample from the input, or posterior, distribution, and sit naturally within a subjective probability (Bayesian) framework. A Markov chain is a series of random variables $\{X(0), X(1), X(2), \ldots, X(N)\}$ for which the conditional probability of a transition from any state $X(i)$ to any other state $X(j)$ depends only on the current state, and not on any previous states. The construction of a Markov chain requires two basic ingredients, namely an initial distribution (a first approximation to the probability of being in each the states X(i)) and a transition matrix [4]. This transition matrix is a matrix of probabilities, defining all the associated probabilities (transition probabilities) of the chain moving from state $X(i)$ to state $X(j)$, $i, j \in (1, N)$.

Markov chain Monte-Carlo methods draw samples from a Markov chain rather than from the probability distribution $f(x)$. When constructed carefully, these can be very efficient approximators. Most Markov chain schemes in use today, such as the popular Gills sampler, are a variant of the Metropolis Hastings approach. For details, see Gilks et al. (1996). The Markov chain is constructed such that its equilibrium distribution is that of the posterior distribution of interest. In this context, such a distribution might be that of the INCA parameters conditional on measured "output" observations and optimality constraints, or the uncertainty present in a prediction given uncertainty in measurands and model structure.

Each state is visited the required number of times to satisfy the conditional distribution of the parameters given the data. This is achieved through satisfying appropriate conditions of reversibility (detailed balance) and ergodicity (Hastings 1970). By giving the microscopic dynamics of the Markov chain (that is, an algorithm that determines $X(i + 1)$ given $X(i)$), the transition matrix and consequently the (unnormalised) input distribution is implicitly fixed. Markov chain methods are capable of sampling from posterior distributions of arbitrary complexity, through the Metropolis Hastings algorithm, which provides simple conditions under which the chain will equilibrate to the required distribution [5]. They have been successfully applied in hydrological modelling by Kuczera and Parent [6].

The draws from the Markov chain are accomplished through variants of the pleasingly simple Metropolis Hastings formulae, involving proposals of candidate values through a proposal function and rejection/acceptance steps. This proposal function is constructed such that it implicitly defines the required conditional distributions, along with satisfying the necessary Markov chain conditions. At any time, it describes the current knowledge regarding parameter distributions, given initial knowledge and information from prior runs. By equating the posterior with Bayes' rule, it also allows potential for converging upon the "true" input distributions through incorporation of learnt information [7].

Bayes' rule is derived from basic axioms of probability. In the context of this work, it is best viewed in terms of updating belief in a hypothesis H given new evidence D. A posterior belief $P(H|D)$, giving the probability of hypothesis H after considering the effect of new data is calculated by multiplying the prior belief $P(H)$ by the likelihood $P(D|H)$ that D will occur if H is true. There is no fundamental distinction between observable quantities and parametric inputs to a model; both can be considered to be random quantities. The theorem can be written as follows,

$$p(H|D) = \frac{p(H)p(D|H)}{p(D)}. \tag{1}$$

A first quantification of $P(H)$ is provided before any data is gathered; this is the prior probability of H. In the context of calibrating a physically based model, the "hypotheses" are the parameter value probability distributions adopted before a simulation commences. These are generally determined subjectively in terms of prior beliefs or knowledge, such as what are realistic ranges of the parameters from previous knowledge of their properties and of the specific scenario situation. Physical constraints, such as non-negativity of concentrations, are also generally included.

In the case of continuous problems, the hypotheses become one continuous parametric distribution. To avoid confusion with the discrete case, this will be denoted by θ, and the outcomes (data) by y. These could be scalars or vectors. Prior beliefs are specified as a probability density function $p(\theta)$, while the outcomes conditional on the hypotheses are the conditional density $p(y|\theta)$, often referred to as the likelihood function [2]. This prior and conditional density fully specify the joint density $p(\theta, y)$ over all hypotheses and data,

$$p(\theta, y) = p(\theta)p(y|\theta) \tag{2}$$

The marginal distribution $p(y)$ of y can be calculated from this joint distribution by integrating over θ,

$$p(y) = \int_\theta p(\theta, y)d\theta = \int_\theta p(\theta)p(y|\theta)d\theta. \tag{3}$$

The posterior probability distribution is therefore given by

$$p(\theta, y) = \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)d\theta}. \tag{4}$$

This is the continuous form of Bayes' Theorem. The denominator, or marginal probability, is easily calculated by recognising that it can, given all other quantities, be considered as a normalising constant.

4. CALIBRATION METHODOLOGIES

The Monte Carlo routines above are suitable for both sensitivity and uncertainty analysis, as they preserve distributional information, and yield information on the total parameter and output spaces. However, in calibration, one is usually interested in locating only a

limited part of the above distributions: generally the global optimum (given appropriate parameter constraints), or, where data error, model structure error, or multiple objectives prevent a single optimum being achievable or meaningful, sets of acceptable local or Pareto-optimal parameters. While full explorations combined with additional information can produce information on "optimal" regions of this space, they are rarely an efficient means of doing so. Therefore, extra routines suitable for calibration have been included in the framework. These are: the Levenburg Marquadt method [8] and the SCEA (Shuffled Complex Evolution Algorithm) [9], neither of which will be considered further in this paper, and finally the addition of simulated annealing to the Markov chain Monte Carlo scheme.

Simulated annealing is a heuristic search procedure based on the metaphor of how annealing works [7]. It aims to reach a global minimum through a procedure that incorporates a decreasing random component to avoid trapping at a local minimum, by allowing a non-improving move to a neighbour with a probability that decreases over time. Since the random component is decreasing, the magnitude of any non-improving change also becomes smaller with time. The rate of this decrease is determined by the cooling schedule, often an exponential decay (in keeping with the thermodynamic metaphor).

To calibrate a model effectively, the "objective" of the calibration must be specified. An objective may be singular, or include several independent criteria that may need to be traded off against each other. They are very application, as well as model, specific, but generally include measures of fit applied to the model output against observed data, and often criteria aimed at minimising risks or costs (economic, environmental, etc). If a model is manually calibrated, the objective may be stated qualitatively: fits may be obtained by eye and intuition then play a part in choosing appropriate calibrated parameter sets. For automated calibration, an 'objective function" or functions giving a mathematical definition of how good a solution is must be formally specified.

Multiple objectives can arise from multiple types of output, emphasising different aspects of model performance, and also from time or space series of one quantity. The latter is generally compacted into one measure through application of a norm (such as least squares). However, different measures of fit will favour different aspects of a series [10]. For example, fits to the variation of stream flow over time may favour the overall water balance, overall shape, or weight the calibration to good agreement of low flow or peak flows. The chosen measure of fit may vary according to the modelling task; flood management may require good estimates of peak flows and overall shape, whereas low flows may be of more important for agricultural management.

One approach is to aggregate the multiple objectives into one single objective function, and optimise to the single-valued best fit. The result is then strongly dependent on the aggregation, or weighting of the objectives. An increasingly common alternative is to employ the concept of Pareto optimality [11]. A set of parameters is said to be Pareto optimal if an improvement in any one criteria will lead to another criteria being degraded; no criteria dominates. This concept does not give a single solution, but rather a set of solutions called the Pareto optimal set. Parameter sets corresponding to these solutions are called non-dominated, and give a visual trade-off between competing objectives. The

user can then choose a solution according to his or her preference. However, in a guided calibration, or to explore posterior distributions conditional upon data, optimality must still be reduced to one measure, perhaps through a normalised weighting of the objectives. One popular means of achieving this is through calculating the coefficient of determination R^2 (widely known in the hydrological literature as the Nash-Sutcliffe efficiency criterion), given by

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{5}$$

where y_i is the simulated value, x_i is measured value, $\overline{x}$ is the mean of the measured values and n is the number of samples. This is the measure used in the forthcoming application, although a variety of approaches are being explored and added to the toolbox capabilities.

5. APPLICATION

To test performance of the integrated model stochastic framework, Markov chain Monte Carlo methods were applied to a model application utilising data from the Kennet catchment in southern England. This is a groundwater-dominated catchment draining an area of 1164 km^2, with a chalk aquifer supplying approximately 95% of its water. As it has been a focus of a variety of water quality and ecological concerns, there is a relatively large amount of data available to compare model response against.

An initial goal was to provide the INCA model with an automated calibration routine in place of the manual calibration procedure used previously, with the aim of enhancing the reliability of calibrated parameters due to a more exhaustive exploration of the parameter space and shortening implementation time. A second purpose was to highlight the most significant parameters for such a calibration, and to identify areas in which model structure could be strengthened. Thirdly, the reliance of results on differing calibration criteria is being investigated.

A one-year simulation period was chosen, from 1^{st} January 1998 to 31^{st} December 1998. Daily precipitation and air temperature were provided, and hydrologically effective rainfall along with soil moisture deficits obtained from MORECS (the U. K. Met Office Rainfall and Evaporation Calculation System). This extracts the relevant quantities from atmospheric data using a Penman-Monteith type routine. Geospatial and other relevant information (e.g. farming practices, proportional land use, base flow indexes, dry deposition data) were obtained from national databases and other sources where available. A hand-calibrated parameter set provided other parameters needed by the model, and was also used as the starting state for the Markov chain Monte Carlo simulations.

To address the first two purposes of the performance test, twenty-six parameters were varied within ranges constrained by existing calibration guidelines and literature values. These parameters are shown in Table 1.

Two scenarios were considered: the first treated the 26 parameters as spatially homogeneous in both land and river phases, and the second allowed for heterogeneity. This heterogeneous scenario had 6 unique land types, and divided the river Kennet into 25 “reaches”, or contiguous lengths. Each reach is then associated with a subcatchment, and

Table 1. Calibration/Sensitivity Parameters examined in Markov Chain Monte Carlo Kennet simulations, January-December 1998.

Name	Units	Distribution	Minimum	Maximum
initial soil flow	$m^3 s^{-1}$	land use	0.0	1.0
initial groundwater flow	$m^3 s^{-1}$	land use	0.0	0.1
initial soilwater nitrate	$mg.l^{-1}$	land use	0.0	10.0
initial groundwater nitrate	$mg.l^{-1}$	land use	0.0	8.0
initial soilwater ammonium	$mg.l^{-1}$	land use	0.0	2.0
initial groundwater ammonium	$mg.l^{-1}$	land use	0.0	1.0
initial soil drainage volume	m^3	land use	10^5	2×10^7
initial ground drainage volume	m^3	land use	10^6	10^8
initial in-stream flow rate	$m^3 s^{-1}$	top reach	0.0	2.0
initial in-stream nitrate	$mg.l^{-1}$	top reach	0.0	10.0
initial in-stream ammonium	$mg.l^{-1}$	top reach	0.0	2.0
denitrification rate	$m.day^{-1}$	land use	0.01	19.0
nitrogen fixation	$kg.ha^{-1}day^{-1}$	land use	0.0	0.0001
plant nitrate uptake	$m.day^{-1}$	land use	0.0	162.0
nitrification rate	$m.day^{-1}$	land use	1.0	54.0
mineralisation	$kg.ha^{-1}day^{-1}$	land use	1.0	292.0
immobilisation rate	$m.day^{-1}$	land use	0.0	1.0
ammonium addition rate	$kg.ha^{-1}day^{-1}$	land use	0.0	100.0
plant ammonium uptake	$m.day^{-1}$	land use	0.0	162.0
Reactive zone residence time	$days$	land use	0.5	5.0
Groundwater residence time	$days$	land use	10.0	200.0
Maximum soil water retention†	m	land use	0.0	1.0
velocity flow a parameter (Qa)	-	by reach	0.001	0.2
velocity flow b parameter Qb	-	by reach	0.3	0.99
Denitrification rate	day^{-1}	by reach	0.04	0.09
Nitrification rate	day^{-1}	by reach	0.1	5.0

this association allows for the coupling of in-river and land processes (for more details, see [12]). This discretisation in space resulted in 215 unique parameters. The objectives for calibration were taken to be least squares fits to in-river flow, nitrate, and ammonium concentrations, with the Nash-Sutcliffe criterion providing a means to weight these appropriately for heuristic guidance.

The first implementation of the tool was only partially successful, as the model was not robust over the entire parametric space. Such problems are common in a first application of a full sensitivity analysis to a model, as an automated routine is likely to discover (by brute force) subtle instabilities in a model formulation, and also extract any parameter constraints that have not been explicitly specified. This interfered with the ability of the heuristically guided Markov chain Monte Carlo and calibration methodologies to explore model input and output distributions.

When tested on sub-sets of parameters where the model was robust, Markov chain

Monte Carlo analysis results were encouraging, and the annealing approach to calibration substantially reduced run-time needed to locate optima. The hand calibrated and automated river flow optima were almost identical; for this particular only two parameters were significant. However, in most reaches the automated calibration improved upon the nitrate concentration optima by a factor of 3 to 6 (using a least squares objective function). An example, showing data, hand and automated calibrations is shown in Figure 1. Ammonium is disregarded for comparison purposes, as the hand-calibration did not seek to optimise this. Figure 2 shows plots of the most sensitive parameter against the

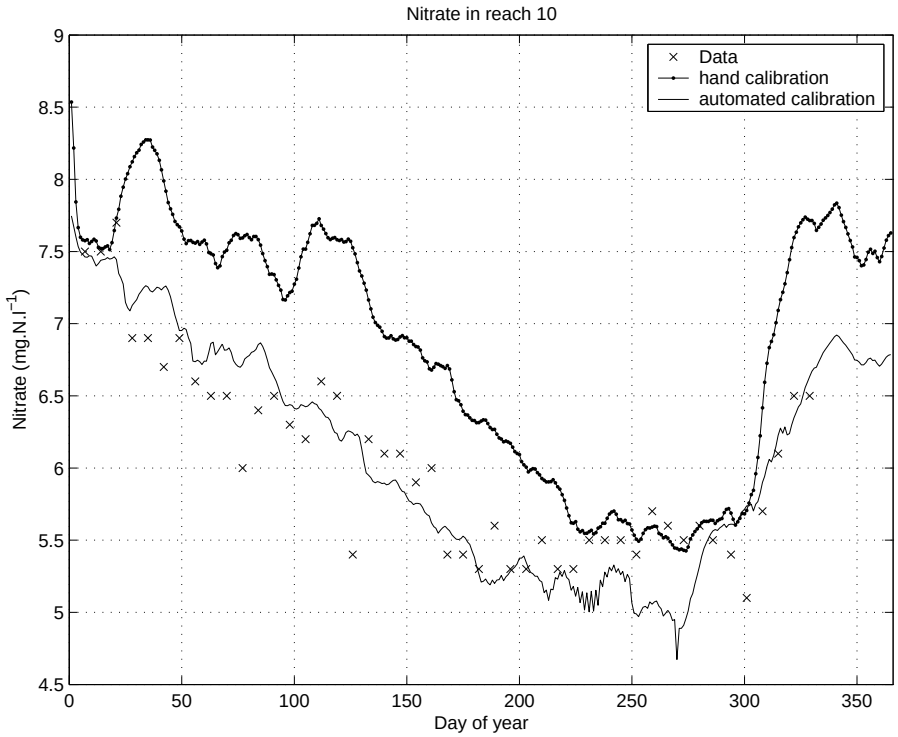


Figure 1. "Hand" versus automated calibration for nitrate in reach 10.

least squares nitrate objective function, for several reaches along the river. Figure 3 shows the combined posterior distribution of two parameters; groundwater drainage volume and initial groundwater nitrate concentration, conditioned on a data fit measure derived from least squares fits to measured in-river nitrate concentrations and flows. Equal importance was given to each reach, and the Nash-Sutcliffe efficiency measure used for normalisation purposes. A flat prior was used with a simple random walk Metropolis proposal function, and the chain was visited two million times. Quantitative convergence diagnostics have not yet been included, partly due to the debates surrounding the issue [**?**]. However, a preliminary diagnostic, dividing the chain into four sub-intervals, showed almost identical distributions. This, along with the long length of the chain, provides a strong argument for acceptance.

The influence of differing data fit measures on the above posterior distribution has also been examined. However, the presence of small instabilities in model responses for all the

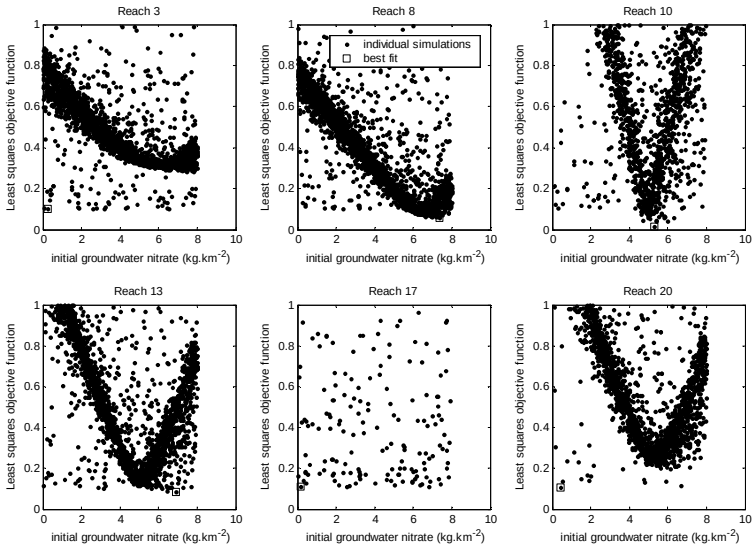


Figure 2. Response of in-river nitrate objective function against initial groundwater nitrate values over space.

posteriors somewhat obscured the biases caused by differing optima. This is continuing to be addressed as the model is refined.

6. CONCLUSIONS

A framework for stochastic analysis of catchment scale modelling scenarios, utilising Markov chain Monte Carlo along with other methodologies, has been developed and integrated with an existing nitrogen in catchments model (INCA).

It has become evident that, while MCMC provides an efficient means of investigating various conditional distributions and model responses, its relevance as an aid to developing models is limited unless it is complemented with cruder, set search methods and tools to identify structural problems and problematic parameter sets. This is a consequence of its reliance, in this context, on heuristic searches, which demand a certain degree of smoothness within the explored response surfaces.

The framework presented here is being extended to include a range of multi-variate analysis tools to investigate and isolate non-viable parameter combinations and structural issues. It is our view that this will be a useful, and arguably necessary, aid to any environmental model that has not already undergone substantial automated testing in its current form.

Implementation of the Markov chains produced successful results in parameter regions with stable model response, and has demonstrated the ability of the Metropolis

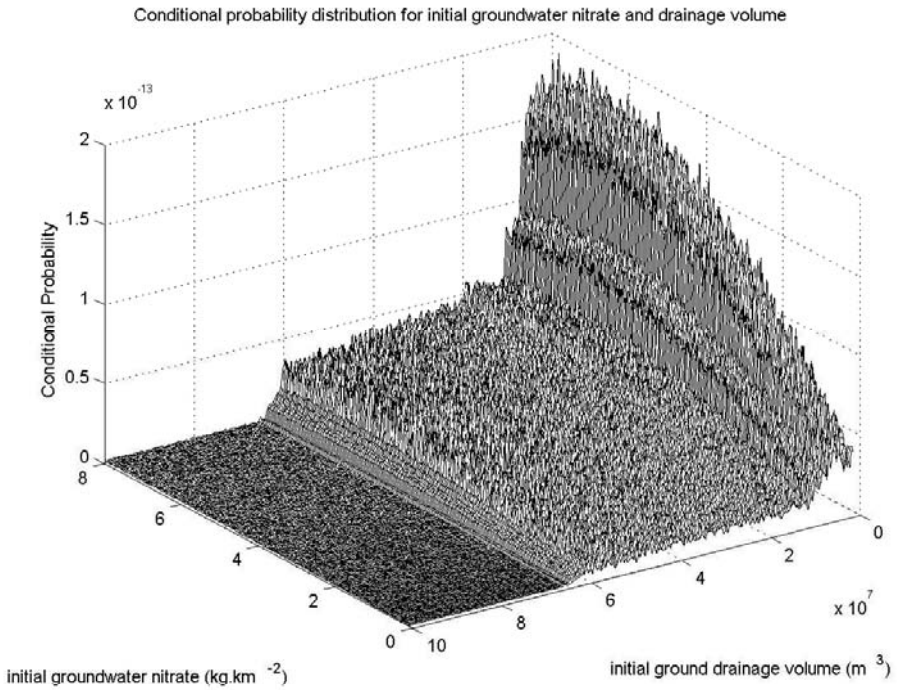


Figure 3. Posterior distribution of initial groundwater nitrate and drainage volume, conditional upon normalised least squares in-river flow and nitrate levels.

Hastings scheme to efficiently recover conditional distributions given appropriate data. To further test its efficiency against other methodologies, and parameterise its heuristics appropriately, model robustness is being tested and improved, and further information on parameter interactions investigated.

Further work is seeking to extend the robustness of the INCA model under automated calibration, through model component changes and further constraints upon parameters, and investigations into how conditional distributions are affected by aspects of model response and the optimality criteria imposed upon an analysis are ongoing. Proposal functions constructed with the aim of allowing sampling to be efficiently weighted towards subsets of the distribution are also being examined.

ACKNOWLEDGMENTS

This work was supported by two United Kingdom bodies: the National Environment Research Council, as part of its safety lowland catchment research programme (LOCAR), and the Environment Agency. Their support is gratefully acknowledged.

REFERENCES

1. A. J. Wade, P. Durand, V. Beaujouan, W. W. Wessel, K. J. Raat, P. G. Whitehead, D. Butterfield, K. Rankinen, and A. Lepisto. A nitrogen model for European catchments: INCA, new model structure and equations. *Hydrol. Earth Syst. Sci.*, 6:559–582, 2002.

2. N. McIntyre, B. M. Jackson, H. S. Wheater, and S. Chapra. Numerical efficiency in Monte Carlo simulations- a case study of a river thermodynamic model. *Journal of Environmental Engineering, ASCE*, (in press).
3. G. S. Fishman. *Monte Carlo: concepts, algorithms, and applications*. Springer Verlag, New York, 1996.
4. R. M. Neal. Probabilistic inference using Markov Chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.
5. W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.
6. G. Kuczera and E. Parent. Monte-Carlo assessment of parameter uncertainty in conceptual catchment models. *Journal of Hydrology*, 211 (1-4):69–85, 1998.
7. C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 1999.
8. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.
9. Q. Duan, S. Soorooshian, and V. Gupta. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, 28(4):1015–1031, 1992.
10. K. J. Beven. Calibration, validation and equifinality in hydrological modelling. In M. G. Anderson and P. D. Bates, editors, *Validation in Hydrological Modelling*. John Wiley and Sons, Chichester, U.K., 2001.
11. K. Miettinen. *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, 1999.
12. P. G. Whitehead, E. J. Wilson, and D. Butterfield. A semi-distributed Integrated Nitrogen model for multiple source assessment in CAtchments (INCA): Part I-model structure and process equations. *Sci. Tot. Environ.*, 210/211:547–558, 1998.

Validating Mathematical Models of Biological Systems: Application of the Concordance Correlation Coefficient

N. R. St-Pierre

The Ohio State University, 2029 Fyffe Rd., Columbus, OH-43210, USA
E-mail: St-Pierre.8@osu.edu

Abstract: The National Research Council is charged with producing mathematical models of nutrient requirements of domestic animals. In ruminants, protein supply is derived from two sources: a fraction of the feed protein unaltered by ruminal fermentation, and microbial protein (MiN) synthesized by the ruminal micro-flora. Measurements of MiN rely on surgically altered animals and inert markers. The prediction of MiN is based on total digestible nutrients, a function of the uncertain composition of feedstuffs. Both observed and predicted MiNs have errors from measurements, parameter estimates, and structural forms. The question is whether predicted MiN can replace measured values when estimating requirements. The concordance correlation coefficient (ρ^c) has been suggested as an omnibus statistic to jointly assess precision and accuracy. Application to a dataset of 256 measured and predicted values of MiN from 56 published studies shows that predictions and measurements are concordant ($\rho^c = 0.476$), have small scale shift (1.54) and location shift (-0.02), and are accurate (0.913) but that they lack precision (0.522). The deviance (0.573) is composed of a small bias (0.0003), a small scale shift (0.095), and a large imprecision (0.479). Little gain in model precision can be expected until more precise methods of measurements are found.

Keywords: concordance correlation coefficient, precision, accuracy, nutrient requirements

1. INTRODUCTION

Mathematical models are now frequently used to quantify complex biological systems [1, 2]. The validation of such models is done by comparing model predictions to observed data. Various statistical methods have been suggested and used to assess a model's validity: the Pearson correlation coefficient, the paired t-test, the least-square analysis of slope (=1) and intercept (=0), and the coefficient of variation or the intraclass correlation coefficient. None of these can completely assess the desired reproducibility characteristics. The Pearson correlation coefficient only measures precision of a linear relationship, not accuracy. Both the paired t-test and least squares analysis can falsely reject (accept) the hypothesis of high agreement when the residual error is small (large). The coefficient of variation and the intraclass correlation coefficient assume a dependent and an independent variable. More importantly, they fail to recognize the duality (interchangeability) of predictions with observations. Both are mathematical transforms of measurements. Both have random errors from measurements and parameter estimates. And both have structural errors due to the simplification of the complex real world. The relevant question is not whether a model predicts observed data but whether the model and the observation method measure the same

thing, whether the methods agree and how good is the agreement. This requires a joint assessment of precision and accuracy.

The Committee on Animal Nutrition of the National Research Council (NRC, [3]) is charged with producing tables of nutrient requirements of various classes of animals. Nutrient requirements are expressed in the form of computerized mathematical models. In a recent publication, the NRC [3] produced a new model for estimating the nutritional requirements of dairy cattle. A key step in the calculation of protein and amino acid requirements is the estimation of the amount of bacterial protein synthesized in the rumen. In ruminants, the net supply of protein and amino acids is derived from two separate fractions: a variable portion of the feed protein not broken down by the ruminal micro-flora passes to the duodenum (small intestine) where it can be digested and absorbed by the animal. The second portion consists of microbial protein synthesized by the ruminal micro-flora using carbon skeletons, ATP, ammonia, amino acids, and short peptides. The quantification of the net supply from each process is very important to the optimal feeding of ruminant animals and for reducing their environment impact from N excretion [4]. The measurements of microbial and undegraded feed protein to the duodenum must rely on surgically altered animals and inert markers [5]. Thus, the measurements of microbial protein (MiN) and non-ammonia-non-microbial protein flows (NANMN) to the duodenum are subject to substantial errors of measurements, plus structural errors (i.e., the non-digestible markers are not perfect markers) and possibly errors in parameter estimates. The prediction of MiN is based on total digestible nutrient intake (TDN) which is a function of the (uncertain) chemical composition of the feedstuffs and their (uncertain) bio-availabilities. Thus, both observed and predicted MiN and NANMN have errors from measurements, parameter estimates, and structural forms. This situation, where predictions and observations are interchangeable is very frequent in biology. The question is whether we can use predictions of MiN and NANMN to replace measured values when estimating nutrient requirements.

In this paper, we first review the model used by NRC [3] to predict MiN in dairy cattle and the proper statistical model linking predictions to observations. Results from applying traditional methods of model validation are presented followed by the application of the concordance correlation coefficient (CCC) of Lin [6].

2. METHODOLOGY

2.1 Prediction of microbial protein synthesis by the National Research Council

In high producing ruminants, microbial protein synthesis is primarily determined by the availability of energy to the micro-organisms [7]. Although various expressions of available energy have been proposed and used to express the availability of feed energy for microbial growth, the total digestible nutrient (TDN) system is still favored in the U.S. due to the considerable literature reporting actual measurements in lactating and non-lactating animals. The measurement of TDN is a tedious process and requires urine and fecal collection in a digestibility study performed over several days (generally 5-7) with multiple animals. The TDN of a feed can also be estimated from its proximate composition using the following system of equations [8]:

$$TDN = tdNFC + tdCP + tdFat + dNDF - 7, \quad (1)$$

$$tdNFC = 0.98 \times (100 - NDFn - CP - Fat - Ash) \times PAF,$$

$$tdCP = EXP(-1.2 \times (ADFIP / CP)) \times CP,$$

$$tdFat = (Fat - 1) \times 2.25,$$

$$dNDF = 0.75 \times (NDFn - L) \times [1 - (L/NDFn)^{0.67}],$$

$$NDFn = NDF - NDFIP,$$

where TDN is the estimated total digestible nutrients (%), tdNFC is true digestible non-fiber carbohydrates (%), tdCP is true digestible crude protein (%), tdFat is true digestible fat (%), dNDF is digestible neural detergent fiber (%), NDFn is NDF corrected for NDFIP (%), CP is the crude protein content (%), Fat is the fat content (%), Ash is the ash content (%), PAF is a processing adjustment factor, ADFIP is the acid detergent insoluble N x 6.25 (%), NDFIP is the neutral detergent insoluble N x 6.25 (%), L is the lignin content (%), and NDF is the neutral detergent content (%) of a given feedstuffs. Although the proximate composition (CP, Fat, Ash, etc.) is determined analytically in a laboratory, this is not done without analytical errors, which typically range between 2 and 10% of the true mean depending on the assay and feedstuff involved. Digestibility coefficients (e.g., 0.98, 0.75) are estimates subject to errors. Also, although the structure of the set of equations in (1) was derived mechanistically, it is nevertheless a simplification to the true, unknown, and far more complex system in nature. Thus, TDN values estimated using the system of equations in (1) are subject to measurement errors (feed composition), parameters in the equation are estimates (thus subject to errors), and the functional form itself is an approximation to the complex world.

In NRC [3], estimated TDN values from the set of equations in (1) are used to estimate MiN according to the following equation:

$$MiN = 130 \times TDN, \quad (2)$$

where MiN is net microbial protein synthesis (expressed in g of N/d). The coefficient 130 was estimated using an independent set of experimental data where both TDN and MiN had been measured. Clearly, it is an estimate also subject to error. By combining Eqs. (1) and (2), the NRC calculates the predicted MiN resulting from a given diet. This prediction is subject to measurement errors (feed composition), as well as errors in estimates of parameters (coefficients in Eqs. (1) and (2)), and errors in functional forms used.

Measurements of MiN are not without errors. Various experimental methods have been suggested in the scientific literature. All have limitations [7]. The prevailing method involves the marking of feeds and fluids with three indigestible markers each associating more predominantly with one of the three major digesta fractions (large particles, small particles, and fluid). Animals must be surgically altered with a large rumen cannula for the infusion or dosage of marker, and a duodenal cannula for sampling digesta leaving the stomach. Multiple samples are taken over time and the concentration of the three markers is then determined in a laboratory for each sample. Assuming first order, steady-state kinetics, forestomach digestibility of feed components can be calculated as well as flow of MiN [9] based on a marker of microbial protein (e.g., purines). It is clear that measured MiN are subject to considerable errors resulting from true measurement errors (concentrations of indigestible markers, concentration of microbial marker) as well as errors in parameter estimates, and error in the functional form (first-order, steady-state kinetics).

In this context, observations and predictions play a symmetric role because they are both functional transforms of other variables. This situation is actually quite frequent when modeling biological systems. The symmetric role of observations and predictions, however, has been largely ignored when models are being validated

2.2 Statistical Model

The following model, which naturally models comparison studies when both observations and predictions are subject to multiple errors, is commonly known as errors-in-variables regression [10, 11]:

$$\begin{aligned} X_i &= \xi_i + \delta_i, \\ Y_i &= \eta_i + \varepsilon_i, \qquad i = 1, \ldots, n, \\ \eta_i &= \alpha + \beta\xi_i, \end{aligned} \tag{3}$$

where X_i is the prediction from the mathematical model and Y_i is the observed value of the i^{th} observation, ξ_i and η_i are the unobserved mean parameters ("true values") of X_i and Y_i respectively, δ_i and ε_i are the errors on the predicted and observed values (generally assumed to be independent, bivariate Gaussian), α is the overall bias of the prediction model, and β is the linear scale difference between the predicted and the observed values. The variance of the two errors, σ^2_δ and σ^2_ε, are the precision parameters for the predictions and observations, respectively. With known or estimable σ^2_δ and σ^2_ε (or more accurately, an unbiased estimate of $\lambda = \sigma^2_\delta / \sigma^2_\varepsilon$), the maximum likelihood estimate of β is [11]:

$$\beta = \frac{S_{YY} - \lambda S_{XX} + ((S_{YY} - \lambda S_{XX})^2 + 4 \lambda S^2_{XY})^{1/2}}{2 S_{XY}}. \tag{4}$$

An estimate of σ^2_ε can be calculated from experimental data. Because of the nonlinearity of the system of equations in (1), an analytical estimate of σ^2_δ does not exist. Numerical methods could possibly be used but would require knowledge about the variances and covariances of all random variables in the equation. This information is currently not available.

2.3 Concordance correlation coefficient

Lin [6] proposed a statistic termed the concordance correlation coefficient (CCC) to evaluate the agreement (reproducibility) between two readings. In short, the degree of concordance between pairs of sample (Y_{i1}, Y_{i2}), i = 1, 2, …, n, can be characterized by the expected value of the squared difference, i.e.,

$$E(Y_1 - Y_2)^2 = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2(1 - \rho)\,\sigma_1\sigma_2, \tag{5}$$

where ρ is the Pearson correlation coefficient. This expectation also represents the expected squared perpendicular deviation from the 45° line, multiplied by 2. Standardizing both sides, we get:

$$\frac{E(Y_1 - Y_2)^2}{2\,\sigma_1\sigma_2} = \frac{(\mu_1 - \mu_2)^2}{2\,\sigma_1\sigma_2} + \frac{(\sigma_1 - \sigma_2)^2}{2\,\sigma_1\sigma_2} + (1 - \rho), \tag{6}$$

which has a sample equivalent:

$$\frac{E(Y_1 - Y_2)^2}{(n-1)\ 2\ s_1 s_2} = \frac{(Y_1 - Y_2)^2}{(n-1)\ 2\ s_1 s_2} + \frac{(s_1 - s_2)^2}{(n-1)\ 2\ s_1 s_2} + (1 - r). \quad (7)$$

In (7), a form that has been called deviance analysis, the total deviance, represented by the left-hand side is partitioned into three right-hand side components: bias (first term), scale difference (second term), and imprecision (third term). The deviance is equal to zero when all (non-negative) terms on the right-hand side are exactly zero, i.e., when the two means are equal, the two variances are equal, and the correlation is equal to 1.

The CCC is defined as follows:

$$\rho^c = 1 - \{E(Y_1-Y_2)^2 / E[(Y_1-Y_2) \mid Y_1, Y_2 \text{ are uncorrelated}]\}, \quad (8)$$

$$\rho^c = 2\ \sigma_{12} / [\sigma^2_1 + \sigma^2_2 + (\mu_1 - \mu_2)^2], \quad (9)$$

$$\rho^c = \rho_{12}\, \chi_{12}, \quad (10)$$

where $\mu_1 = E(Y_1)$, $\mu_2 = E(Y_2)$, $\sigma^2_1 = Var(Y_1)$, $\sigma^2_2 = Var(Y_2)$, and $\sigma_{12} = Cov(Y_1, Y_2) = \sigma_1\ \sigma_2\ \rho_{12}$. The CCC is a product of two components: precision (ρ_{12}) and accuracy (χ_{12}), where $\chi_{12} = 2\ \sigma_1\ \sigma_2 / [\sigma^2_1 + \sigma^2_2 + (\mu_1 - \mu_2)^2] = [(\nu_{12} + 1/\nu_{12} + u^2_{12}) / 2]^{-1}$, with $\nu_{12} = \sigma_1 / \sigma_2$ representing scale shift, and $u_{12} = (\mu_1 - \mu_2) / (\sigma_1\ \sigma_2)^{1/2}$ representing location shift relative to the scale. The CCC is an omnibus statistic used to test simultaneously and jointly for accuracy and precision.

2.4 Dataset

The data used are described at length in the NRC publication [3]. In short, feed composition and measured MiN were gathered from 56 published, peer-reviewed studies of which 27 involved growing cattle and 29, lactating dairy cows. In total, the dataset comprised 256 records of observed MiN (oMiN, g/d) and predicted MiN (pMiN, g/d).

3. RESULTS

3.1 Pearson correlation

The Pearson correlation, which measures the degree of linear association (relationship) between two random variables has been used for comparing mathematical model predictions to observed values. In our application, this correlation is equal to: $r_{pMiN,\ oMiN} = 0.52$, $P < 0.0001$. This statistic shows that oMiN and pMiN have a significant association. The Pearson correlation, however, is invariant to location and scale. Agreement is a much more stringent concept than correlation because both the scale of the measurements and the slopes are important. Also, observations are not random samples from a population (i.e., the sample of observed and predicted values was not drawn at random from the population of all cows in the world). Thus, the Pearson correlation coefficient fails to determine whether pMiN and oMiN are equivalent.

3.2 Paired t-test

Applying the paired t-test on the data (mean oMiN = 244.91, mean pMiN = 246.36, $SE_{diff} = 4.16$, $t_{255} = -0.35$, $P = 0.73$), we conclude that there is no significant difference between the mean oMiN and the mean pMiN. This test provides information only for the overall bias (location shift). Because of its structure, the t-test can falsely reject the null hypothesis of

high agreement when the residual error is small. That is, the larger the precision, the more likely you are to conclude that the two methods are not equivalent.

3.3 Least-squares analysis

The linear regression of oMiN on pMiN is presented in Figure 1. The model:

$$oMiN = B_0 + B_1\, pMiN + e \qquad (11)$$

is theoretically incorrect because both oMiN and pMiN have errors. Under least-squares analysis, the null hypothesis is that the two methods are concordant. Thus, small datasets will generally lack power resulting in the conclusion that the two methods are concordant. Likewise, large datasets will result in rejecting the null hypotheses for the intercept ($B_0 = 0$) and the slope ($B_1 = 1$) when differences are relatively trivial. This is what occurs with the dataset at hand where the two null hypotheses are rejected. A casual inspection of the regression line in Figure 1 reveals the trivial difference between the regression line and the line of unity when the spread of the data points from either line is considered.

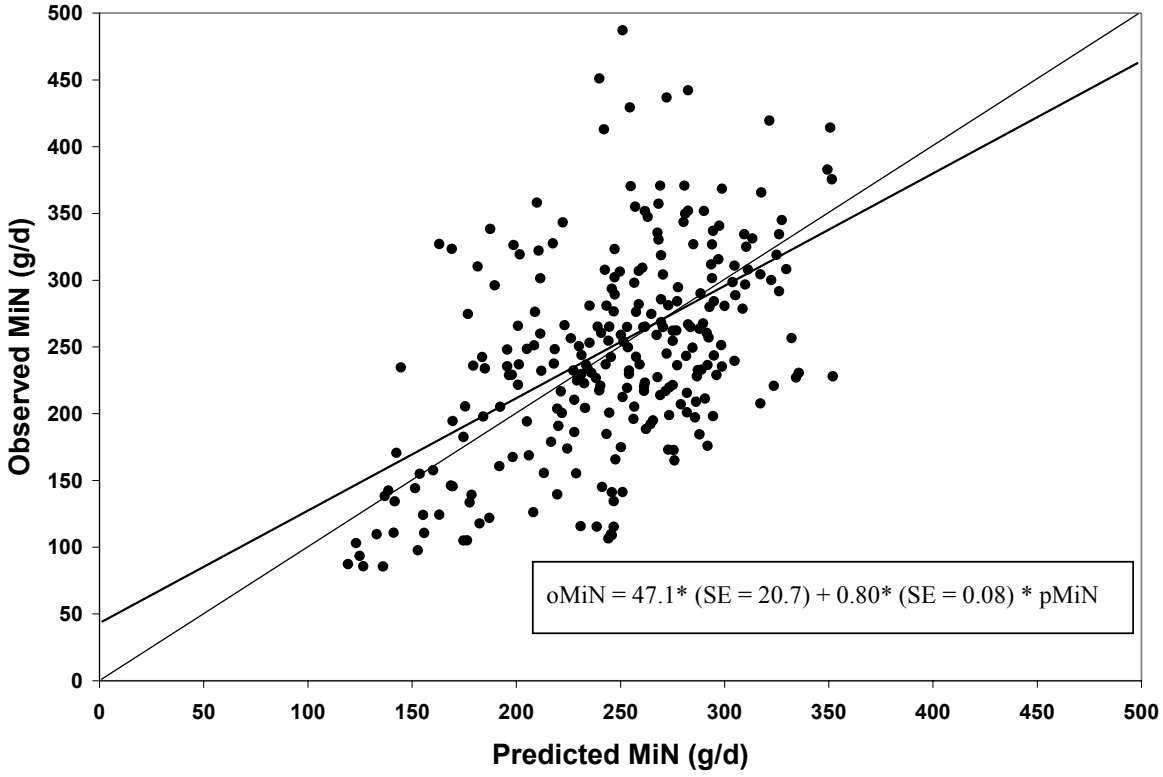


Figure 1. Linear regression of observed microbial flow to the duodenum (oMiN) on predicted microbial N flow (pMiN) using the National Research Council model.

This is quite clear when the differences between oMiN and pMiN are plotted against pMiN as in Figure 2. This plot, however, raises the legitimate question as to which variable should be used on the X-axis?

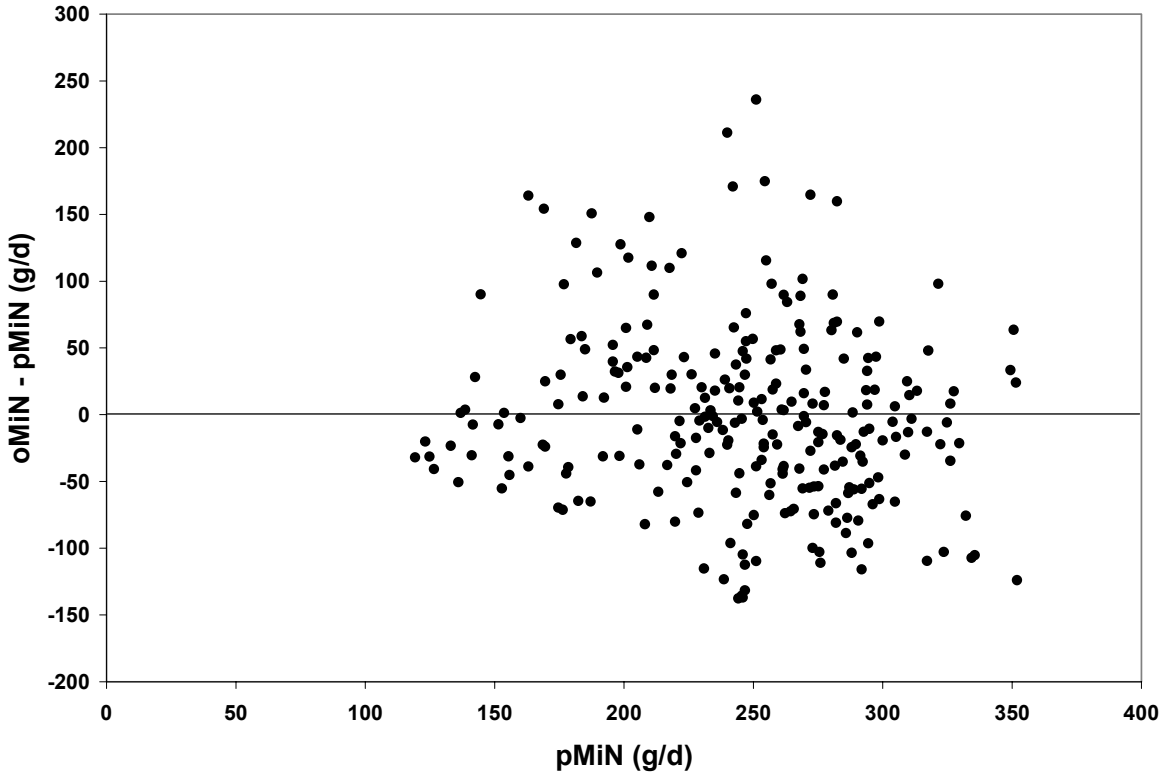


Figure 2. Plot of residuals vs. predicted microbial N flow to the duodenum (pMiN)) using the National Research Council model to calculate predicted microbial N flow (pMiN).

Recall that both oMiN and pMiN are measurement with errors. In Figure 2, pMiN was chosen on the X-axis because this is the correct variable to use in residual plots when the independent variable is assumed to be errorless, as in the linear regression paradigm [12]. Because of the duality of oMiN and pMiN, one could have chosen oMiN for the X-axis, resulting in a different conclusion regarding the presence or absence of bias (Figure 3).

Recognizing this problem, Altman and Bland [13] suggested using the mean of oMiN and pMiN for the X-axis (Figure 4). In fact, this is the correct axis if, and only if the precisions of both methods are equal (i.e., when $\sigma^2_\delta = \sigma^2_\varepsilon$, or simply that $\lambda = 1$). In the data at hand, however, the precision of pMiN is unknown. Thus the correct residual plot lies somewhere between the two extremes presented in Figures 2 and 3. Unless a satisfactory estimator for σ^2_δ can be identified, residual plots will invariably lead to the paradox depicted in Figures 2, 3, and 4, where one cannot decide whether a linear bias is present or not.

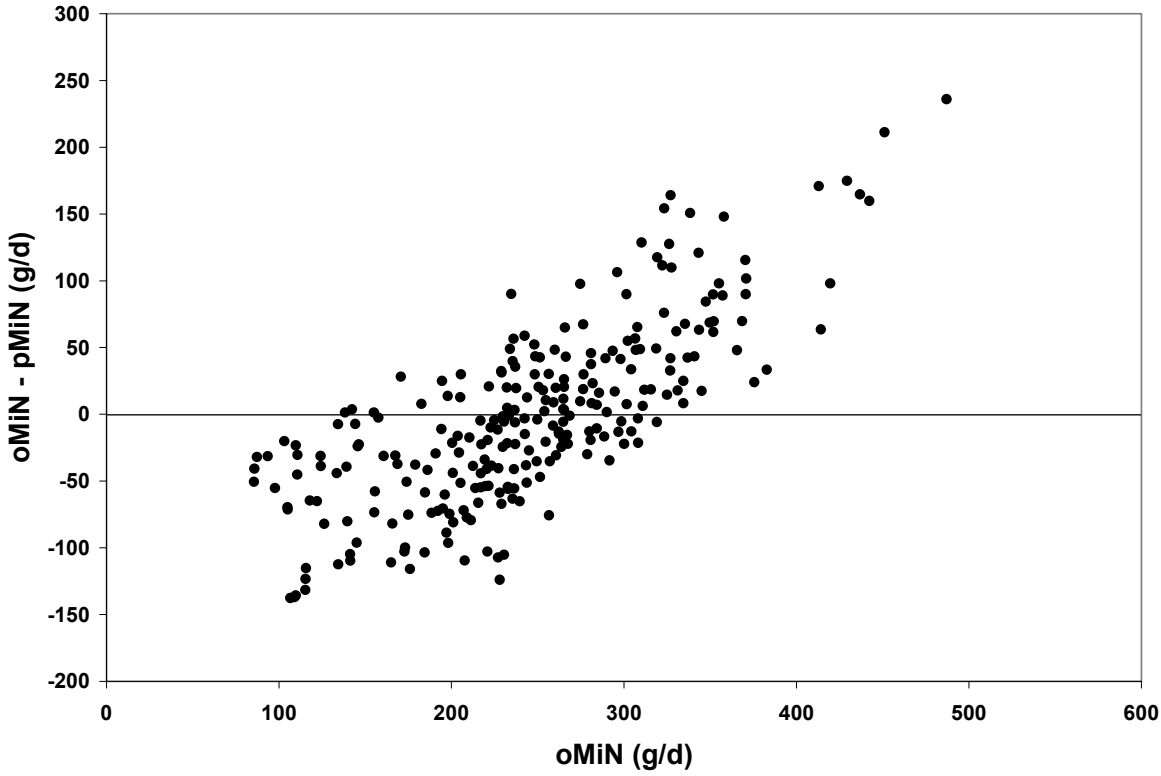


Figure 3. Plot of residuals vs. observed microbial N flow to the duodenum (pMiN) using the National Research Council model to calculate predicted microbial N flow (pMiN).

3.3 Deviance analysis

Application of equation (7) using the following estimates: $s_1 = 50.16$, $s_2 = 77.23$, $s_{12} = 2020.2$, mean $(Y_1) = 246.4$, mean $(Y_2) = 244.9$, and $r = 0.522$ (where the subscript 1 refers to pMiN and the subscript 2, to oMiN) results in the following:

$$\begin{array}{lllll} 0.5733 & = & 0.0003 + & 0.0945 + & 0.4785 \\ \text{Deviance} & = & \text{Bias} + & \text{Scale difference} + & \text{Imprecision} \end{array} \quad (12)$$

The deviance is composed of a very small bias (0.0003; or 0.05% of the deviance), a small scale shift (0.0.95; or 16.5% of the deviance), and a large imprecision (0.479; or 83.5% of the deviance). Thus, it is clear that most of the deviance is the result of imprecision. The expression of deviance in (7) is in the form of the mean of squared deviations standardized by the product of standard deviations. The unit for deviance does not correspond to the unit of the physical variables being measured or predicted. Thus, although the method is useful, biologists struggle with the physical interpretation of the analysis. However, biologists are very familiar with the Pearson correlation coefficient, so that the expression of deviance re-scaled with a lower bound of -1 and an upper bound of 1 is certainly appealing. In essence, this is what is accomplished by the CCC.

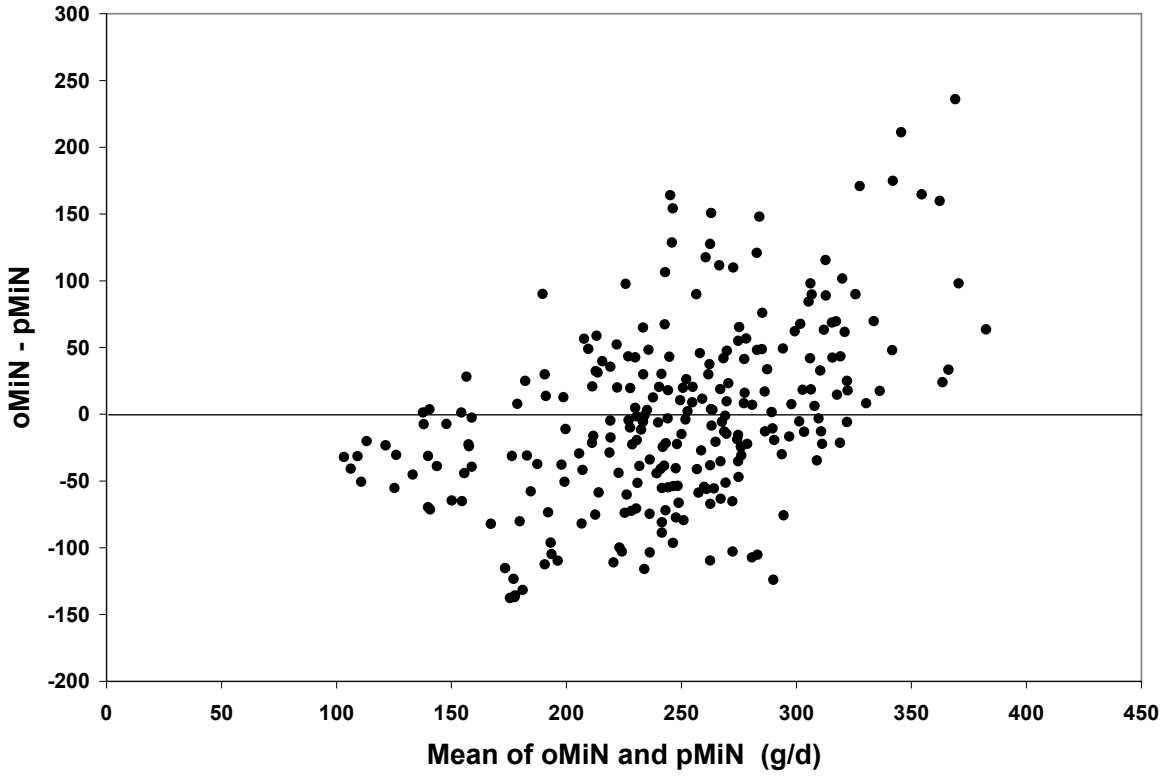


Figure 4. Plot of residuals vs. the mean of observed microbial N flow to the duodenum (pMiN) and predicted microbial N flow (PMiN) using the National Research Council model.

3.3 Concordance correlation coefficient

Application of equation (10) to our dataset results in $\rho^c = 0.476$. Using the inverse hyperbolic tangent transformation (or Z-transformation) suggested by Lin [6], and under the assumption of asymptotic normality, one concludes that predictions and measurements are concordant ($P = 0.22$). The accuracy statistic (χ_{12}) is equal to 0.913, whereas the precision statistic (ρ_{12}) is equal to 0.522. Recalling that $\rho^c = 0.476 = 0.913 \times 0.522$, it becomes evident that precision and not accuracy is the issue. The CCC is equal to 1 when there is no location differential, no scale differential, and perfect correlation between the two variables. It is an omnibus statistic that tests jointly precision and accuracy. In our application, measurements are too imprecise to allow the development of a model with acceptable prediction error. Thus, gains in the prediction of MiN can only be achieved with the development of superior methods of measurements, with much greater precision than the methods currently in use.

4. CONCLUSIONS

The validation of quantitative biological models is not a simple problem. Methods must account for the multiplicity of errors in both the observed and the predicted values. That is, methods must recognize the symmetric role of observations and predictions because both are algebraic transforms of other variables. The CCC shows potential in this regard.

REFERENCES

1. J. France, and J. H. M. Thornley. *Mathematical Models in Agriculture*. Butterworth, London, 1884.
2. J. D. Murray. *Mathematical Biology*. Springer, Berlin. 1993.
3. National Research Council. *Nutrient Requirements of Dairy Cattle.* 7th Ed. National Academy of Sciences, Washington. 2001.
4. N. R. St-Pierre, and C. S. Thraen. Animal grouping strategies, sources of variation, and economic factors affecting nutrient balance on dairy farms. *J. Dairy. Sci.* 82 (Suppl. 2):72-83, 1999.
5. G. J. Faichney. Digesta Flow. In J. M. Forbes, and J. France, editors, *Quantitative Aspects of Ruminant Digestion and Metabolism*, pages 53-86. CABI, Wallingford (UK), 1993.
6. L. I. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255-268, 1989.
7. J. L. Firkins, M. S. Allen, B. S. Oldick, and N. R. St-Pierre. Modeling ruminal digestibility of carbohydrates and microbial protein flow to the duodenum. *J. Dairy Sci.* 81:3350-3369, 1998.
8. W. P. Weiss, H. R. Conrad, and N. R. St-Pierre. A theoretically-based model for predicting total digestible nutrient values of forages and concentrates. *Animal. Feed Sci. Tech.* 39:95-110, 1992.
9. J. France, and R. C. Siddons. Determination of digesta flow by continuous marker infusion. *J. Theoretical Biology* 121:105-119, 1986.
10. G. Casella, and R. L. Berger. *Statistical Inference*. Wadsworth, Pacific Grove, CA, 1990.
11. C. Y. Tan, and B. Iglewicz. Measurement-methods comparisons and linear statistical relationship. *Technometrics* 41:192-201, 1999.
12. N. R. St-Pierre. Reassessment of biases in predicted nitrogen flows to the duodenum by NRC 2001. *J. Dairy Sci.* 86:344-350, 2003.
13. D. G. Altman, and J. M. Bland. Measurement in medicine: the analysis of method comparison studies. *The Statistician* 32:307-317, 1983.

On a Strategy of Reduction of the Lack of Knowledge (LOK) in Model Validation

P. Ladevèze[a]*,G. Puel*[a] *and T. Romeuf*[b]

[a] LMT-Cachan (ENS Cachan/CNRS/Paris 6 University)
61, avenue du Président Wilson 94235 CACHAN Cedex FRANCE;

[b]European Aeronautic Defence and Space Company (EADS Space Transportation)
Route de Verneuil BP 96 78133 LES MUREAUX Cedex FRANCE.

E-mail: pierre.ladeveze@lmt.ens-cachan.fr, guillaume.puel@lmt.ens-cachan.fr,
thierry.romeuf@launchers.eads.net

Abstract: The quantification of the quality of a structural mechanical model remains a major issue today, with the use of an increasing number of methods in order to validate a model in comparison with an experimental reference. This paper presents a new theory based on the concept of Lack of Knowledge combining convex uncertainty models with probabilistic features by introducing for each substructure two bounds of the strain energy as stochastic variables. A general strategy of reduction of the lack of knowledge is discussed and applied to academic as well as industrial cases.

Keywords: Lacks of Knowledge, reduction, model validation, structural dynamics

1. INTRODUCTION

Today, the problem of quantifying the quality of a structural mechanical model remains a major issue. As far as the comparison with an experimental reference is concerned, many approaches can be used to update a deterministic, dynamic model (stiffness, mass and damping) based on free or forced vibration tests [1],[2]. After this process, there may be some phenomena that still cannot be described properly: some uncertainties remain in the material properties, or the model of some parts (e.g. joints) may be simplified. In order to describe these uncertainties, the use of probabilistic methods has become increasingly popular: generally, these methods consist in studying the effects of the uncertainties which affect the input on the variability of the output. This can be done in various ways and has led to major improvements: for example, meta-models have been built by spanning the space of the most influential parameters and applying a specific technique to reduce the computational effort drastically [3],[4].

In [5], we introduced the concept of *Lack of Knowledge* (LOK), which combines convex uncertainty models [6],[7] with probabilistic features. The basic principle consists in globalizing the uncertainties on a substructure by means of an internal variable, called the *basic Lack of Knowledge* (*basic LOK*), which is included within an interval whose upper and lower bounds are stochastic variables. From these basic LOKs, one can derive, for the whole structure, the *effective Lack of Knowledge* (*effective LOK*) of a quantity of interest α, which leads to a stochastically bounded interval which can be compared with

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

experimental values derived from a family of similar real structures. In [8], this theory was successfully applied to a simple problem, which proved its identification and prediction capabilities. In this paper, we establish the first bases of a general strategy of reduction of the lacks of knowledge and we present applications on academic as well as industrial cases.

2. BASIC LOKS

2.1. Basic concept

Each similar structure can be divided into several substructures; by the way, joints can also be treated as substructures. Only the errors concerning structural stiffnesses are considered, hence the use of substructural strain energies in the following definition; indeed, we associate to any substructure E a lack of knowledge m located *anywhere* within an interval whose the two bounds are two internal variables m_E^+ and m_E^- defined by

$$\left(1 - m_E^-\right) \overline{e}_E \leq e_E \leq \left(1 + m_E^+\right) \overline{e}_E, \tag{1}$$

where $\overline{e}_E$ and e_E are the strain energies associated respectively with the deterministic, theoretical model and with one of the real structures. m_E^+ and m_E^- are the *upper basic LOK* and the *lower basic LOK* respectively.

The basic LOKs m_E^+ and m_E^- are sampled using a probabilistic law; the nature of this law is chosen *a priori* and its characteristics are defined by two values $\overline{m}_E^+$ and $\overline{m}_E^-$:

- for example, if the distribution chosen is uniform, these two values include all possible sampled values of m_E^+ and m_E^-;
- in some particular cases of imperfect modelings (e.g. nonlinear joints represented with linear models), characterized by a severe lack of information, one cannot determine precisely the distribution of lack of knowledge and it can only be stated that m is somewhere within $[-\overline{m}_E^-; \overline{m}_E^+]$.

In the absence of specific information, it is reasonable to choose the previous description. We can also consider that a normal distribution is appropriate in cases in which the sources of errors are material uncertainties.

2.2. Illustration

Let us consider the case of a lack of knowledge of the material properties: for a family of similar real structures, we assume that the lack of knowledge m of a substructure E is defined by a centered normal distribution whose Probability Density Function (PDF) is written as follows:

$$m \in [-m_E^-; m_E^+] \text{ with PDF } p(m) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{m^2}{2\sigma^2}}. \tag{2}$$

The standard deviation σ can be associated to the values $\overline{m}_E^+$ and $\overline{m}_E^-$ by stating for example that $\int_{-\overline{m}_E^-}^{\overline{m}_E^+} p(m)\mathrm{d}m = 0.99$; the PDF is then set to zero below $-\overline{m}_E^-$ and beyond $\overline{m}_E^+$, and can be normalized again.

The probability of having m within a given interval $[-m_E^-; m_E^+]$ is

$$P(-m_E^- \leq m \leq m_E^+) = \int_{-m_E^-}^{m_E^+} p(m)\mathrm{d}m. \tag{3}$$

Since the basic LOKs are defined on both sides of the theoretical model, this occurrence can be described by two independent events:

- $m \in [0; m_E^+]$, *i.e.* the event $(m_E^- = 0, m_E^+)$ occurs with probability $P^+(m_E^+)$;
- $m \in [-m_E^-; 0]$, *i.e.* the event $(m_E^-, m_E^+ = 0)$ occurs with probability $P^-(m_E^-)$.

Of course, one has $P^+(\infty) + P^-(\infty) = 1$. In this special case of a centered distribution, one even has $P^+(\infty) = P^-(\infty) = \frac{1}{2}$. This situation is depicted in Figure 1. This case illustrates how the basic LOKs should be sampled: depending on the value of m obtained, one gets two distinct types of intervals: $[0; m_E^+]$ and $[-m_E^-; 0]$.

2.3. Definition of an Interval Probability

Since the use of two distinct probabilities P^+ and P^- is rather impractical, we developed in [9] some mathematical tools in order to circumvent this difficulty.

Let us consider a family of intervals $[-m_E^-; m_E^+] \ni m$ with $m_E^+ + m_E^- = L$. An interval $[-m_E^-; m_E^+]$ is called a *standard interval* $I(L)$ if, for a given interval length L, the probability of m being in $I(L)$ is the greatest of all such intervals of length L, *i.e.*

$$I(L) = \arg \max_{\substack{[-m_E^-; m_E^+] \\ m_E^+ + m_E^- = L}} P^+(m_E^+) + P^-(m_E^-). \tag{4}$$

From this definition, we can introduce the concept of *interval probability* $P(L)$ by stating that for a given length L, $P(L)$ is the probability of having m in $I(L)$, *i.e.*

$$P(L) = P(m \in I(L)) = \max_{\substack{[-m_E^-; m_E^+] \\ m_E^+ + m_E^- = L}} P^+(m_E^+) + P^-(m_E^-). \tag{5}$$

One interpretation of these definitions is that if one wants to determine an interval such that m has a given probability P of being inside, one has to select the standard interval $I(L)$ whose probability interval $P(L)$ is equal to P, and one can show that this interval is the smallest interval $[-m_E^-; m_E^+]$ such that $P^+(m_E^+) + P^-(m_E^-) = P$, *i.e.*

$$I(L) = \arg \min_{\substack{[-m_E^-; m_E^+] \\ P^+(m_E^+) + P^-(m_E^-) = P}} m_E^+ + m_E^-. \tag{6}$$

One can also prove that the bounds of $I(L)$ verify the equality: $p(m_E^+) = p(m_E^-)$. These remarks are summarized in Figure 2 in the case of a non-centered normal law.

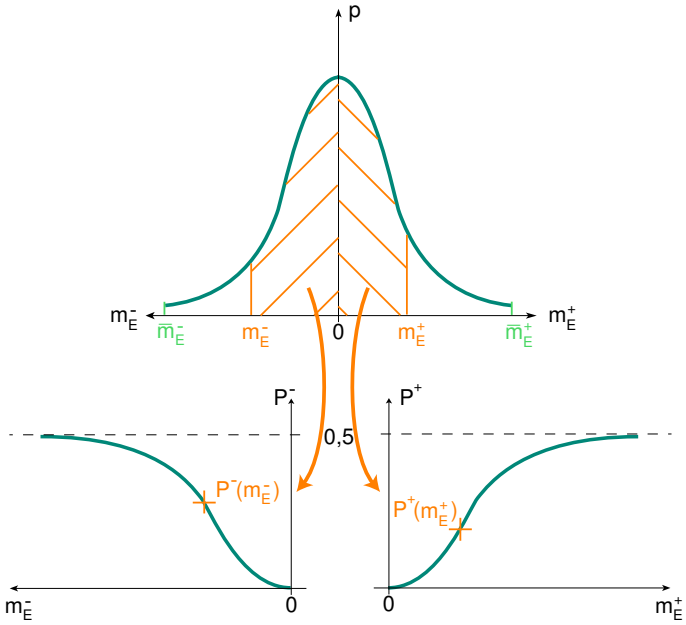


Figure 1. Example of a centered normal law: (top) PDF of m; (bottom) P^- and P^+.

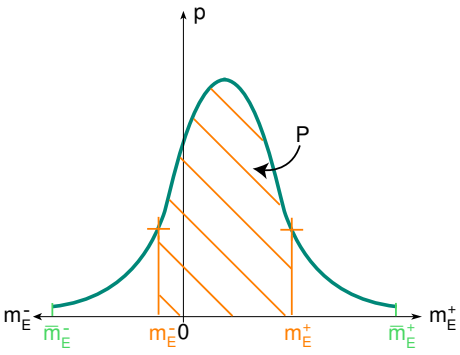


Figure 2. Illustration of the concepts of standard interval and interval probability.

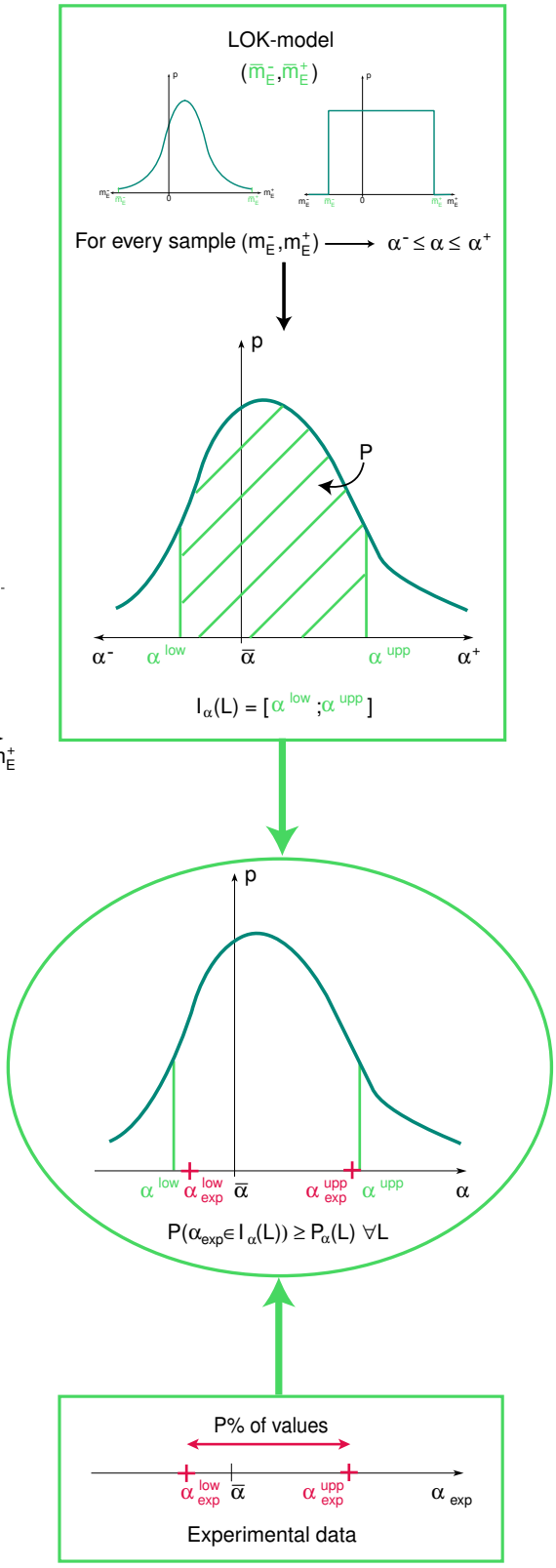


Figure 3. Illustration of the concept of Lack of Knowledge.

3. THE USE OF LOK

3.1. Principle

Let us consider a certain quantity of interest α.

- for every sample of $(m_E^-, m_E^+)_{E\in\Omega}$, one can calculate two bounds α^- and α^+ of the quantity of interest α_{mod} relative to the model, as will be shown in Section 3.2. If one knows the stochastic laws for the basic LOKs, one can obtain the probabilistic distribution of these bounds α^- and α^+ by means of an interval probability $P_\alpha(L)$ such that $P(\alpha_{mod} \in I_\alpha(L)) = P_\alpha(L)\,\forall L$. Remembering the previous interpretation of an interval probability, one can get for a given probability value P the associated standard interval $I_\alpha(L)$ such that $P(\alpha_{mod} \in I_\alpha(L)) = P$. We will refer to the two bounds of this interval as the *effective Lack of Knowledge* (*effective LOK*), and denote them α^{low} and α^{upp}.

- based on the similar real structures, one can derive in the same way two bounds α_{exp}^{low} and α_{exp}^{upp} which include $P\%$ of the experimental values α_{exp}.

The experimental data and the values obtained from the LOK model are then compared as in Figure 3. In order for the model to be conservative, the basic LOKs should be such that

$$P(\alpha_{exp} \in I_\alpha(L)) \geq P_\alpha(L)\,\forall L. \tag{7}$$

This means that one should have $\alpha^{low} \leq \alpha_{exp}^{low} \leq \alpha_{exp}^{upp} \leq \alpha^{upp}$ for any given probability value P. Note that this last interpretation is a generalization of the 99%-bounds described in [5] and [8].

3.2. Effective LOKs

The comparison between the results of the model and reality is made using quantities which are standard in the field of modal analysis: in this paper, we use free-vibration tests; therefore, our quantities of interest α are eigenfrequencies and eigenmodes. The previously defined pair of quantities α^{low} and α^{upp} is called the *effective LOK* and the corresponding values for eigenfrequencies and eigenmodes are reviewed below.

3.2.1. Effective LOK of an Eigenfrequency

If the modes $\overline{\phi}_i$ are mass-normalized, a first-order approximation ($\phi_i \simeq \overline{\phi}_i$) gives

$$\omega_i^2 - \overline{\omega}_i^2 = \phi_i^T K \phi_i - \overline{\phi}_i^T \overline{K\phi}_i \simeq \overline{\phi}_i^T (K - \overline{K})\overline{\phi}_i = 2\sum_{E\in\Omega} \left(e_E(\overline{\phi}_i) - \overline{e}_E(\overline{\phi}_i)\right). \tag{8}$$

From relationship (1), one has for a given sample $(m_E^-, m_E^+)_{E\in\Omega}$

$$\omega_i^{2-} \leq \omega_i^2 \leq \omega_i^{2+} \tag{9}$$

with

$$\omega_i^{2-} = \overline{\omega}_i^2 - 2\sum_{E\in\Omega} m_E^- \overline{e}_E(\overline{\phi}_i), \tag{10}$$

$$\omega_i^{2+} = \overline{\omega}_i^2 + 2\sum_{E\in\Omega} m_E^+ \overline{e}_E(\overline{\phi}_i) \tag{11}$$

Thus, for a given probability value P, one can derive the two bounds $\omega_i^{2\,low}$ and $\omega_i^{2\,upp}$ of the associated standard interval $I_{\omega_i^2}(L)$, i.e. the effective LOK of an eigenfrequency.

3.2.2. Effective LOK of an Eigendisplacement

For small values of the basic LOKs, we can approximate the variation of an eigendisplacement (defined as the value at a Degree of Freedom of an eigenmode) by writing

$$\phi_{ki} - \overline{\phi}_{ki} \simeq U^T \Delta K \overline{\phi}_i = \sum_{E\in\Omega} U^T \left(K_E - \overline{K}_E\right) \overline{\phi}_i \tag{12}$$

where U is a given vector. Using $U^T K_E \overline{\phi}_i = \frac{1}{2} e_E(U + \overline{\phi}_i) - \frac{1}{2} e_E(U - \overline{\phi}_i)$ and relationship (1), one gets for a given sample $(m_E^-, m_E^+)_{E\in\Omega}$

$$\phi_{ki}^- \leq \phi_{ki} \leq \phi_{ki}^+ \tag{13}$$

with

$$\phi_{ki}^- = \overline{\phi}_{ki} - \frac{1}{2}\sum_{E\in\Omega} \left\{ m_E^- \overline{e}_E(U + \overline{\phi}_i) + m_E^+ \overline{e}_E(U - \overline{\phi}_i) \right\}, \tag{14}$$

$$\phi_{ki}^+ = \overline{\phi}_{ki} + \frac{1}{2}\sum_{E\in\Omega} \left\{ m_E^+ \overline{e}_E(U + \overline{\phi}_i) + m_E^- \overline{e}_E(U - \overline{\phi}_i) \right\} \tag{15}$$

Thus, for a given probability P, one can derive the two bounds ϕ_{ki}^{low} and ϕ_{ki}^{upp} of the associated standard interval $I_{\phi_{ki}}(L)$, i.e. the effective LOK of an eigendisplacement.

4. DETERMINATION OF THE BASIC LOKS

The purpose of determining the basic LOKs is to find the values of m_E^+ and m_E^- which are the most representative of the dispersion. The process we introduce here is based on the idea that the more abundant the experimental data, the better we can reduce the LOK-level within the structure. Therefore, the first step of the process consists in setting initial, overestimated values of the basic LOKs for all the substructures; this can be done by applying one's *a priori* knowledge or experience of the structure being studied. Indeed, it is not vital to use accurate estimates; the most important point is to use overestimated values $(\overline{m}_E^{+\,0}, \overline{m}_E^{-\,0})_{E\in\Omega}$ of the basic LOKs for each substructure.

The reduction process consists in using relevant experimental data to reduce the LOK-level individually for each substructure. Let us consider a given substructure E^*. One has to find smaller values of $\overline{m}_{E^*}^+$ and $\overline{m}_{E^*}^-$, which, in terms of interval probabilities, yields the following relationship:

$$P_{E^*}^0(L) \leq P_{E^*}(L) \quad \forall L. \tag{16}$$

This reduction should be carried out with the constraint created by the experimental information selected:

$$\alpha^{low} \leq \alpha_{exp}^{low} \leq \alpha_{exp}^{upp} \leq \alpha^{upp}. \tag{17}$$

In fact, as one is interested in the minimization of the lack of knowledge of Substructure E^*, one intends to take into account the worst happening case concerning all the other substructures. We can write formally for each given sample $(m_E^-, m_E^+)_{E\in\Omega}$:

$$\alpha^{worst+} = \overline{\alpha} + S_{E^*}\Delta\alpha_{E^*}^{+} + \sum_{E\neq E^*} \Delta\alpha_E^{worst+} \tag{18}$$

$$\alpha^{worst-} = \overline{\alpha} + S_{E^*}\Delta\alpha_{E^*}^{-} + \sum_{E\neq E^*} \Delta\alpha_E^{worst-}. \tag{19}$$

This worst-case analysis is completed by the introduction of a coefficient quantifying whether the experimental information is more or less representative of the behavior of the structure; this value $S_{E^*} \in]0;1]$ is called *test severity coefficient* for Substructure E^* and is maximal when the test fits perfectly the global mechanics of the structure. Then we can associate to these bounds α^{worst+} and α^{worst-} an interval probability $P_{\alpha^{worst}}(L)$ and derive the two bounds $\alpha^{worst\,upp}$ and $\alpha^{worst\,low}$ of the associated standard interval $I_{\alpha^{worst}}(L)$ for a given probability P.So the following constraints are introduced:

$$\alpha^{worst\,low} \leq \alpha_{exp}^{low} \leq \alpha_{exp}^{upp} \leq \alpha^{worst\,upp}. \tag{20}$$

So as a summary, the problem consists in finding

$$\max P_{E^*}(L, \overline{m}) \quad \forall L \tag{21}$$

with the previous constraints, and for several given values of L.

5. APPLICATION TO A SIMPLE PROBLEM

5.1. Definition of the Structure

5.1.1. Deterministic Theoretical Model

The structure being considered is a plane truss similar to that studied in [8]; it consists of six bars connected by spherical joints, as shown in Figure 4. We assume that the bars are solicited only in traction-compression and that the connections between the ground and the structure at Nodes 1 and 2 are perfectly rigid links. The material properties of the associated theoretical model are given in Table 1.

5.1.2. Experimental Data

A family of such actual trusses is simulated and their eigenfrequencies and eigenmodes constitute the data which is then used in the reduction process described in Section 4: the 'experimental' data are simulated by using the theoretical model and introducing some stochastic distributions in the stiffness characteristics of the substructures; these changes are summarized in Table 1. Note that material "X" is considered to be imperfectly known; hence the uniform law chosen for the simulation. For each of these 'real' structures, we are able to calculate eigenfrequencies and eigenmodes and, thus, derive experimental distributions of the eigenfrequencies or eigendisplacements associated to the real structures (see for example the distribution of ω_{exp}^2 for Mode 6 in Figure 5).

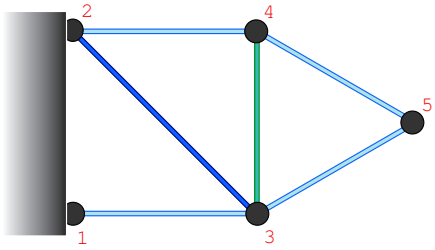


Figure 4. Plane truss example

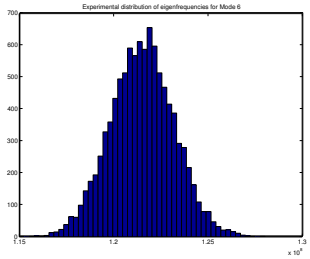


Figure 5. Experimental distribution of ω^2_{exp} for Mode 6

Table 1. Properties of the deterministic plane truss and of the simulated structures, with Bars 1-3, 3-5, 4-5 and 2-4 constituting Group g1, bar 2-3 as Group g2 and bar 3-4 as Group 3.

Gp	Material	Young's modulus	Density	Law	Mean/Range	Simulated stiffnesses
g1	aluminium	$\overline{E}_{g1} = 72GPa$	$2700kg/m^3$	normal	0%/5%	$\in [0.95\overline{K}_{g1}; 1.05\overline{K}_{g1}]$
g2	steel	$\overline{E}_{g2} = 210GPa$	$7800kg/m^3$	normal	−5%/10%	$\in [0.85\overline{K}_{g2}; 1.05\overline{K}_{g2}]$
g3	"X"	$\overline{E}_{g3} = 10GPa$	$1500kg/m^3$	uniform	5%/15%	$\in [0.90\overline{K}_{g3}; 1.20\overline{K}_{g3}]$

5.2. Reduction of the Basic LOKs of the Structure Being Considered

The reduction process is carried out by assuming *a priori* an initial LOK-level of 50% for each substructure, which guarantees overestimated starting values. We also assume a normal LOK-distribution for the aluminum and steel bars and select a uniform LOK-distribution for the "X" bar. From the measured distributions of eigenfrequencies, we decide to keep in mind the values $\omega^{2\,upp}_{i\,exp}$ and $\omega^{2\,low}_{i\,exp}$ that include 99% of the experimental eigenfrequencies; this means that we do not care any more about the distribution of these experimental eigenfrequencies within the two 99%-values. If we wanted a more precise description, we could also take the 50%-values in order to have an estimation of the standard deviation of the experimental values.

Next, it is important to select the most relevant experimental tests to carry out the successive processes. An effective method consists in using the fact that the sensitivities of the effective LOKs to the basic LOKs are directly related to the modal strain energy of the theoretical, deterministic model (see Section 3.2 for more details). The most relevant modal tests for reducing the basic LOK of Substructure E^* are those in which the modal strain energy is contained mainly within Substructure E^*. As we are interested with experimental eigenfrequencies, the modal strain energies $\overline{e}_E(\overline{\phi}_i)$ are considered for Modes 1 to 6 and are listed in Table 2 where the largest substructural energies are emphasized.

Table 2. Modal strain energies for Modes 1 to 6.

$\overline{e}_E(\overline{\phi}_i)$	i=1	i=2	i=3	i=4	i=5	i=6
E=g1	$\mathbf{3.3.10^5}$	$\mathbf{1.3.10^6}$	$\mathbf{7.6.10^6}$	$3.8.10^6$	$\mathbf{2.5.10^7}$	$\mathbf{6.0.10^7}$
E=g2	$\mathbf{1.4.10^5}$	$6.7.10^4$	$9.9.10^3$	$\mathbf{1.0.10^7}$	$2.0.10^6$	$1.7.10^5$
E=g3	$\mathbf{2.5.10^5}$	$\mathbf{1.7.10^6}$	$6.1.10^5$	$4.7.10^5$	$6.9.10^4$	$1.9.10^5$

The reduction is achieved by selecting as relevant experimental tests $\omega^{2\,upp}_{i\,exp}$ ans $\omega^{2\,low}_{i\,exp}$ derived from Modes 6, 4 and 2 for Groups 1, 2 and 3 respectively, and by considering

that these data are representative of the global behavior of the structure (test severity coefficients equal to one). The results come out as

$$\overline{m}^{+}_{g1} = 0.032 \qquad \overline{m}^{+}_{g2} = 0.034 \qquad \overline{m}^{+}_{g3} = 0.205$$
$$\overline{m}^{-}_{g1} = 0.034 \qquad \overline{m}^{-}_{g2} = 0.092 \qquad \overline{m}^{-}_{g3} = 0.101.$$

In this very special case, with a first-order assumption, these results are to be compared directly with the stiffness distributions introduced into the deterministic model to simulate the experimental data: $[(1-0.05)\overline{K}_{g1}; (1+0.05)\overline{K}_{g1}]$, $[(1-0.15)\overline{K}_{g2}; (1+0.05)\overline{K}_{g2}]$ and $[(1-0.10)\overline{K}_{g3}; (1+0.20)\overline{K}_{g3}]$. We can conclude that the agreement is rather good. The choice of the relevant experimental data is crucial; if one tried to reduce the LOKs of any group using Mode 1, the minimization process would not lead to any reduction because the influence of the other two groups is not small enough.

5.3. Capacity of Prediction

With the values just obtained, we are able to calculate the effective LOKs for the three other modes (1, 3 and 5) in order to evaluate the results of the reduction process. The basic LOKs are sampled with the values determined and the probabilistic laws chosen; the corresponding calculated 99%-values are listed and compared with the experimental 99%-values in Table 3 below. The constraints are successfully respected for Modes 1, 3 and 5, which shows the consistency of the results obtained with Modes 2, 4 and 6.

Table 3. Comparison of eigenfrequencies and eigendisplacements (99%-values) for Modes 1, 3, 5.

i	$\omega_i^{2\,low}$	$\omega_{i\,exp}^{2\,low}$	$\overline{\omega}_i 2$	$\omega_{i\,exp}^{2\,upp}$	$\omega_i^{2\,upp}$	ϕ_{ki}^{low}	$\phi_{ki\,exp}^{low}$	$\overline{\phi}_{ki}$	$\phi_{ki\,exp}^{upp}$	ϕ_{ki}^{upp}
1	$1.36.10^6$	$1.35.10^6$	$1.43.10^6$	$1.53.10^6$	$1.54.10^6$	0.85	0.88	0.95	0.99	1.01
3	$1.58.10^7$	$1.58.10^7$	$1.64.10^7$	$1.71.10^7$	$1.70.10^7$	−1.00	−0.98	−0.95	−0.91	−0.90
5	$5.28.10^7$	$5.29.10^7$	$5.51.10^7$	$5.68.10^7$	$5.69.10^7$	−0.74	−0.72	−0.68	−0.62	−0.62

6. STUDY OF A STRUCTURE WITH A MODELLING ERROR

6.1. Presentation of the Structure

6.1.1. Deterministic Theoretical Model

In this example, we want to study the ability of our theory to describe a modelling error in the theoretical model. The studied structure is a beam clamped at one end; we are interested with its bending vibrations. The theoretical model consists of 100 standard Bernoulli-Euler elements based on a cubic interpolation of displacements.

6.1.2. Experimental Data

The experimental structure is simulated by inserting a joint in the middle of the beam: the two corresponding ends of the two half beams are linked by two linear springs: one concerning the vertical translation ($k = 7.10^7 N/m$) and a rotational one ($K = 1000 N.m/rad$), as in Figure 6. We then compute the eigenfrequencies and eigenmodes of this structure.

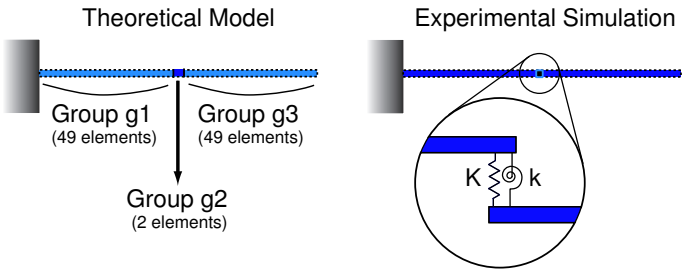


Figure 6. Theoretical model and experimental simulation of the clamped beam.

6.2. Calculation of the Basic LOKs

Before determining the Lacks of Knowledge, the model is updated with the first 15 modes of the experimental structure according [2]. This method leads to the correction of the stiffnesses of the two elements at both sides of the joint (with a factor of 0.41). After updating, we still have a global residual error of 4%, and no further improvement can be made, which means that the model cannot represent the experimental data in a better way.

The determination of the Lacks of Knowledge is achieved on three different groups: Group g2 corresponds to the two elements at both sides of the joint, Group g1 and Group g3 to the other elements located before and after the joint respectively, as indicated in Figure 6. We use the eigenfrequencies of Modes 4, 8 and 12, by considering them as extreme values describing the distribution coming from the reality; moreover, the test severity coefficients are set to one. With an initial LOK-level of 50% and a normal law assumption for each group, we obtain the following results:

$$\begin{array}{lll} \overline{m}^{+}_{g1} = 0.003 & \overline{m}^{+}_{g2} = 0 & \overline{m}^{+}_{g3} = 0 \\ \overline{m}^{-}_{g1} = 0 & \overline{m}^{-}_{g2} = 0.040 & \overline{m}^{-}_{g3} = 0, \end{array}$$

which means that the actual structure is perfectly described by the theoretical model, excepted in the neighborhood of the joint where we find a lack of knowledge of 4%. This example shows that the theory of the Lacks of Knowledge is useful to indicate the areas where the model is not good enough to represent the global behavior of the whole structure, and gives an estimate of its accuracy.

7. STUDY OF A REAL CASE

7.1. Description of the Structure

We will now present the application of the method to an actual, industrial structure: the Sylda5 satellite support developed by the EADS company is capable of carrying two individual satellites and is represented in Figure 7. Vibration tests were performed by IABG for DASA/DORNIER under contract with CNES: the test setup consisted of 5 exciters and 260 sensors. The model proposed by EADS represents both the support itself and a payload simulating the presence of a satellite; it consists of 38 substructures with various materials, including orthotropic sandwiches, aluminum and steel. The first

tests have shown that it was essential to take the ground into account in the model; this was done using 3 rotational springs, one translation spring and a rigid-body-movement constraint for all the bottom nodes. In the end, the model consists of 27648 DOFs and 9728 elements.

We consider as experimental data the extreme values of the eigenfrequencies and eigenmodes measured from a series of tests, without caring about their distribution.

7.2. Determination of the Basic LOKs

First, the model is updated with the first 12 modes using the method described in [2]. At this point, we want to describe the remaining lacks of knowledge. In order to do that, we divide the whole structure into 4 main groups of substructures, depicted in Figure 8:

- Group g1 is associated with the payload substructure;
- Group g2 represents the junction between the payload substructure and the Sylda support itself;
- Group g3 is the Sylda support;
- Group g4 is associated with the ground model.

Figure 7. The Sylda5 satellite support.

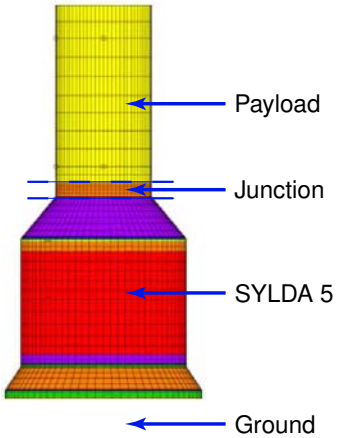


Figure 8. Model associated to Sylda5.

The objective is to carry out the reduction of the most influential lacks of knowledge. An initial value of 50% is assumed for each pair $(\overline{m}_E^{+0}, \overline{m}_E^{-0})$, where $E \in \{g1, g2, g3, g4\}$. In the first 8 modes, only Group 1 and Group 3 have significant modal strain energy levels and, thus, they are for the moment the only ones involved in the reduction process. We consider as experimental data the extreme values of the eigenfrequencies measured from a series of tests, without caring about their distribution.

With such values from Modes 7 and 8 on the one hand, and from Modes 4 and 5 on the other hand, we reduce the basic LOKs of Group 1 and Group 3 respectively to

$$\overline{m}^{+}_{g1} = 0.154 \qquad \overline{m}^{+}_{g3} = 0.001$$
$$\overline{m}^{-}_{g1} = 0.009 \qquad \overline{m}^{-}_{g3} = 0.012,$$

by using test severity coefficients equal to one. As a conclusion, we can stress that these results corroborate the quality of the updated Sylda support model (Group 3) and give an estimation of the accuracy of the model used to describe the payload (Group 1).

8. CONCLUSION

We showed in this paper some applications of the theory of the Lacks of Knowledge which combines convex uncertainty models with probabilistic features. The method is able to quantify local uncertainties by using quantities of interest defined on the whole structure and it can also be useful to the estimation of modelling accuracy. The reduction process that we introduced in this paper consider experimental data as information usable to reduce the overestimated basic LOKs assumed for each substructure. This approach should lead the way to the development of a general method for reducing the lacks of knowledge for predetermined families of parameters by designating what tests should be performed or which substructure models should be improved.

REFERENCES

1. J.E. Mottershead and M.I. Friswell. Model updating in structural dynamics: A survey. *Journal of Sound and Vibration*, 167(2):347–375, 1993.
2. P. Ladevèze and A. Chouaki. Application of a posteriori error estimation for structural model updating. *Inverse Problems*, 15:49–58, 1999.
3. J. F. Schultze, F. M. Hemez, S. W. Doebling, and Hoon Sohn. Statistical based non-linear model updating using feature extraction. In *Proceedings of the 19th International Modal Analysis Conference (IMAC-XIX)*, pages 18–26, Kissimmee, Florida, Feb. 5-8 2001.
4. F. M. Hemez, A. C. Wilson, and S. W. Doebling. Design of computer experiments for improving an impact test simulation. In *Proceedings of the 19th International Modal Analysis Conference (IMAC-XIX)*, pages 977–985, Kissimmee, Florida, Feb. 5-8 2001.
5. P. Ladevèze. On a theory of the lack of knowledge in structural computation. Technical Note SY/XS 136 127, EADS Launch Vehicles, April 2002. in French.
6. Y. Ben-Haim. Design certification with information-gap uncertainty. *Structural Safety*, 21:269–289, 1999.
7. P. Vinot, S. Cogan, and Y. Ben-Haim. Reliability of structural dynamics models based on info-gap models. In *Proceedings of the 20th International Modal Analysis Conference (IMAC-XX)*, pages 1057–1063, Los Angeles, California, Feb. 4-7 2002.
8. D. Barthe, A. Deraemaeker, P. Ladevèze, and G. Puel. On a theory of the quantification of the lack of knowledge in structural computation. In *Proceedings of the 21st International Modal Analysis Conference (IMAC-XXI)*, Kissimmee, Florida, Feb. 3-6 2003.
9. P. Ladevèze. The theory of the lack of knowledge in structural dynamics: Development and first examples. Technical Note LY22 139 423, EADS Launch Vehicles, May 2003. in French.

Modeling uncertainty in population biology: how the model is written does matter

Janos G. Hajagos*

March 2, 2004

1 An alternative to second-order Monte Carlo

Monte Carlo based approaches are used to calculate the risk of extinction for threatened species. In the risk assessment the exact values of the statistical moments of the input distributions need to be known. At best, the mean and variance for the growth rate of the population might be known plus or minus ten percent of the estimated value. The normal course of action is to perform a second-order Monte Carlo analysis. In such an analysis, a second statistical distribution is sampled for the moments of the first distribution. Second-order Monte Carlo adds an additional factor of computation time and makes more assumptions about the distribution of moments; when data is sparse, like in the case of endangered species, these additional probabilistic assumptions might not be supported.

An alternative to second-order Monte Carlo analysis is presented in this paper. Instead of sampling from a second statistical distribution, the uncertainty around the moments will be bound, and then propagated through a numerical simulation of population dynamics using interval analysis. With interval analysis no additional assumptions except that the moments are bounded need to be made. It will be shown that there are two ways to write the equation for population growth. The correct equation to use will depend on what is meant by an interval. If one believes that an interval represents a bounded set of possible values then Equation 8 should be used, but if one believes that an interval represents uncertainty of not knowing a fixed value then Equation 9 should be used. The choice is not without consequences: the bounds on the quasi-extinction decline risk will be tighter with Equation 9.

*Department of Ecology and Evolution, State University of New York at Stony Brook, Stony Brook, NY 11794-5245, U.S.A.; email: jhajagos@life.bio.sunysb.edu; fax: 631-632-7626

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

2 Population models

The basic model for growth of an animal population is the exponential growth function, written here in its continuous form

$$f(t) = N_0 \exp(rt), \tag{1}$$

where N_0 is the initial population size, t is time, and r is the per capita rate of growth. This function arises from a solution to the simple differential equation

$$\frac{dN}{dt} = rN, \tag{2}$$

where N is the population size.

Discrete deterministic population models are normally written in the form

$$N_{t+1} = RN_t, \tag{3}$$

where R is a per unit time multiplier, N_t is the population at time t. For predicting N_T, such that, $T \in \{0, 1, 2, \ldots\}$, one has

$$N_{t+T} = R^T N_t. \tag{4}$$

An important relationship exists between R and r, the finite rate of increase and the per capita rate of growth, that is,

$$R = \exp(r). \tag{5}$$

From this point the notation used to write a discrete function of population growth will change. We will now consider the population abundance at time T to be a function of the size of the population at time 0, the time horizon T (the length of the simulation), and the per-capita growth rate r. The equation of population growth rewritten in terms of the new notation is

$$f(N_0, r, T) = N_0 \exp(rT) = N_T. \tag{6}$$

3 Adding stochasticity

For real biological populations, that is, those that are observed in nature, the per-capita rate of population growth is not fixed through time but varies. Equation 6 can be rewritten to take into account varying rates of r

$$f(N_0, \{r_1, \ldots, r_T\}, T) = N_0 \exp\left(\sum_{i=1}^{T} r_i\right) = N_T, \tag{7}$$

where r_i is a random variate from G, a statistical distribution. It is assumed here that G is a normal distribution with a mean $\bar{r}$ and with a standard deviation of σ_r; $r_i = g(\bar{r}, \sigma_r)$ is a random variate from the normal distribution $G(\bar{r}, \sigma_r)$ [Lewontin and Cohen, 1969].

To simulate the potential dynamics to the populations, we make K runs or realizations of the model.

4 Adding measurement uncertainty

To propagate epistemic uncertainty, that is, uncertainty which can reduced through effort, interval analysis [Moore, 1966] will be used. An interval $\mathbf{X}$ is defined as a closed set on the real line, such that, $x \in \mathbf{X} \subseteq \mathbb{R}$ where $\underline{X} \leq x \leq \overline{X}$, and $\underline{X}$ and $\overline{X}$ are the infinimum and supremum, respectively of $\mathbf{X}$. The set of all intervals on the real line is denoted $\mathbb{IR}$. Given intervals $\mathbf{X}$ and $\mathbf{Y}$ addition is defined as

$$\mathbf{X} + \mathbf{Y} = [\underline{X} + \underline{Y}, \overline{X} + \overline{Y}] = \{x + y : x \in \mathbf{X}, y \in \mathbf{Y}\}$$

There are interval definitions for a wide range of basic mathematical operators, such as, $\{-, \times, /, ^2\}$, and for functions, such as, $\{\exp, \log, \sin, \cos\}$. To propagate epistemic errors through a simulation of population dynamics two additional operators need to be defined:

$$\mathbf{X} \times \mathbf{Y} = [\min(\underline{XY}, \overline{XY}, \underline{X}\overline{Y}, \overline{X}\underline{Y}), \max(\underline{XY}, \overline{XY}, \underline{X}\overline{Y}, \overline{X}\underline{Y})] = \{xy : x \in \mathbf{X}, y \in \mathbf{Y}\}$$

$$\exp(\mathbf{X}) = [\exp(\underline{X}), \exp(\overline{X})] = \{\exp(x) : x \in \mathbf{X}\}.$$

By outwardly rounding the endpoints of an interval operation the interval is guaranteed to contain the true value. For the simulation of population dynamics the Intlab toolbox [Rump, 1999b, Rump, 1999a] for Matlab is used.

The algebra on intervals differs from the algebra on real numbers. For example,

$$\mathbf{C} \times (\mathbf{A} + \mathbf{B}) \subseteq \mathbf{C} \times \mathbf{A} + \mathbf{C} \times \mathbf{B}$$

this is known as the subdistributive law [Moore, 1979]. In the non-strict inequality, equality will hold when $\underline{A}, \underline{B} > 0$. Of more importance is Moore's single use theorem which states that if each variable in a mathematical expression occurs only once then the resulting bounds from applying interval operators will be optimal [Hansen, 1997]. The effect of repeated variables is that, in some cases, the bounds on the evaluated expression will be conservatively suboptimal or too wide [Kreinovich et al., 2002]. In the continuous and discrete models of exponential growth, Equations 1 & 5, each variable appears only once, therefore interval arithmetic can be naively applied.

A statistical distribution can have uncertain moments, for example, bounds on the mean or standard deviation (c.f. [Ferson, 2002]). To propagate epistemic uncertainty through a Monte Carlo simulation interval analysis is used.

Equation 7 can be written in two *intervalized* forms

$$h(\mathbf{N}_0, \bar{\mathbf{r}}, \sigma_{\mathbf{r}}, T) = \mathbf{N}_0 \exp\left(\sum_{i=1}^{T} g_i(\bar{\mathbf{r}}, \sigma_{\mathbf{r}})\right) = \mathbf{N}_T \tag{8}$$

$$j(\mathbf{N}_0, \bar{\mathbf{r}}, \sigma_{\mathbf{r}}, T) = \mathbf{N}_0 \exp\left(T\bar{\mathbf{r}} + \sigma_{\mathbf{r}} \sum_{i=1}^{T} g_i(0, 1)\right) = \mathbf{N}_T \tag{9}$$

If all the parameters for Equations 8 & 9 are degenerate intervals then the two functions are equivalent given the same set of random deviates. A degenerate interval is defined as $\mathbf{X} = [\underline{x}, \overline{x}]$, where $\underline{x} = \overline{x}$. If $\mathbf{N}_0 \in \mathbb{IR}$ is the only non-degenerate parameter the expressions are still equivalent because $\mathbf{N}_0$ appears only once in each of the expressions. When $\bar{\mathbf{r}} \in \mathbb{IR}$ or $\sigma_{\mathbf{r}} \in \mathbb{IR}$ then the expressions do not give equivalent results, and it follows from subdistibutivity of interval arithmetic $j(\mathbf{N}_0, \bar{\mathbf{r}}, \sigma_{\mathbf{r}}, T) \subseteq h(\mathbf{N}_0, \bar{\mathbf{r}}, \sigma_{\mathbf{r}}, T)$.

In Equation 8 the dependency between the statistical moments for the individual variates in the sum $g_1(\bar{r}, \sigma_r) + g_2(\bar{r}, \sigma_r) + \ldots + g_T(\bar{r}, \sigma_r)$ is not accounted for. The dependency occurs in that the $\bar{\mathbf{r}}$ and $\sigma_{\mathbf{r}}$ occur repeatedly in the expression as statistical moments for g. Due to the ability to factor out the mean and variance from a normal variate the sum of variates can be algebraically rearranged to take into account that $\bar{r} \in \bar{\mathbf{r}}$ and $\sigma_r \in \sigma_{\mathbf{r}}$ are fixed values:

$$\bar{r} + \sigma_r g_1(0,1) + \bar{r} + \sigma_r g_2(0,1) + \ldots + \bar{r} + \sigma_r g_T(0,1) = T\bar{r} + \sigma_r \sum_{i=1}^{T} g_i(0,1).$$

The question then becomes which of the formulations, Equations 8 or 9, is correct. The answer to this question depends on one's philosophical view of what an interval is. If the belief is that there exists a single fixed value bounded by an infinimum and supremum which bounds the uncertainty about ones estimate of the fixed value, then Equation 9 gives the optimal answer. However, if one thinks of an interval as representing a closed bounded set then there is no reason to believe that the $\bar{r}$ is fixed at each point in time. Allowing $\bar{r}$ or σ_r not to be fixed leads to widening bounds on N_T.

5 Quasi-extinction risk

The study of population viability is focused on quantifying the risk of a population falling below a critical period over a fixed time period. Rather then focusing entirely on total extinction, $N = 0$, the concept of quasi-extinction risk has been developed [Ginzburg et al., 1982]. Quasi-extinction risk is the probability that a population will fall below a given threshold during the simulation. Because intervals were used to propagate uncertainty through the simulation upper and lower bounds on the quasi-extinction risk curve must also be generated.

For Monte Carlo simulations of population dynamics the quasi-extinction decline curve is generated from the minimum of each k series of abundance.

$$N_{\min_k} = \min\left(N_{1,k}, N_{2,k} \ldots, N_{T,k}\right). \tag{10}$$

Note that the initial abundance $N_{0,k}$ is not included in the calculation of the minimum [Akçakaya et al., 1999]. For a sorted list of abundances $N_{\min_1} \leq N_{\min_2} \leq \ldots \leq N_{\min_K}$, where K is the total number of simulations, a cumulative probability $p_k = k\frac{1}{K}$ is associated with each $N_{\min_k}$.

For interval data the minimum is defined as

$$\min\left(\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}\right) = \left[\min(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n), \min(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n)\right] \tag{11}$$

$$\mathbf{N}_{\min_k} = \min\left(\mathbf{N}_{1,k}, \mathbf{N}_{2,k} \ldots, \mathbf{N}_{T,k}\right). \tag{12}$$

To generate the quasi-extinction decline risk curve — the cumulative distribution function of minimum abundances — for interval data the infinimum and supremum are sorted separately

$$\underline{N}_{\min_1} \leq \underline{N}_{\min_2} \leq \cdots \leq \underline{N}_{\min_K}$$
$$\overline{N}_{\min_1} \leq \overline{N}_{\min_2} \leq \ldots \leq \overline{N}_{\min_K}.$$

A probability mass $p_k = k\frac{1}{K}$ is associated with each sorted $\underline{N}_{\min_k}$ and $\overline{N}_{\min_k}$. To conservatively bound the quasi-extinction decline curve a step function is used. The bounds on the infinimum of the CDF are

$$\underline{\mathrm{CDF}}(x) = \begin{cases} \text{if } \underline{N}_{\min_1} \leq x < \underline{N}_{\min_2} \text{ then } 1/K \\ \text{if } \underline{N}_{\min_2} \leq x < \underline{N}_{\min_3} \text{ then } 2/K \\ \vdots \\ \text{if } \underline{N}_{\min_{K-1}} \leq x < \underline{N}_{\min_K} \text{ then } 1 \end{cases} \tag{13}$$

and the bounds on the supremum are

$$\overline{\mathrm{CDF}}(x) = \begin{cases} \text{if } \overline{N}_{\min_1} < x \leq \overline{N}_{\min_2} \text{ then } 1/K \\ \text{if } \overline{N}_{\min_2} < x \leq \overline{N}_{\min_3} \text{ then } 2/K \\ \vdots \\ \text{if } \overline{N}_{\min_{K-1}} < x \leq \overline{N}_{\min_K} \text{ then } 1 \end{cases}. \tag{14}$$

6 Acknowledgments

I would like to thank Dr. Scott Ferson of Applied Biomathematics for his comments on this paper, and Dr. Lev Ginzburg of Stony Brook University for his support.

References

[Akçakaya et al., 1999] Akçakaya, H. R., Burgman, M. A., and Ginzburg, L. R. (1999). *Applied Population Ecology: Principles and Computer Excercises using RAMAS Ecolab.* Sinauer Associates, Inc., Sunderland, Massachusetts.

[Ferson, 2002] Ferson, S. (2002). *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers.* Lewis Publishers, Boca Raton, Florida.

[Ginzburg et al., 1982] Ginzburg, L. R., Slobodkin, L. B., Johnson, K., and Bindman, A. G. (1982). Quasiextinction probabilities as a measure of impact on population growth. *Risk Analysis*, 21:171–181.

[Hansen, 1997] Hansen, E. (1997). Sharpness in interval computations. *Reliable Computing*, 3:7–29.

[Kreinovich et al., 2002] Kreinovich, V., Longpre, L., and Buckley, J. J. (2002). Are there efficient necessary and sufficient conditions for straightforward interval computations to be exact? In *Extended Abstracts of the 2002 SIAM Workshop on Validated Computing*, pages 94–96, Toronto, Canada.

[Lewontin and Cohen, 1969] Lewontin, R. C. and Cohen, D. (1969). On population growth in a randomly varying environment. *Proceedings of the National Academy of Sciences*, 62:1056–1060.

[Moore, 1966] Moore, R. (1966). *Interval Analysis*. Prentice-Hall, Englewood Cliffs, New Jersey.

[Moore, 1979] Moore, R. E. (1979). *Methods and Applications of Interval Analysis*. Society for Industrial and Applied Mathematics, Philadelphia.

[Rump, 1999a] Rump, S. M. (1999a). Fast and parallel interval arithmetic. *BIT*, 39(3):534–554.

[Rump, 1999b] Rump, S. M. (1999b). Intlab – interval laboratory. In Csendes, T., editor, *Developments in Reliable Computing*, pages 77–104. Kluwer Academic Publishers.

ENHANCING THE MORRIS METHOD

F. Campolongo[(1)], J. Cariboni[(1),(2)], A. Saltelli[(1),] and W. Schoutens[(2)]

1. European Commission, Joint Research Centre, Via Fermi 1, Ispra 21020, Italy
2. K.U.Leuven, U.C.S., W. De Croylaan 54, B-3001 Leuven, Belgium

E-mail: francesca.campolongo@jrc.it; jessica.cariboni@jrc.it; andrea.saltelli@jrc.it; Wim.Schoutens@wis.kuleuven.ac.be; Jessica.Cariboni@wis.kuleuven.ac.be

Abstract: The screening method proposed by Morris (1991) and recently improved by Campolongo et al. (2003) is very effective to screen a subset of few important input factors among a large number contained in a model. In this work the enhanced Morris method is first confronted with the variance based methods and then employed to assess the sensitivity of a financial model for option pricing.

Keywords: sensitivity analysis, screening designs, Morris method, variance based sensitivity indices, the Heston model, option pricing.

1. INTRODUCTION

A sensitivity analysis method widely used to screen factors in models of large dimensionality is the design proposed by Morris [1]. The Morris method deals efficiently with models containing hundreds of input factors without relying on strict assumptions about the model, such as for instance additivity or monotonicity of the model input-output relationship.

The Morris method is simple to understand and implement, and its results are easily interpreted. Furthermore it is economic in the sense that it requires a number of model evaluations that is linear in the number of model factors. The method can be regarded as global as the final measure is obtained by averaging a number of local measures (the elementary effects), computed at different points of the input space.

In very recent work [2] Campolongo and coworkers proposed an improved version of the Morris measure μ, denoted as μ^*, which is more effective in ranking factors in order of importance. Furthermore, they extended to the Morris measure a desirable property of the variance based methods: the capability to treat group of factors as if they were single factors.

Here we extend the study in [2] by testing the performance of μ and μ^* by groups on an analytical test function recently proposed by O'Hagan [3]. Results of the enhanced Morris method are also compared with those obtained by making use of the variance based sensitivity measures. The motivation for this comparison lies in the present trend that sees the variance based methods as particularly apt to sensitivity analysis, because of their desirable properties in terms of model independence, global nature, ease of interpretation and others [4]. At the same time the method of Morris is considerably cheaper than the variance based methods in terms of model evaluation, hence the interest in this comparison.

Results confirm that the Morris method, in its new version μ*, is as efficient as the variance based techniques in identifying irrelevant factors, i.e. those factors that can be fixed at any given value within their range of uncertainty without significantly affecting the total output variance. Hence, it is recommendable as a valid alternative to the variance based when the problem is such that the cost of the variance based techniques is too high.

A theoretical link between the Morris and the variance based measures is also argued for by expressing the measure μ* in terms of conditional variances.

Section 4 of this work is dedicated to an application of the enhanced Morris method to a real test case, a financial model here used to price a European call option. Results of the sensitivity analysis confirm the good quality of the model and encourage to extend its use to more delicate problems such those of pricing exotic options.

2. METHODOLOGY

2.1. The Morris method and its improved version

The experimental plan proposed by Morris is composed of individually randomized 'one-factor-at-a-time' experiments: the impact of changing one factor at a time is evaluated in turn. Each input factor may assume a discrete number of values, called levels, which are chosen within the factor range of variation.

The sensitivity measures proposed in the original work of Morris [1] are based on what is called an elementary effect. The elementary effect for the ith input is defined as follows. Let Δ be a predetermined multiple of $1/(p-1)$. For a given value of $\mathbf{x}$, the elementary effect of the ith input factor is defined as

$$EE_i(\mathbf{x}) = \frac{[y(x_1,\ldots,x_{i-1},x_i+\Delta,x_{i+1},\ldots,x_k) - y(\mathbf{x})]}{\Delta}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ is any selected value in Ω such that the transformed point $(\mathbf{x} + \mathbf{e}_i \Delta)$, where $\mathbf{e}_i$ is a vector of zeros but with a unit as its ith component, is still in Ω for each index i=1,..,k. The finite distribution of elementary effects associated with the ith input factor, is obtained by randomly sampling different $\mathbf{x}$ from Ω, and is denoted by F_i.

In Morris [1], two sensitivity measures were proposed for each factor: μ, an estimate of the mean of the distribution F_i, and σ, an estimate of the standard deviation of F_i. A high value of μ indicates an input factor with an important overall influence on the output. A high value of σ indicates a factor involved in interaction with other factors or whose effect is non-linear. Here we consider a third sensitivity measure, μ^*, which is an estimate of the mean of the distribution (here denoted as G_i) of the absolute values of the elementary effects [2].

We believe that μ^* is better than μ to rank factors in order of importance. The reason is that if the distribution F_i contains elements of opposite sign, which occurs when the model is non-monotonic, when computing its mean some effects may cancel each other out. Thus a

factor which is important but whose effect on the output has an oscillating sign may be erroneously considered as negligible, thus generating a mostly undesirable Type II error.

The performance of μ^* is tested on the analytical function presented in Section 3 and compared with that of the variance based methods described in the following subsection.

2.2. The variance based measures

Variance based methods choose as a measure of the main effect of a factor X_i on the output, an estimation of quantity $\frac{V_{X_i}\left(E_{\mathbf{X}_{-i}}\left(Y|X_i\right)\right)}{V(Y)}$, which is known in the literature as the "first order effect" of X_i on Y, and denoted by S_i. Reasons for this choice are detailed in [4].

Another sensitivity measured based on the variance decomposition is the total sensitivity index, S_{T_i}. The total index is defined as the sum of all effects involving the factor X_i. S_{Ti} is estimated by the quantity $\frac{E_{\mathbf{X}_{-i}}\left(V_{X_i}\left(Y|\mathbf{X}_{-i}\right)\right)}{V(Y)}$.

The total index is the appropriate measure when the problem is that of Factors Fixing [4], i.e. that of identifying those factors that can be fixed to any given value within their range of variation because they are non influent on the total output variance. A necessary and sufficient condition for factor X_i to be totally non-influent is that $S_{Ti} = 0$. In fact, if factor X_i is totally non-influent, then all the variance is due to $\mathbf{X}_{-i}$, and fixing this vector results in $V_{X_i}\left(Y|\mathbf{X}_{-i}\right) = 0$, as well as in $E_{\mathbf{X}_{-i}}\left(V_{X_i}\left(Y|\mathbf{X}_{-i}\right)\right) = 0$. The reverse is also true: if $V_{X_i}\left(Y|\mathbf{X}_{-i}\right) = 0$ at all fixed points in the space of $\mathbf{X}_{-i}$, then X_i is non-influent, so that $S_{Ti} = 0$.

Variance based techniques have several desirable properties. They are "model free", in the sense of independent from assumptions about the model such as linearity, additivity and so on. They are global, i.e. they explore the entire interval of definition of each factor and the effect of each factor is taken as an average over the possible values of the other factors. They are usually quantitative, which is they can tell how much factor *a* is more important than factor *b*. They are able to treat grouped factors as if they were single factors, a property of synthesis that may be essential for the agility of the interpretation of the results.

The main drawback of the variance based measures is their computational cost, as they require a number of model evaluation such as $N \times (k+2)$ where k is the number of input factors and N is of the order of N = 500, 5000…, [4] which in some instances may result to be unaffordable. Note that this number can be lowered considerably if one desires to compute only the first order sensitivity indices, as shown by Ratto et al. [6]. Design based strategies to estimate sensitivity indices at low sample size are also proposed in [3].

In this work it is shown that the improved Morris measure represents a valid alternative to the variance based one when the aim of the analysis is that of screening few important factors among a large number, and the cost of applying variance based techniques would be excessive.

3. TESTING THE METHOD

In this Section we propose to test the performance of m* and that of the Morris strategy extended for groups on the analytical function recently proposed by Oakley and O'Hagan [3].

The test function is the following:

$$h(\mathbf{x}) = \mathbf{a}_1^T \mathbf{x} + \mathbf{a}_2^T \cos(\mathbf{x}) + \mathbf{a}_3^T \sin(\mathbf{x}) + \mathbf{x}^T \mathbf{M} \mathbf{x}$$

where $\mathbf{x}$ is a fifteen dimensional input vector while $\mathbf{a_1}$, $\mathbf{a_2}$, $\mathbf{a_3}$ and $\mathbf{M}$ are respectively three (1´ 15) row vectors and a (15´ 15) matrix of parameters (Table 1). The unknown input factors are assumed to be independent and to follow a normal distribution N(0,1). In [3] the emphasis in on computing first order sensitivity measures, and the test case is designed to have three groups of factors, with respectively high (x_{11} - x_{15}), medium (x_6 - x_{10}) and low (x_1 - x_5) values of S_i.

Table 1: Parameters of the analytical function proposed by O'Hagan.

a1	a2	a3
0.01	0.43	0.10
0.05	0.09	0.21
0.23	0.05	0.08
0.04	0.32	0.27
0.12	0.15	0.13
0.39	1.04	0.75
0.39	0.99	0.86
0.61	0.97	1.03
0.62	0.90	0.84
0.40	0.81	0.80
1.07	1.84	2.21
1.15	2.47	2.04
0.79	2.39	2.40
1.12	2.00	2.05
1.20	2.26	1.98

M														
-0.02	-0.19	0.13	0.37	0.17	0.14	-0.44	-0.08	0.71	-0.44	0.50	-0.02	-0.05	0.22	0.06
0.26	0.05	0.26	0.24	-0.59	-0.08	-0.29	0.42	0.50	0.08	-0.11	0.03	-0.14	-0.03	-0.22
-0.06	0.20	0.10	-0.29	-0.14	0.22	0.15	0.29	0.23	-0.32	-0.29	-0.21	0.43	0.02	0.04
0.66	0.43	0.30	-0.16	-0.31	-0.39	0.18	0.06	0.17	0.13	-0.35	0.25	-0.02	0.36	-0.33
-0.12	0.12	0.11	0.05	-0.22	0.19	-0.07	0.02	-0.10	0.19	0.33	0.31	-0.08	-0.25	0.37
-0.28	-0.33	-0.10	-0.22	-0.14	-0.14	-0.12	0.22	-0.03	-0.52	0.02	0.04	0.36	0.31	0.05
-0.08	0.004	0.89	-0.27	-0.08	-0.04	-0.19	-0.36	-0.17	0.09	0.40	-0.06	0.14	0.21	-0.01
-0.09	0.59	0.03	-0.03	-0.24	-0.10	0.03	0.10	-0.34	0.01	-0.61	0.08	0.89	0.14	0.15
-0.13	0.53	0.13	0.05	0.58	0.37	0.11	-0.29	-0.57	0.46	-0.09	0.14	-0.39	-0.45	-0.15
0.06	-0.32	0.09	0.07	-0.57	0.53	0.24	-0.01	0.07	0.08	-0.13	0.23	0.14	-0.45	-0.56
0.66	0.35	0.14	0.52	-0.28	-0.16	-0.07	-0.20	0.07	0.23	-0.04	-0.16	0.22	0.00	-0.09
0.32	-0.03	0.13	0.13	0.05	-0.17	0.18	0.06	-0.18	-0.31	-0.25	0.03	-0.43	-0.62	-0.03
-0.29	0.03	0.03	-0.12	0.03	-0.34	-0.41	0.05	-0.27	-0.03	0.41	0.27	0.16	-0.19	0.02
-0.24	-0.44	0.01	0.25	0.07	0.25	0.17	0.01	0.25	-0.15	-0.08	0.37	-0.30	0.11	-0.76
0.04	-0.26	0.46	-0.36	-0.95	-0.17	0.003	0.05	0.23	0.38	0.46	-0.19	0.01	0.17	0.16

The total variance of the output can be computed analytically and decomposed as the sum of first and second order effects:

$$V = \sum_i V_i + V^{2nd}$$

$$V_i = a_{1,i}^2 * E(x_i^2) + a_{2,i}^2 * E(\sin^2(x_i)) + a_{3,i}^2 * E(\cos^2(x_i)) + M_{i,i}^2 * E(x_i^4) +$$
$$2 * a_{1,i} * a_{2,i} * E(x_i \sin(x_i)) - \left[a_{3,i} * E(\cos(x_i)) + M_{i,i} * E(x_i^2)\right]^2$$

$$V^{2nd} = \sum_{i=1}^{15} \sum_{\substack{j=1 \\ j>i}}^{15} (\mathbf{M} + \mathbf{M}^T)_{i,j}^2$$

where $E(\cdot)$ indicates the mean operator. All the terms of higher order are zero.

Table 2 shows the rank of importance for the 15 input factors according to the revised Morris measure m^* and to the Sobol' total effect index. The analytical values of the total effects are also reported. The total number of model evaluations needed to estimate each set of measures is reported in the first row. Results confirm that, with just 1024 model evaluations, the Morris revised measure is capable of identifying the subset of important factors (x_{11} - x_{15}). Note that when the total sensitivity indices are used, the factors end up partitioned in just two sets, that of the most influential factors (x_{11} - x_{15}), and that of the less influential ones (all others). Strictly speaking, none of the input factors of this test case can be fixed unless a rather high threshold is imposed. The least important factor's bottom marginal variance is in fact as high as 2.6 %. Four factors could be fixed if the threshold were 5%, while ten could be fixed if the threshold were 10%.

Table 2: Sensitivity analysis results for the test function in [3]. The analytical values of the total indices are reported together with the Sobol' estimates. The correspondent ranks are compared with that obtained through the Morris experiment.

Factor	ST(i) Analytics	ST(i) N=65563	Analytics Rank	ST(i) Rank N=65563	Morris Rank N=1024
X1	0.059	0.034	9	11	8
X2	0.063	0.032	8	12	9
X3	0.036	0.026	13	14	12
X4	0.055	0.035	11	10	10
X5	0.026	0.01	15	15	15
X6	0.041	0.038	12	9	13
X7	0.058	0.047	10	8	11
X8	0.082	0.067	7	7	7
X9	0.097	0.073	6	6	5
X10	0.036	0.027	14	13	14
X11	0.151	0.14	2	5	3
X12	0.148	0.172	3	2	2
X13	0.142	0.152	4	3	4
X14	0.141	0.143	5	4	6
X15	0.155	0.175	1	1	1

4. THE FINANCIAL PROBLEM

The problem is that of pricing a European call option. Different scenarios are assumed, corresponding to different possible strike prices and times to maturity. The dynamic of the underlying stock price is modeled according to the Heston Stochastic Volatility model (HEST, [5]), where the stock price follows the Black-Scholes stochastic differential equation SDE in which the volatility behaves stochastically over time:

$$\frac{dS_t}{dt} = (r - q)dt + s_t dW_t \qquad S_0 \, ³ \, 0 .$$

The (squared) volatility follows the Cox-Ingersoll-Ross process:

$$ds_t^2 = k(h - s_t^2)dt + q s_t d\tilde{W}_t \qquad s_0 \geq 0,$$

where $W = \{W_t, t \geq 0\}$ and $\tilde{W} = \{\tilde{W}_t, t \geq 0\}$ are two correlated standard Brownian motions such that $Cov[dW_t, d\tilde{W}_t] = r\,dt$.

Here we also consider an extension of the HEST model that introduces jumps in the asset price [5]. Jumps here are assumed to occur as a Poisson process and the percentage jump-sizes are log-normally distributed.

In the Heston Stochastic Volatility model with jumps (HESJ), the SDE of the stock price process is extended to yield:

$$\frac{dS_t}{dt} = (r - q - l m_J)dt + s_t dW_t + J_t dN_t \qquad S_0 \geq 0,$$

where $N = \{N_t, t \geq 0\}$ is an independent Poisson process with intensity parameter $l > 0$, i.e. $E[N_t] = lt$. J_t is the percentage jump size (conditional on a jump occurring) that is assumed to be log-normally, identically and independently distributed over time, with unconditional mean m_J. The standard deviation of $\log(1 + J_t)$ is s_J:

$$\log(1 + J_t) \sim N\left(\log(1 + m_j) - \frac{s_J^2}{2}, s_J^2\right).$$

The SDE of (squared) volatility process remains unchanged. J_t and N are assumed to be independent, as well as of W and of ~ W.

Sensitivity analysis is performed first on the HEST model and then on its extended version with jumps HESJ. For HEST the input variables considered in the analysis are s_0, k, h, q and r. In the case where jumps are present l, m_J, s_J are added. The initial condition for the underlying price S_0 is fixed at 100, while the interest rate *r* and the dividend yields *q* of the stock are respectively at 1.9% and 1.2%. The distributions chosen for the inputs are listed in Table 3. Both the Morris measure μ* and the total sensitivity indices S_{T_i} are computed for the input factors in 42 different scenarios, a scenario being determined by a different value of the option strike price and of the time to maturity.

Table 3: Distributions for the inputs of the HEST and HESJ models.

Input	Distribution	Minimum	Maximum	Input	Distribution	Minimum	Maximum
s_0	Uniform	0.04	0.09	r	Uniform	-1	0
k	Uniform	0	1	l	Uniform	0	2
h	Uniform	0.04	0.09	mj	Uniform	-0.1	0.1
q	Uniform	0.2	0.5	sj	Uniform	0	0.2

Tables 4 and 5 show the ranking of the input factors obtained according to the two measures for the two versions of the model, HEST and HESJ, in some of the scenarios. For the variance based method, which is a quantitative method (as each index represent the fraction of the output variance due to the effect of that factor), we also reported values of the indices. The total number of model executions for the total sensitivity indices is 20480. For

the Morris experiment four levels are considered and 60 model executions performed to obtain the distribution of elementary effects for each input. The variance based method has also been repeated doubling the sample size to verify the convergence of the obtained sensitivity indices. Results confirm that the sample size 20480 can be considered sufficient for the estimation of the indices.

Table 4: Sensitivity analysis results of the HEST model for six selected scenarios

	ST(i) N=14336	**ST(i) Rank N=14336**	**Morris Rank N=60**	**ST(i) N=14336**	**ST(i) Rank N=14336**	**Morris Rank N=60**	**ST(i) N=14336**	**ST(i) Rank N=14336**	**Morris Rank N=60**
	Strike = 80			**Strike = 100**			**Strike = 120**		
	Time to maturity = 1y								
s_0	0.742	1	1	0.821	1	1	0.411	2	2
k	0.026	4	5	0.050	4	4	0.033	5	5
h	0.055	3	3	0.075	3	3	0.045	4	4
q	0.013	5	4	0.084	2	2	0.110	3	3
r	0.194	2	2	0.009	5	5	0.448	1	1
	Time to maturity = 3y								
s_0	0.493	1	1	0.348	1	1	0.201	3	2
k	0.158	3	3	0.273	2	2	0.231	2	3
h	0.362	2	2	0.267	3	3	0.182	4	4
q	0.065	4	4	0.165	4	4	0.149	5	5
r	0.042	5	5	0.049	5	5	0.324	1	1

Table 5: Sensitivity analysis results of the HESJ model for six selected scenarios.

	ST(i) N=2080	**ST(i) Rank N=2080**	**Morris Rank N=90**	**ST(i) N=2080**	**ST(i) Rank N=2080**	**Morris Rank N=90**	**ST(i) N=2080**	**ST(i) Rank N=2080**	**Morris Rank N=90**
	Strike = 80			**Strike = 100**			**Strike = 120**		
	Time to maturity = 1y								
s_0	0.342	2	3	0.317	2	3	0.191	3	3
k	0.012	7	6	0.019	7	7	0.013	8	8
h	0.025	6	5	0.028	6	5	0.018	7	7
q	0.001	8	8	0.034	4	4	0.055	5	5
r	0.060	4	4	0.010	8	8	0.151	4	4
l	0.264	3	1	0.308	3	1	0.297	2	1
m_j	0.040	5	7	0.030	5	6	0.044	6	6
s_j	0.366	1	2	0.379	1	2	0.369	1	2
	Time to maturity = 3y								
s_0	0.166	3	3	0.125	3	3	0.086	4	5
k	0.061	5	5	0.092	5	6	0.084	5	7
h	0.120	4	4	0.097	4	4	0.076	6	6

q	0.028	7	6	0.060	6	5	0.065	7	4
r	0.007	8	8	0.032	7	7	0.118	3	3
l	0.314	2	1	0.316	2	1	0.307	2	1
m_J	0.031	6	7	0.030	8	8	0.032	8	8
s_J	0.424	1	2	0.404	1	2	0.388	1	2

From both Tables it emerges that the rankings obtained with μ* and with S_{Ti} are very similar in each of the scenarios, and in some cases even identical (especially for the HEST model), confirming reliability of the results.

The few cases where Morris inverts the ranking of two factors are those where their sensitivity indices values are very similar. In the worst case (in all 84 simulations we performed) the Morris design inverts 2 factors whose difference in the total indices represents nearly 18% of the total output variance. In general Morris can be considered successful in its goal of screening a subset of factors that can be fixed, as it never confounds groups of important and unimportant factors. If a factor is high ranked according to S_{Ti} it is also high ranked for Morris and vice versa.

The Morris method has the great advantage of a low computational cost. However, as a drawback it is not quantitative; the value of its measures can only be used to rank factors but cannot be interpreted as percentages of output variance. For this reason the S_{Ti} indices are used for analyzing the behavior of each input factors in different scenarios for example in the case of absence of jumps (Fig.1). In Figure 1 each dot refers to a scenario. The dot size highlights the importance of the factor in that scenario. The following conclusions can be drawn from the Figure:

- The three model parameters k, h, q are not very relevant at low times to maturity, but their importance increases with increasing the time horizon.
- The initial condition s_0 is the most important factor when the time to maturity is rather small and its importance decreases with time.
- The correlation r is also an influential parameter, especially when the option is not at-the-money.

When jumps are present the same conclusions can be drawn. Moreover the overall influence on the model outcomes of the three parameters related to jumps is relatively high, confirming the importance of the jumps inclusion. In particular, results show that l and s_J are very much influential at all time horizons and strike prices (they are always among the three most important factors).

We also applied the Morris method to work with groups. Four groups were considered: the group of model parameters (k, h, q) relative to modeling the stochastic volatility; the initial condition s_0, the group of model parameters (l, m_J, s_J) relative to jumps; and the correlation r. Results are plotted in Figure 2. In the plots the relative importance of the four groups is shown for all the considered scenarios. The total number of model evaluations for each scenario is N=50

The group of the jumps' parameters results to be always the most important, while the influence of s_0 and r depends upon the scenario characteristics. As expected, the group (k, h, q) is negligible at low times to maturity, but its importance increases with time horizon.

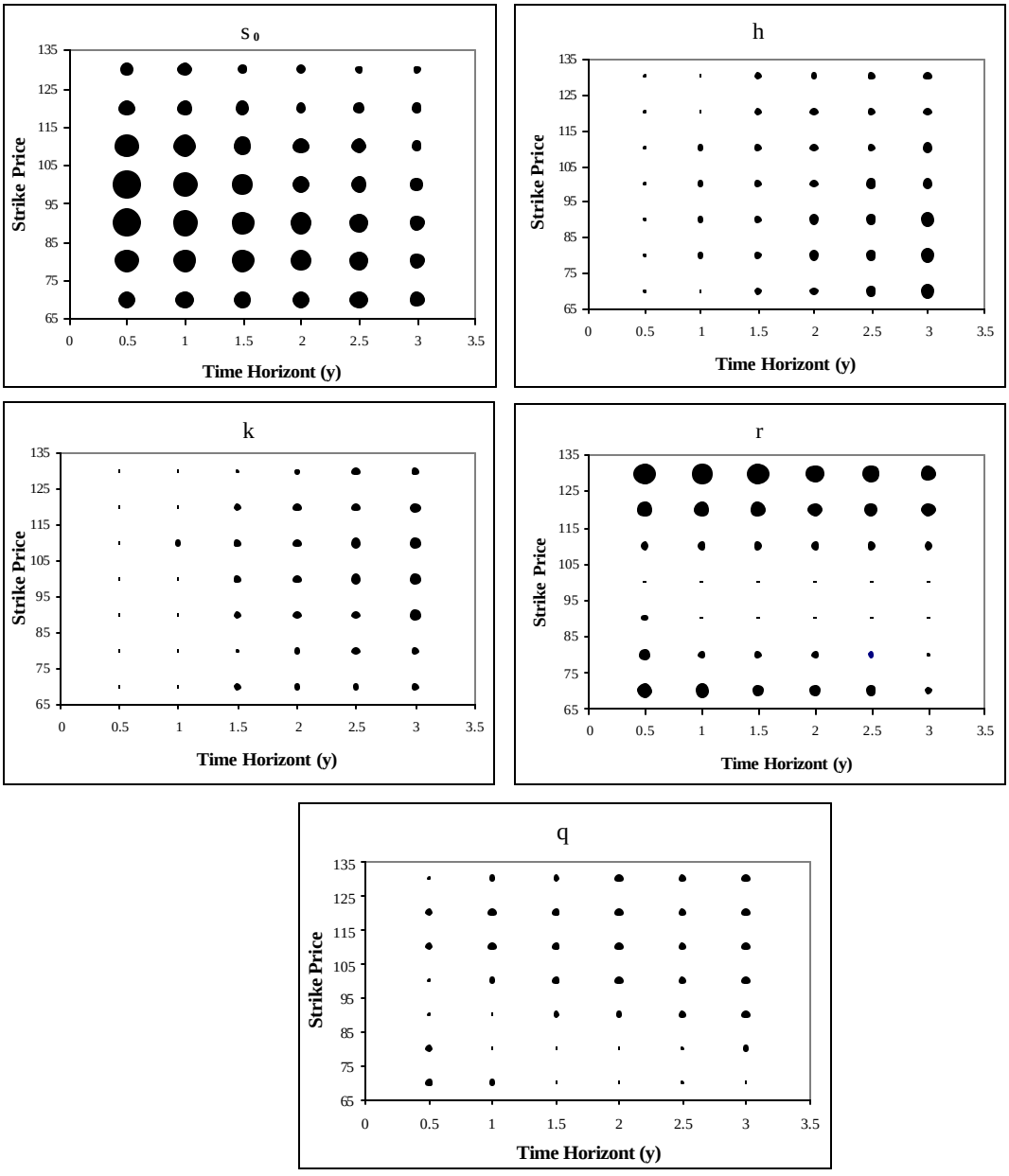
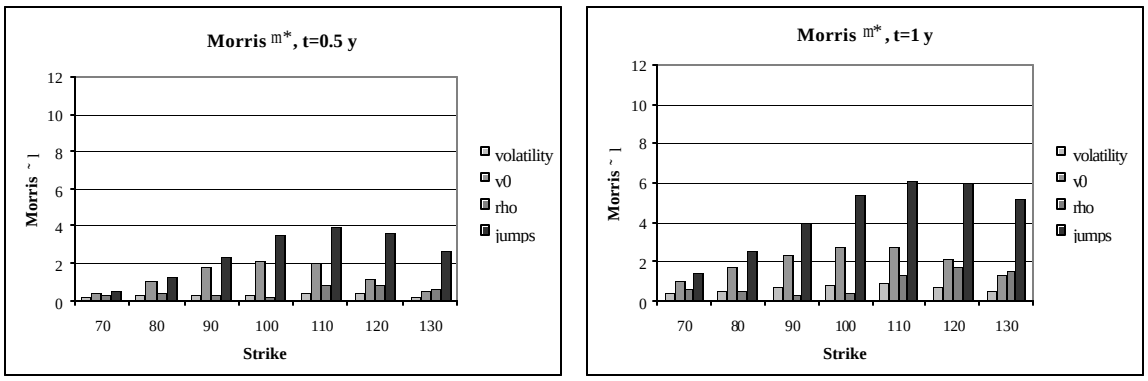


Figure 1: S_{Ti} results for the HEST model. The differences in the size of the dots represent the differences in the importance of the fixed input factors in all the considered scenarios.



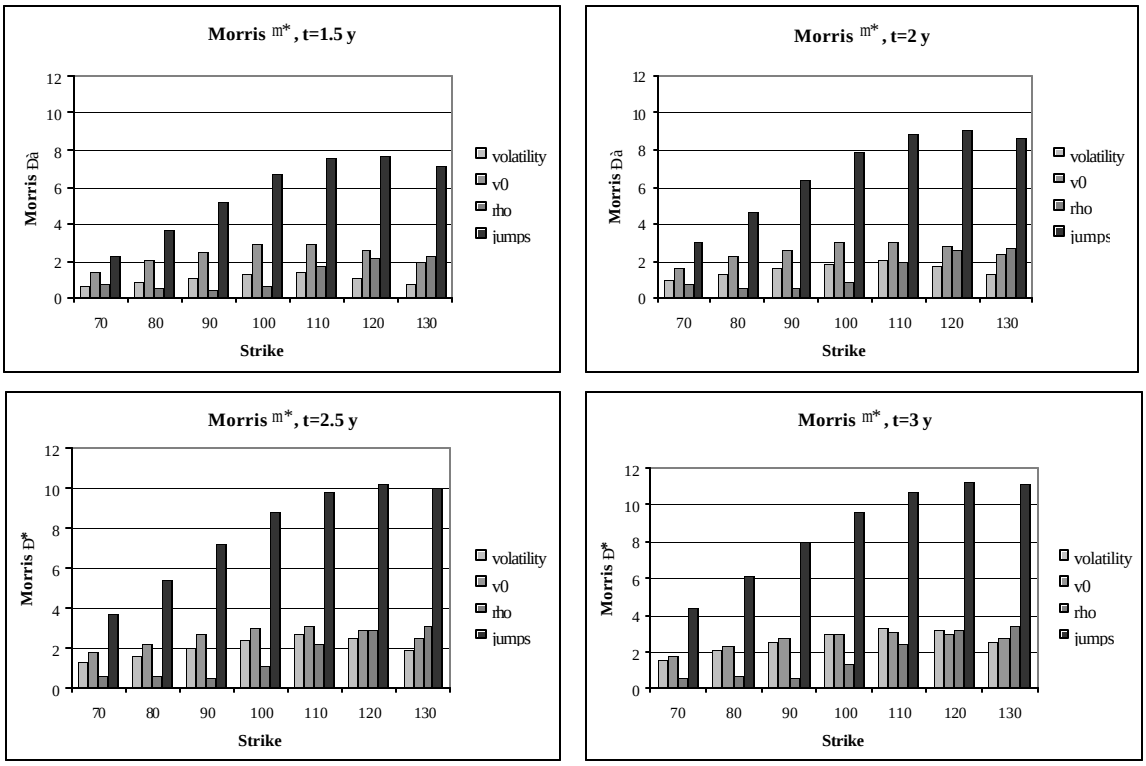


Figure 2: Screening results obtained through the Morris method for the HESJ model. 50 is total number of model evaluations for the Morris experiments. The bars plot the Morris revised μ^*, which can be used to screen the negligible factors in the model.

5. CONCLUSIONS

In this work we have confirmed the capability of the sensitivity measure μ^*, an improved version of the Morris measure introduced by Campolongo et al. [2], to distinguish between important and negligible model input factors at low computational cost. Also the updated measure has proved to be effective when factors are grouped.

Results of sensitivity analysis on the Heston model for pricing European option has allowed to concluding that jumps play a major role in determining the option price, thus stressing the need of including them in the model formulation. Furthermore results have underlined that, as expectable, at low time to maturity the initial condition for volatility needs to be accurately determined as the resulting option price is highly affected by its value. Its importance decreases as the time to maturity increases. Finally, it emerged that the correlation between the two Brownian motions needs to be carefully defined, especially when the option is not at the money, while the other model parameters are less important.

REFERENCES

1. M. D. Morris *Factorial Sampling Plans for Preliminary Computational Experiments*, *Technometrics*, **33**, 161-174, 1991.
2. F. Campolongo, J. Cariboni, and A. Saltelli. *Sensitivity analysis: the Morris method versus the variance based measures*. Submitted to Technometrics, 2003.

3. J. Oakley, and A. O'Hagan. *Research Report* No. 525/02 Department of Probability and Statistics, University of Sheffield. http://www.shef.ac.uk/st1jeo/#_Papers. Submitted to Journal of the Royal Statistical Society, Series B, 2003.
4. A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*. John Wiley & Sons publishers, Probability and Statistics series, to appear March 2004.
5. W. Schoutens, E. Simons, J. Tistaert. *A Perfect calibration! Now What?* 2003
6. M. Ratto , S. Tarantola, A. Saltelli, P. C. *Young Accelerated estimation of sensitivity indices using state dependent parameter models*, SAMO 2004

Initial Evaluation of Pure and "Latinized" Centroidal Voronoi Tessellation for Non-Uniform Statistical Sampling*

Vicente J. Romero[§]
Sandia National Laboratories†
Albuquerque, NM

John V. Burkardt, Max D. Gunzburger[◊]*, Janet S. Peterson*
School of Computational Science and Information Technology
Florida State University, Tallahassee, FL

Abstract

A recently developed Centroidal Voronoi Tessellation (CVT) sampling method is investigated here to assess its suitability for use in statistical sampling applications. CVT efficiently generates a highly uniform distribution of sample points over arbitrarily shaped M-Dimensional parameter spaces. On several 2-D test problems CVT has recently been found to provide exceedingly effective and efficient point distributions for response surface generation. Additionally, for statistical function integration and estimation of response statistics associated with uniformly distributed random-variable inputs (uncorrelated), CVT has been found in initial investigations to provide superior points sets when compared against Latin-Hypercube and Simple-Random Monte Carlo methods and Halton and Hammersley quasi-Monte-Carlo sequence methods. In this paper, the performance of all these sampling methods and a new variant ("Latinized" CVT) are further compared for non-uniform input distributions. Specifically, given uncorrelated normal inputs in a 2-D test problem, statistical sampling efficiencies are compared for resolving various statistics of response: mean, variance, and exceedence probabilities.

Keywords: Centroidal Voronoi tessellation, statistical sampling methods, uncertainty propagation

1. Introduction and Background

It is often beneficial in statistical sampling and function integration to sample "uniformly" over the applicable parameter space. Such uniformity, while conceptually simple and intuitive on a qualitative level, is on a quantitative level somewhat complicated to describe and characterize mathematically. Quantitative aspects of uniformity involve: 1) the equality with which points are spaced relative to one another in the parameter space (are they all nominally the same distance from one another?); 2) uniformity of point density over the entire domain of the parameter space (i.e., uniform "coverage" of the whole domain by the set of points, and not just good uniformity within certain regions of the space); and 3) isotropy in the point placement pattern. Each of these aspects of uniformity can be quantified by several mathematical measures as described in reference [2]. We will not discuss these measures further here, but mention them to indicate that quantitative measures do exist for the notion of uniformity.

---

*This paper is declared a work of the United States Government and is not subject to copyright protection in the U.S.

§corresponding author email contact: vjromer@sandia.gov

†Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

◊M. Gunzburger also supported in part by the Computer Science Research Institute, Sandia National Laboratories, under contract 18407.

We find that for 2-D data sets the eye is an excellent integrator of the different aspects of uniformity listed above. The intuitive sense of uniformity obtained from viewing sample sets in a unit square (2-D hypercube) usually correlates very strongly with the quantitative measures. Thus, for 2-D data sets like the ones we present later, fairly accurate visual judgements can be made about whether one particular layout of sample points is mathematically more uniform than another, or whether the uniformity varies significantly over the parameter space.

Achieving high sampling uniformity over generic domains is an area of active research. Much effort has been applied to the problem of achieving uniform placement of N samples over M-dimensional hypercubes, where M and N are both arbitrary. It is well recognized that Simple-Random sampling (SRS) Monte Carlo does not do a particularly good job of uniformly spreading out the sample points. The popular Latin Hypercube Sampling (LHS) method ([5]) generally does a much better job of uniformly spreading out the points. This is due to the greater sampling regularity over each individual parameter dimension before the individually generated parameter values are randomly combined into parameter sets which define the coordinates of the sampling points.

Recent efforts to modify LHS to get an even more uniform distribution of points over the parameter space have included Distributed Hypercube Sampling (DHS, [12]) and Improved [Distributed] Hypercube Sampling (IHS, [1]). The fundamentals and history of these are reviewed briefly in [18]. Though the quantitative measure of uniformity used for comparisons in [1] and [12] was somewhat flawed, it does appear that DHS gives better sampling uniformity than LHS, and IHS gives better sampling uniformity than DHS (but is increasingly more computationally expensive as the dimensionality of the parameter space increases). We have recently become aware of another LHS variant, "Optimal Symmetric LHS" (OSLHS, [21]) which also seems to improve the spatial uniformity of LHS samples. Its computational cost and performance relative to DHS and IHS are not yet known, however.

A number of other potential approaches for achieving uniform point placement that are not evolved from an LHS basis are reviewed (and some new ones are presented) in [7]. There, some quantitative metrics related to visual/sensory perception of point uniformity in 2-D are reviewed and some new ones are presented. Many of these non-LHS-based approaches appear to work very well in 2-D, but it is said in [7] that some of the methods may not be applicable or may not perform well in more than two dimensions, and some clearly will not scale up to high dimensions affordably. Others seem more promising for high dimensions, but have not yet been investigated enough.

The so-called "Quasi- Monte Carlo" (QMC, see *e.g.* [14]) sub-random low-discrepancy sequence methods can often achieve reasonably uniform sample placement in hypercubes. The strength of these sequence methods (Halton, Hammersley, Sobol, etc.), is that they can produce fairly uniform point distributions even though samples are added one at a time to the parameter space. The one-at-a-time incremental sampling of QMC (and SRS) enables these methods to have better efficiency prospects than CVT and LHS-type methods in the area of error estimation and control. Not only this, the results achieved are often quite good. For resolving the mean and standard deviation of response measures, Hammersley sequences were found in [11] to converge to within 1% of exact results 3 to 100 times faster than LHS over a large range of test problems. For resolving response probabilities, Hammersley and modified-Halton were found in [15] to perform roughly the same as LHS on balance over several test problems.

However, when the hyperspace dimension becomes moderate to large and/or the sampling density becomes high, some (perhaps all?) sequences suffer from spurious correlation of the samples. This is shown for standard Halton sequences in 16-D (ref. [12]) and 40-D (ref. [15]). Sometimes a modification can be found to suppress or delay the onset of spurious correlation – as a fix from the literature implemented in [15] shows for Halton sequences.

Recently, a long-recognized approach for achieving uniformity of point placement in M-dimensional volumes, called "**Centroidal Voronoi Tessellation**" (CVT), has been made computationally efficient ([10]) for implementing the principles of Centroidal Voronoi diagrams ([6],[13]). These diagrams subdivide arbitrarily shaped domains in arbitrary-dimensional space into arbitrary numbers of nearly uniform subvolumes, or Voronoi cells/regions. Given a set of N points $\{z_i\}$ (i=1,...,N) in an M-dimensional hypercube, the Voronoi region or Voronoi cell V_j (j=1,...,N) corresponding to z_j is defined to be all points in the hypercube that are closer to z_j than to any of the other z_i's. The set $\{V_i\}$ (i=1,...,N) is called a Voronoi tessellation or Voronoi diagram of the hypercube, the set $\{z_i\}$ (i=1,...,N) being the generating points or generators. A *centroidal* Voronoi tessellation (CVT) is a special Voronoi tessellation with the property that each generating point z_i is itself the mass centroid of the corresponding Voronoi region V_i.

Although CVTs are deterministic, they can be converged to with probabilistic sampling methods. In [10], new probabilistic CVT construction algorithms were introduced, implemented, and tested. These methods are generally much more computationally efficient than previous deterministic and probabilistic methods for constructing CVTs.

The CVT concept and the algorithms in [10] for their construction can be generalized in many ways (see [3] for details). For example, instead of a hypercube, general regions in M-dimensional space can be treated. This feature has been exploited with great success (see [6]) for discretizing arbitrary 2-D and 3-D domain volumes for computational mechanics analysis with meshless analogues of finite element methods. Furthermore, points can be distributed non-uniformly according to a prescribed density function over the space (like the bi-normal density function that Figure 7 corresponds to).

In initial investigations ([2]) for 2-D, 7-D, and 20-D test cases, CVT has provided greater sampling uniformity than Halton, Hammersley, Sobol, SRS, LHS, DHS, and IHS according to a meaningful subset of non-flawed quantitative quality measures. Additionally, no degradation of sampling uniformity has been detected in higher dimensions (*i.e.*, for the 20-D case).

It is therefore natural to ask whether CVT can be applied for: A) statistical sampling over arbitrary-dimensional spaces of input random variables to calculate various statistics of output response behavior; B) function integration over arbitrarily shaped domains; and C) whether it can serve as a method for generating favorable point distributions for improved response-surface accuracy.

A preliminary positive indication regarding item C) for response surface generation is presented in [18]. There, CVT was shown on several 2-D test problems to provide superior point distributions for generating locally-conforming Moving Least Squares response surfaces. Point distributions by CVT, SRS, LHS, and a structured sampling method with deterministically uniform point placement ([17]) were tried in the study.

Reference [19] compared the above sampling methods for sampling performance in 2-D test problems of statistical function integration and estimation of response statistics associated with

uniformly distributed random-variable inputs (uncorrelated). By the same weighted measure of sampling effectiveness defined and used in Section 3.3 of this paper, CVT handily outperformed SRS, LHS, Halton, and Hammersley in resolving various statistics of response: mean, variance, and exceedence probabilities.

In this paper we take a first step toward examining the potential of CVT for improved statistical sampling given *non*-uniform inputs. Specifically, the performance of the above sampling methods and a new CVT variant ("Latinized" CVT) are compared for non-uniform uncorrelated input distributions in a 2-D test problem. Statistical sampling efficiencies are compared for calculating response mean, variance, and exceedence probabilities.

2. Uniformly Distributed Test Point-Sets and Their Mapping to Binormal Joint Densities

Figure 1 shows three LHS and three corresponding CVT point sets for 100 samples in a 2D unit hypercube. The three LHS point sets were generated with the software [9] for different initial seeds (Seed1 = 123456789, Seed2 = 192837465, Seed3 = 987654321) and a uniform joint probability density function (JPDF) over the unit-square parameter space. The three corresponding CVT point sets were generated with the software [4] by using the LHS sets as initial conditions (starting point locations) from which the CVT iterations begin. In all cases each CVT set is much more uniform visually (and quantitatively, see [2]) than its associated LHS set. All three CVT sets are relatively similar visually and quantitatively, even though starting from three very different initial conditions given by the LHS sets.

The LHS sets exhibit significantly more clustering and non-uniformity of the points than the CVT sets. For a visual indicator of sampling uniformity, Figure 5 compares a 25-sample LHS set and a 25-sample CVT set started from the LHS set. Non-overlapping circles are drawn in each domain, where each sample point has a circle centered about it having a radius proportional to the distance from the point to its nearest neighbor. The surrounding circles for the CVT set are all fairly uniform in size, whereas the variance in circle size is very large for the LHS set. Thus, the LHS point sets are relatively non-uniform in their "coverage" of the domain.

Besides the three LHS sets and three corresponding CVT sets shown in Figure 1, three SRS sets generated from initial seeds 1, 2, and 3 will also be tested here. These point sets can be seen in reference [18]. They exhibit even less uniformity than the LHS sets in Figure 1. Three CVT sets derived from the three different SRS sets as initial conditions can also be seen in [18]. The different LHS and SRS initial conditions do not have much of an impact on final CVT point uniformity, so the CVT algorithms appear to be robust in this regard.

Figures 2 and 3 show Halton and Hammersley point sets and the corresponding CVT sets derived from them. Again, the resulting CVT sets are of essentially equivalent uniformity. The Halton point set is noticeable and quantitatively more uniform than any of the LHS sets; the Hammersley set is even more uniform than the Halton set; and the CVT sets in Figures 1, 2, and 3 are even more uniform than the Hammersley set.

In reference [19] we compared the mentioned point sets for effectiveness in 2-D test problems of statistical function integration and estimation of response statistics for the case of uniformly distributed input random variables (uncorrelated). The CVT point sets performed best, as will be summarized in Section 4 of this paper. In this paper we focus on comparing the per-

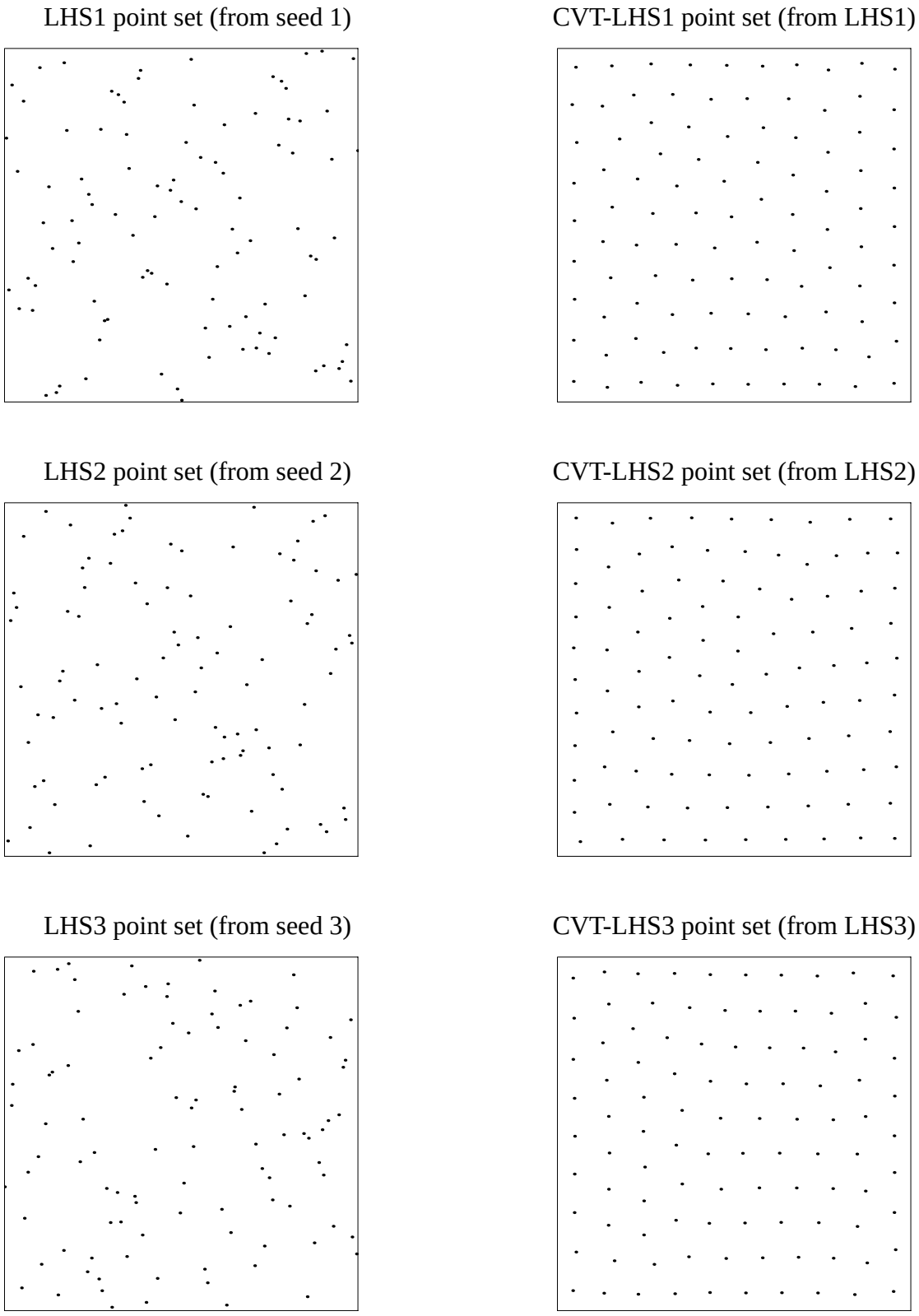


Figure 1. 100-point sample sets on a 2-D unit hypercube for: A) Left Column– uniform JPDF LHS Monte Carlo with three different initial seeds; and B) Right Column– corresponding uniform JPDF CVT sets starting from LHS sets as initial conditions.

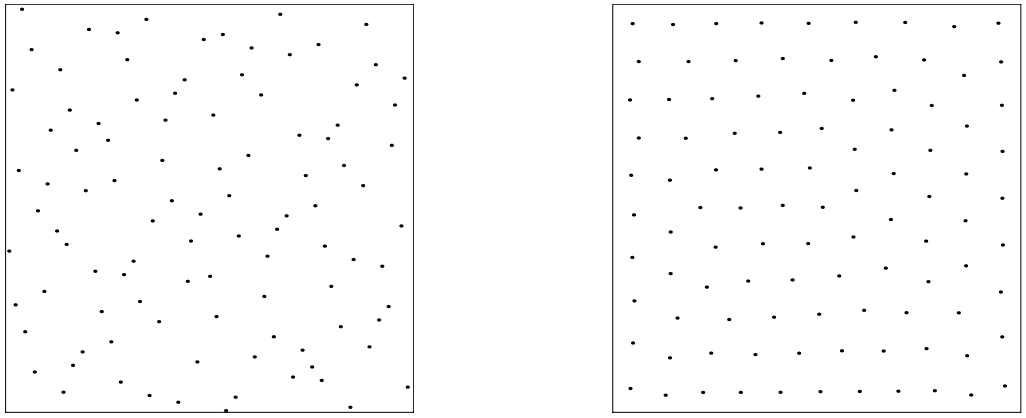


Figure 2. 100-point sample sets on 2-D unit hypercube for:
A) Left plot– Halton QMC sequence;
B) Right plot– corresponding CVT set starting from the Halton set as initial conditions.

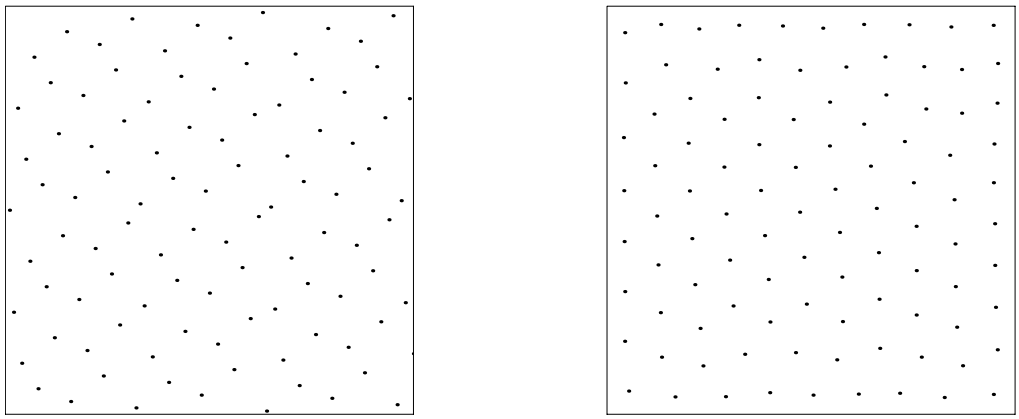


Figure 3. 100-point sample sets on 2-D unit hypercube for:
A) Left plot– Hammersley QMC sequence;
B) Right plot– corresponding CVT set starting from the Hammersley set as initial conditions.

formance of the above sampling methods and a new CVT variant ("Latinized" CVT, see [2]) as starting sets for mapped non-uniform point distributions intended to reflect a JPDF of uncorrelated normal inputs.

Our 2-D test problem has two random inputs *p1* and *p2* from independent normal distributions having means 0.5 and standard deviations σ=0.5/3. The corresponding JPDF is shown in Figure 4 after truncation of the function beyond the unit *p1-p2* parameter space and renormalization to integrate to one over the space.

The following procedure is used to map a set of uniformly distributed points to a set that reflects the desired non-uniform JPDF. First, for each random variable *p* in the problem, we consider its cumulative distribution function CDF(*p*), where

$$\mathrm{CDF}(p) = \int_{0}^{p \le 1} \mathrm{PDF}(p')dp' \qquad \text{EQ 1}$$

and PDF(p) is the probability density function of the random input p. We note that the value of the CDF ranges from 0 to 1 as the coordinate p ranges from 0 to 1 over our unit hypercube domain. It can also be shown that realizations $\{p_i\}$ drawn at random from a density function PDF(p) map through EQ 1 (setting $p=p_i$) into a uniformly distributed set of realizations $\{\mathrm{CDF}(p_i)\}$. This set is therefore distributed uniformly between 0 and 1.

Hence, we recognize that the above properties can be used to inverse-map numbers uniformly distributed between 0 and 1 (produced, e.g., by a random number generator), into realizations $\{p_i\}$ that would appear to be drawn from the density function PDF(p). In a multidimensional problem, we inverse-map the coordinates of points uniformly distributed in the hypercube into transformed point sets that reflect the individual or "marginal" PDFs of the random inputs contributing to the Joint PDF. Figure 5 helps visualize the multidimensional mapping process. The tick marks on the coordinate axes indicate the projections of the points onto the coordinate axes. The (hopefully) uniformly distributed tick marks ranging from 0 to 1 on each coordinate axis give the random values that are inverse-mapped through the marginal CDFs into transformed tick locations ranging from 0 to 1 on the coordinate axes of the JPDF space. Thus, uniformly distributed points in a unit hypercube are transformed to new locations in the unit hypercube. The transformed coordinate sets define point locations distributed according to the target JPDF. (Our transformation algorithm for mapping uncorrelated uniformly distributed points sets to binormally distributed point sets was verified as described in Section 3.2.) Correlation between random variables can be imparted with the rank correlation procedure described in [8].

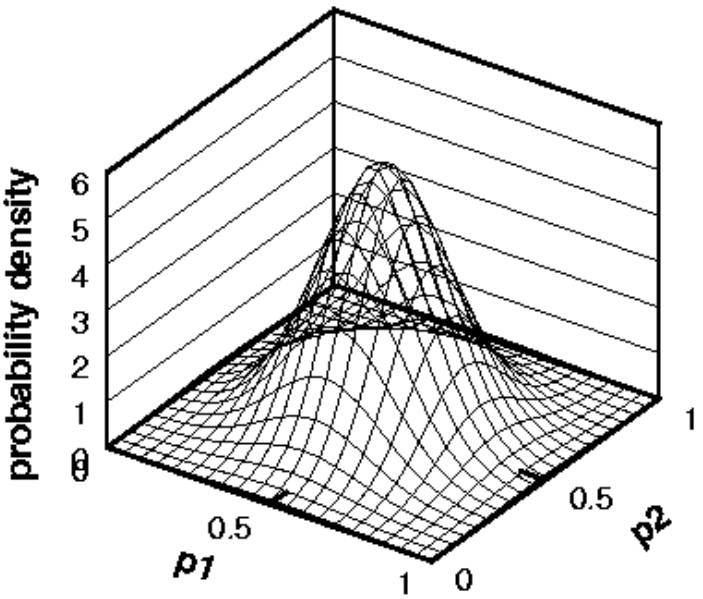


Figure 4. Joint Probability Density Function describing the random variables in the problem: normally distributed parameters *p1* and *p2* with means 0.5, std. deviations σ=0.167, and truncation of the unit square parameter space at 3σ above and below the mean values.

Figure 5. LHS and CVT sample sets showing relative uniformities of point spacing and discrepancies of point projections onto coordinate axes.

The mapping transformation presupposes a point set in an M-dimensional unit hypercube with point locations that project with uniform spacing onto all coordinate axes. However, consider the point sets in Figure 5. Though the CVT set is more uniform *volumetrically* than the LHS set, the LHS points clearly project more uniformly onto the coordinate axes. The projections of the CVT points occur in clusters that portray a "banded" distribution over the 0 to 1 range on each axis, as opposed to the desired uniform distribution. In the limit of a perfectly volumetrically uniform distribution of points over the domain, say a 5x5 rectangular array of points on the unit square, the points would project onto the coordinate axes making 5 uniformly spaced tick marks. These marks would inverse map through the marginal CDFs into only 5 different values or samples of each input variable. Thus, out of twenty-five sampling opportunities, each input variable is sampled at only five values. However, this is not *automatically* bad; the 25 particular *sets* or combinations of the five values of each input variable (when the uniform 5x5 grid of points is mapped to the JPDF space) may pose certain advantages over other point layouts. We are presently striving to understand the particular benefits and disadvantages that arise here.

The LHS point set, on the other hand, would sample each of the input variables at 25 different values. By the nature of LHS ([5]), a sample value would be picked at random from within each of the 25 equal intervals on the 0 to 1 range of each marginal CDF. These would map to 25 points in the JPDF space that each sample a different value of the input variables.

One measure of a point set's uniformity of projection onto all the coordinate axes is called its *discrepancy.* As uniformity increases, discrepancy decreases. LHS is a lower-discrepancy sampling method than CVT is. Methods specifically designed with low discrepancy in mind are the quasi- or sub- random low-discrepancy sequence methods Halton, Hammersley, Sobol, etc. ([14]). These can have both lower discrepancy than standard LHS *and* higher volumetric uniformity. Though CVT tends to have better volumetric uniformity than the sequence methods, which helps its relative performance in other areas (*cf.* [18], [19]), it also has much higher discrepancy, which hurts its relative performance as a sampling basis for non-uniformly-random

inputs. Therefore, a hybrid of CVT and LHS has recently been formulated ([2]) with appears to have both lower discrepancy than pure CVT and higher volumetric uniformity than pure LHS. In the next section we compare the performance of this hybrid "Latinized" CVT (LCVT) against pure CVT and the other sampling methods.

Figure 6 shows uniformly distributed point sets from SRS, LHS, CVT, LCVT, Halton, and Hammersley methods, and corresponding mapped bi-normal point sets. The SRS, LHS, CVT, and LCVT uniform and mapped sets are typical of the three sets obtained from three different initial seeds described at the start of this Section. The SRS, LHS, CVT, and LCVT sets plotted in Figure 6 correspond to Seed 1. Our mapping process was checked by verifying that our bi-normal results mapped from the Seed 1, 2, and 3 uniform LHS sets in Figure 1 were essentially identical to bi-normal LHS sets generated directly from the LHS code ([8]) that produced the three uniform LHS sets. Thus, our mapping process corresponds almost exactly to the mapping process used in the well-pedigreed code [8].

The effect of high discrepancy in uniform CVT sets is immediately apparent in the mapped set in Figure 6. The mapped CVT set has a rectangular shaped layout of points rather than a circularly oriented layout seemingly more appropriate for the circularly symmetric bi-normal JPDF targeted (Figure 4). Unexpectedly, we find in the next section that this non-intuitive rectangular shaped set of points actually performs relatively well among the six types of mapped sets shown in Figure 6. This rectangular-shaped set performs much better, in fact, than the much more likely looking set shown in Figure 7, which was generated directly with density-weighted CVT. The set mapped from uniform Latinized CVT appears much closer to a bi-normal density than the rectangular mapped CVT set, but actually doesn't perform quite as well. The performance of the various mapped sets is examined more closely in the next Section.

3. Evaluation of Statistical Sampling Effectiveness of the Methods

3.1. 2-D Model Response Function and Statistical Measures of Response in Performance Evaluation

Figure 8 shows an analytic multi-modal function describing system response *r* as a function of two system inputs p1 and p2:

$$r(\mathrm{p1},\mathrm{p2}) = \left[0.8\kappa + 0.35\sin\left(2.4\pi\frac{\kappa}{\sqrt{2}}\right)\right][1.5\sin(1.3\theta)] \qquad \text{EQ 2}$$

on the domain $0 \le p1 \le 1$ and $0 \le p2 \le 1$, where $\kappa = \sqrt{(p1)^2 + (p2)^2}$, $\theta = \operatorname{atan}\left(\frac{p2}{p1}\right)$.

A statistical problem arises if p1 and p2 are random variables. In that case, any particular realization $p1_i$ and $p2_i$ of the stochastic variables yields a deterministic response r_i as given by the above functional relationship. An ensemble of responses accompanies the different realizations of p1 and p2 as they vary stochastically or randomly according to their individual propensities, or joint propensities if the two variables are correlated.

The JPDF likelihood function for attaining various input combinations of p1 and p2 maps through the response function *r*(p1,p2) into a corresponding likelihood function for response values. Operationally, the resulting response probability density function, PDF(*r*), can be approached closer and closer via Monte Carlo sampling as more and more parameter sets or real-

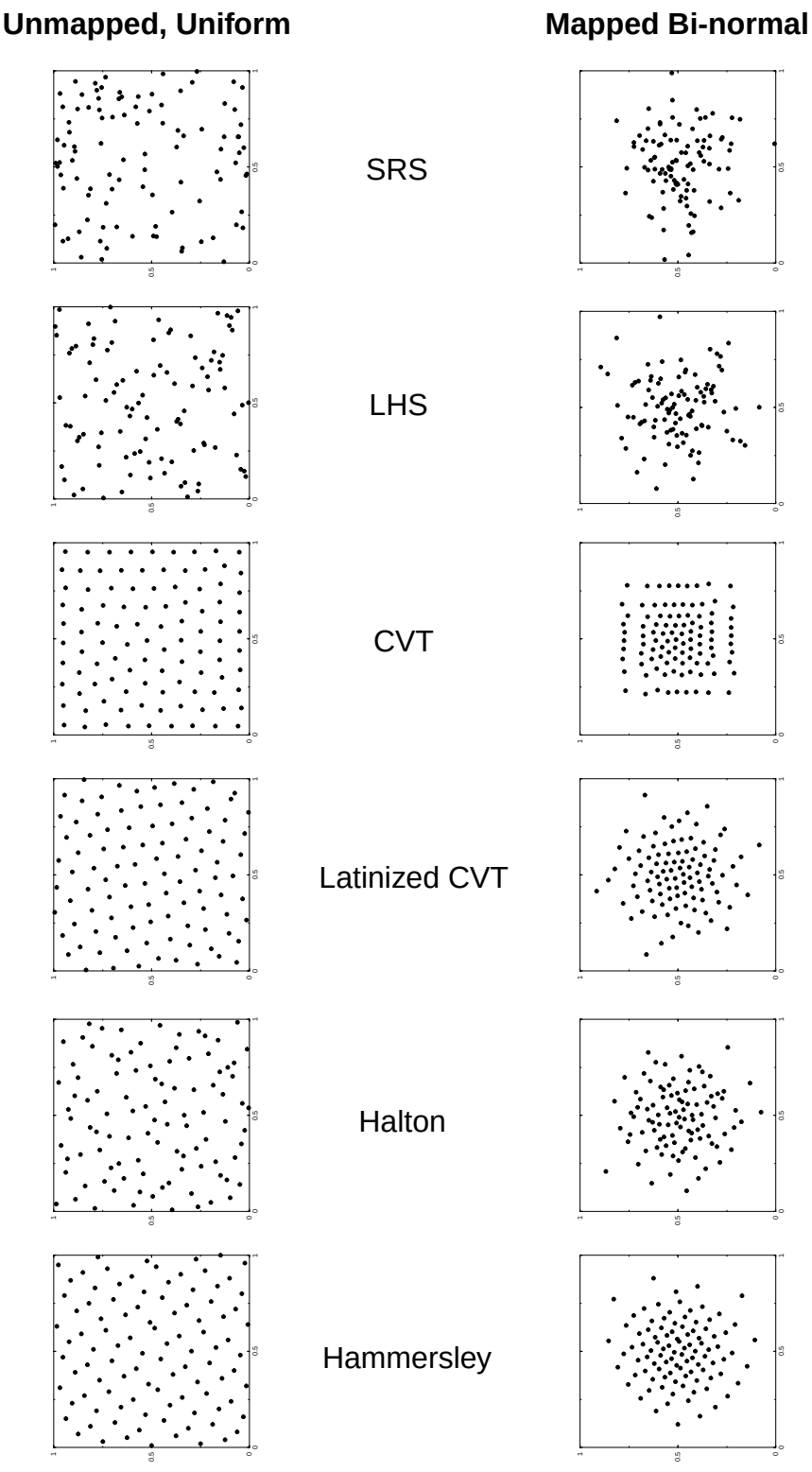


Figure 6. 100-point sample sets on a 2-D unit hypercube for:
Left Column– uniformly distributed point sets
Right Column– corresponding bi-normally distributed point sets mapped fromuniform sets.

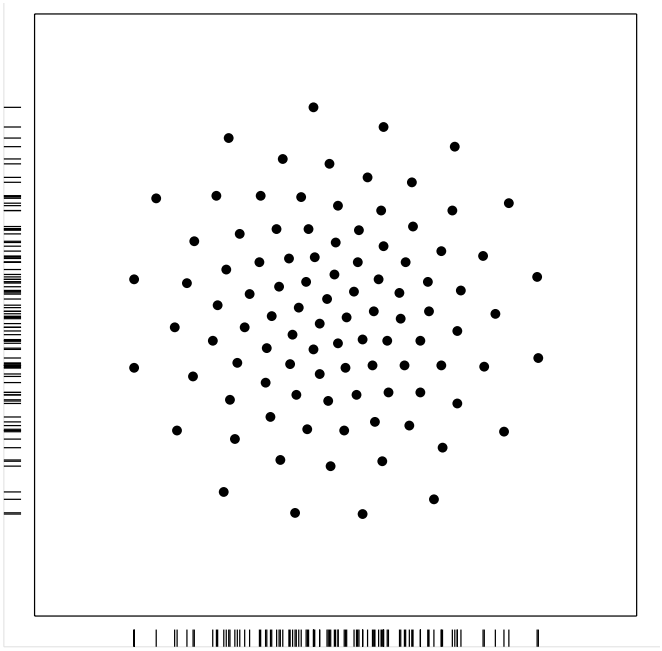


Figure 7. 100-point set in a unit square, generated directly with density-weighted CVT to model the bi-normal joint probability density function shown in Figure 4.

izations $(p1,p2)_i$ are randomly generated from the governing input JPDF and are propagated through the response function $r(p1,p2)$ into response realizations r_i. The response realizations are distributed in the response space (*i.e.*, along the response coordinate axis r) with a density that, as more and more samples are added, trends toward the exact PDF of response.

Very often, only certain statistical measures of the PDF of response are desired or can be reasonably estimated. Response mean, μ_r, and standard deviation, σ_r, can be estimated directly from the mean $\hat{\mu}_r$ and standard deviation $\hat{\sigma}_r$ of the population or set $\{r_i\}$ of realizations. We have the following definitions:

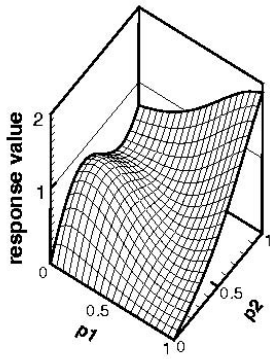


Figure 8. 2-D model function for system response as a function of input parameters p1 and p2.

$$\hat{\mu}_r = \frac{1}{N}\sum_{i=1}^{N} r_i \qquad \text{EQ 3}$$

$$\hat{\sigma}_r = \left[\frac{1}{N-1}\sum_{i=1}^{N} (r_i - \hat{\mu}_r)^2\right]^{\frac{1}{2}} \qquad \text{EQ 4}$$

where N is the number of realizations or samples of response.

Also of interest is the probability of response exceeding (or not exceeding) some particular threshold value r_T. The former is equivalent to the volume integral of the joint probability density function JPDF(p1,p2) integrated over the region of the p1-p2 domain where response exceeds the stipulated threshold r_T. Three such regions corresponding to thresholds r_T=1.0, 0.5, and 0.2 are shown in Figure 9 (as shaded areas on the cutting planes r_T=1.0, 0.5, 0.2) for our model function EQ 2.

Exceedence probability is very simply estimated as the ratio of the number of calculated response values at or above the given threshold value, to the total number of samples N drawn from the JPDF. As the number of response realizations increases, the estimate (quotient) trends toward greater accuracy, *i.e.*, toward the actual exceedence probability. This is of course also true for the estimates $\hat{\mu}_r$ and $\hat{\sigma}_r$ of response mean and standard deviation.

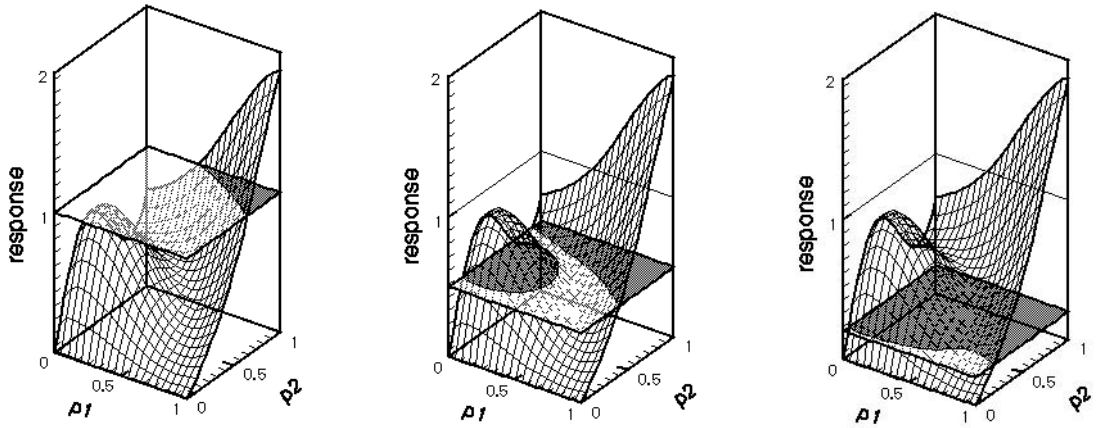


Figure 9. Cutting planes through exact function showing associated exceedence (shaded) and complement (unshaded) regions of the p1-p2 parameter space for response threshold values of 1.0, 0.5, and 0.2, respectively.

3.2. Comparison of Response Statistics from Various Sampling Methods

Here we compare estimates of response mean, standard deviation, and exceedence probabilities as obtained from the various sample sets represented by the right column in Figure 6. We map these $\{(p1,p2)_i\}$ sets through our response function EQ 2 to obtain corresponding response sets, and then calculate the aforementioned statistics of the response populations. We then compare the calculated statistics of each response set to reference values obtained from using three million SRS samples at parameter values generated by the sampling code [9]. The reference val-

ues are actually averages of three results, each obtained from one million samples generated from random initial seeds “X”, “Y”, and “Z” (different from seeds 1, 2, and 3 used to generate the 100-sample sets).

Three “replicate” sets of one million samples each are used in preference to one set of three million samples so that *empirical* confidence intervals (CI) on the calculated averages could be compared against their classical CI to reaffirm or caveat them. (Recent research ([16], [20]) has shown that for SRS, empirical CI appear to be somewhat more accurate than classical CI.) Empirical CI are formed by assuming the calculated statistic (response mean, standard deviation, or exceedence probability) is a random realization from a normal or nearly normal distribution about the exact result. Hence a T-distribution with 3 - 1 = 2 degrees of freedom can be used to get confidence intervals about the small-sample average of the three replicates. Thus, for 95% empirical CI the following formula is used:

$$\text{95\% confidence half-interval} = 4.303\frac{\hat{\sigma}_{est}}{\sqrt{3}} \qquad \text{EQ 5}$$

where $\hat{\sigma}_{est}$ is the sample standard deviation (*cf.* EQ 4) of the three estimates.

Table 1 shows various estimates of response mean, standard deviation, and exceedence probabilities calculated from the three one-million-sample SRS sets. The average and standard deviation of the estimates is also shown in the table.

Table 1. Calculated response statistics for reference values (10^6 samples, Bi-normal JPDF, SRS

		response statistic			
		$\hat{\mu}_r$	$\hat{\sigma}_r$	$\hat{P}_{0.2}$	$\hat{P}_{0.5}$
REALIZATION	1	0.511872	0.162834	0.984429	0.448457
	2	0.511824	0.162733	0.984585	0.447915
	3	0.511940	0.162737	0.984511	0.449029
	average	0.511879	0.162768	0.984508	0.448467
	std. dev.	5.829E-05	5.720E-05	7.803E-05	0.000557

3.2.1. Mean of Response

The average of the three 10^6-sample estimates of mean response is taken as the reference value, $\hat{\mu}_{ref}$ =0.511879 from Table 1. Empirical confidence intervals on this reference mean are obtained by substituting the standard deviation of the estimates, $\hat{\sigma}_{est}$ =5.829E-05 from Table 1, into EQ 5. Thus, empirical 95% half-CI are 0.000145. When the reference mean is calculated based on the entire population of N=3x10^6 samples, the value doesn’t change from the averaged value based on three separate 10^6-sample sets, but the classical CI can be computed. The classical 95% half-CI from standard statistical formulas is somewhat larger, at 0.000185. Using the larger (classical) CI here to be conservative, we say that with at least 95% certainty the true re-

sponse mean μ lies within the range $\hat{\mu}_{ref} \pm 0.000185$ = (0.511694, 0.512064). The CI range ±0.000185 is typically very small compared to the *nominal* differences listed in Tables 2 and 3 between $\hat{\mu}_{ref}$ and the estimates of mean response from the 100-sample sets.

We take the differences from $\hat{\mu}_{ref}$ in Tables 2 and 3 as nominal measures of the error of the estimates from the 100-sample sets. For SRS, LHS, CVT, and LCVT methods there is no unique 100-sample set. For SRS and LHS the sets depend on the initial seed and the particular pairing of the 0 to 1 random variates on the p1 and p2 axes (of a uniform JPDF set). The CVT, and LCVT sets further depend –fairly insensitively if enough iterations are performed to stabilize certain uniformity measures, see [2]– on the starting sample set (initial condition). We therefore use three instantiations of SRS, LHS, CVT, and LCVT sets to begin to obtain a representative picture of the errors we might expect from a random realization of each of these types of sets. For each of these methods we average the individual errors from the three instantiations to determine an average magnitude of error. This measure reflects contributions from both the average error (bias) in the three estimates, as well as the variance of the three results. (This error measure is zero only if both the average error (bias) is zero and the variance of the estimates is zero.) Furthermore, this error measure applies as well to the Halton and Hammersley results which consist of only one instantiation because they are deterministic sampling methods.

To also obtain a broad picture of the each method's sampling efficacy across the different types of statistics calculated, we just use a simple ranking scheme for method accuracy for each of the various calculated statistics (response mean, variance, and exceedence probabilities). This allows us to compare method performance across the different types of statistics calculated. This is perhaps somewhat more satisfying than the piecemeal comparisons in, e.g., [15] and [19] that fail to give an explicit impression (quantitative balanced indicator) of the overall performance of the various sampling methods across a matrix of test problems. Hence, the accuracy ranking of each method with respect to average magnitude of error is given on the final lines in Tables 2 and 3. Rank 1 indicates the method was the most accurate and therefore ranked first in performance. Rank 6 indicates the method was the least accurate among the sampling schemes tried.

Table 2. Calculated response means (100 samples, Normal 2D JPDF)

		SRS		LHS		Latinized CVT	
		$\hat{\mu}_r$	$\hat{\mu}_r$ error	$\hat{\mu}_r$	$\hat{\mu}_r$ error	$\hat{\mu}_r$	$\hat{\mu}_r$ error
REALIZATION	1	0.49980	-0.01208	0.51466	+0.00278	0.51238	+0.00050
	2	0.51058	-0.00130	0.51110	-0.00079	0.50780	-0.00408
	3	0.50315	-0.00873	0.50752	-0.00436	0.51182	-5.867E-05
	average	0.50451	-0.00737	0.51109	-0.00079	0.51067	-0.00121
	std. dev.	0.00552	0.00552	0.00357	0.00357	0.00250	0.00250
	avg. error mag.		0.00737 Rank 6		0.00264 Rank 4		0.00155 Rank 1

Table 3. Calculated response means (100 samples, Normal 2D JPDF)

		CVT		Halton		Hammersley	
		$\hat{\mu}_r$	$\hat{\mu}_r$ error	$\hat{\mu}_r$	$\hat{\mu}_r$ error	$\hat{\mu}_r$	$\hat{\mu}_r$ error
REALIZATION	1	0.50868	-0.00320	0.50565	-0.00623	0.51029	-0.00159
	2	0.51148	-0.00040				
	3	0.50840	-0.00348				
	average	0.50952	-0.00236	0.50565	-0.00623	0.51029	-0.00159
	std. dev.	0.00170	0.00170				
avg. error mag.			0.00236 Rank 3		0.00623 Rank 5		0.00159 Rank 2

3.2.2. Standard Deviation of Response

Tables 4 and 5 show the estimates of the standard deviation of response. Nominal errors from the reference value $\hat{\sigma}_{ref}$ =0.162768 are also shown. This value is the average of the three standard deviations in Table 1 calculated from the three 10^6 SRS sets. The standard deviation of these three estimates is $\hat{\sigma}_{est}$ =5.719E-05. Empirical 95% half-CI by EQ 5 are 0.000142. Accordingly, we say that with 95% confidence the true response standard deviation σ lies within the range $\hat{\sigma}_{ref}$ ±0.000142 = (0.162626, 0.162910). The CI are negligibly small.

Table 4. Calculated response standard deviations (100 samples, Normal 2D JPDF)

		SRS		LHS		Latinized CVT	
		$\hat{\sigma}_r$	$\hat{\sigma}_r$ error	$\hat{\sigma}_r$	$\hat{\sigma}_r$ error	$\hat{\sigma}_r$	$\hat{\sigma}_r$ error
REALIZATION	1	0.16874	+0.00597	0.18054	+0.01777	0.15782	-0.00495
	2	0.16265	-0.00012	0.15570	-0.00707	0.15532	-0.00745
	3	0.15699	-0.00578	0.14191	-0.02086	0.15554	-0.00723
	average	0.16279	2.533E-05	0.15938	-0.00339	0.15623	-0.00654
	std. dev.	0.00588	0.00588	0.01958	0.01958	0.00138	0.00138
avg. error mag.			0.00396 Rank 2		0.01523 Rank 6		0.00654 Rank 3

Table 5. Calculated response standard deviations (100 samples, Normal 2D JPDF)

		CVT		Halton		Hammersley	
		$\hat{\sigma}_r$	$\hat{\sigma}_r$ error	$\hat{\sigma}_r$	$\hat{\sigma}_r$ error	$\hat{\sigma}_r$	$\hat{\sigma}_r$ error
REALIZATION	1	0.14804	-0.01473	0.15415	-0.00862	0.16077	-0.00200
	2	0.14906	-0.01371				
	3	0.14615	-0.01662				
	average	0.14775	-0.01502	0.15415	-0.00862	0.16077	-0.00200
	std. dev.	0.00148	0.00148				
	avg. error mag.		0.01502 Rank 5		0.00862 Rank 4		0.00200 Rank 1

3.2.3. Response Exceedence Probability for r_T=0.2

Tables 6 and 7 show the estimates of the exceedence probability (EP) corresponding to a response threshold level of r_T=0.2. Nominal errors from the reference value $\hat{P}_{0.2,\,ref}$ =0.984508 are also shown. This value is the average of the three EPs in Table 1 calculated from the three 10^6 SRS sets. The standard deviation of these three estimates is $\hat{\sigma}_{est}$ =7.803E-05. Empirical 95% half-CI by EQ 5 are 0.000194. When the reference EP is calculated based on the entire population of N=3x10^6 samples, the value doesn't change from the averaged value based on three separate 10^6-sample sets, but classical CI can be computed. The classical 95% half-CI from standard statistical formulas is somewhat smaller, at 0.000140. Using the larger (empirical) 95% half-CI for conservatism, we say that to 95% confidence the true probability $P_{0.2}$ of response exceeding the threshold value r_T=0.2 lies within the range $\hat{P}_{0.2,\,ref}$ ±0.000194 = (0.984314, 0.984702). The CI are negligibly small. We note that the SRS and LHS results are both ranked at 4.5 because together they occupy the 4th and 5th ranks and both have the same error magnitude.

Table 6. Calculated response exceedence probabilities, threshold=0.2 (100 samples, Normal 2D JPDF)

		SRS		LHS		Latinized CVT	
		$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error
REALIZATION	1	0.97	-0.01451	0.98	-0.00451	0.99	+0.00549
	2	0.99	+0.00549	0.99	+0.00549	0.98	-0.00451
	3	0.99	+0.00549	1.00	+0.01549	0.98	-0.00451
average		0.98333	-0.00117	0.99	+0.00549	0.98333	-0.00118
std. dev.		0.01155	0.01155	0.01	0.01	0.00577	0.00577
avg. error mag.			0.00850 Rank 4.5		0.00850 Rank 4.5		0.00484 Rank 2

Table 7. Calculated response exceedence probabilities, threshold=0.2 (100 samples, Normal 2D JPDF)

		CVT		Halton		Hammersley	
		$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error	$\hat{P}_{0.2}$	$\hat{P}_{0.2}$ error
REALIZATION	1	0.99	+0.00549	0.97	-0.01451	0.98	-0.00451
	2	0.99	+0.00549				
	3	0.99	+0.00549				
average		0.99	+0.00549	0.97	-0.01451	0.98	-0.00451
std. dev.		0.0	0.0				
avg. error mag.			0.00549 Rank 3		0.01451 Rank 6		0.00451 Rank 1

3.2.4. Response Exceedence Probability for r_T=0.5

Tables 8 and 9 show the estimates of the exceedence probability (EP) corresponding to a response threshold of r_T=0.5. Nominal errors from the reference value $\hat{P}_{0.5,\,ref}$ =0.448467 are also shown. This value is the average of the three EPs in Table 1 calculated from the three 10^6 SRS sets. The standard deviation of these three estimates is $\hat{\sigma}_{est}$ =0.000557. Empirical 95% half-CI by EQ 5 are 0.001384. When the reference EP is calculated based on the entire population of N=3x10^6 samples, the value doesn't change from the averaged value based on three separate 10^6-sample sets, but classical CI can be computed. The classical 95% half-CI from standard sta-

tistical formulas is considerably smaller, at 0.000563. Using the larger (empirical) 95% half-CI for conservatism, we say that to 95% confidence the true probability $P_{0.5}$ of response exceeding the threshold value r_T=0.5 lies within the range $\hat{P}_{0.5,\,ref}$ ±0.001384 = (0.447083, 0.449851). The CI are negligibly small.

Table 8. Calculated response exceedence probabilities, threshold=0.5 (100 samples, Normal 2D JPDF)

		SRS		LHS		Latinized CVT	
		$\hat{P}_{0.5}$	$\hat{P}_{0.5}$ error	$\hat{P}_{0.5}$	$\hat{P}_{0.5}$ error	$\hat{P}_{0.5}$	$\hat{P}_{0.5}$ error
REALIZATION	1	0.43	-0.01847	0.44	-0.00847	0.46	+0.01153
	2	0.41	-0.03847	0.45	+0.00153	0.43	-0.01847
	3	0.43	-0.01847	0.45	-0.00847	0.47	+0.02153
	average	0.42333	-0.02513	0.44667	-0.0018	0.45333	+0.00487
	std. dev.	0.01155	0.01155	0.00577	0.00577	0.02082	0.02082
avg. error mag.			0.02513 Rank 6		0.00384 Rank 1		0.01718 Rank 5

Table 9. Calculated response exceedence probabilities, threshold=0.5 (100 samples, Normal 2D JPDF)

		CVT		Halton		Hammersley	
		$\hat{P}_{0.5}$	$\hat{P}_{0.5}$ error	$\hat{P}_{0.5}$	$\hat{P}_{0.5}$ error	$\hat{P}_{0.5}$	$\hat{P}_{0.5}$ error
REALIZATION	1	0.44	-0.00847	0.46	+0.01153	0.44	-0.00847
	2	0.46	+0.01153				
	3	0.46	+0.01153				
	average	0.45333	+0.00487	0.46	+0.01153	0.44	-0.00847
	std. dev.	0.01155	0.01155				
avg. error mag.			0.01051 Rank 3		0.01153 Rank 4		0.00847 Rank 2

3.3. Weighted Measure of Statistical Sampling Merit

The performance rankings for the sampling schemes and statistical quantities tested are summarized in Table 10. The last column contains a normalized weighted figure of merit which is a broad measure of each method's sampling performance across the different types of statistics calculated. This figure of merit is obtained by first averaging the rankings for the (two) exceedence probabilities calculated, and then averaging this rank for EPs in with the ranks for the mean and standard deviation calculations. These averages are then divided by the number of sampling methods involved. Hence, the normalized ranks in this column add up to unity. This type of normalization allows comparison to other investigations such as those in, e.g., [11], [15] and [19], if their results are also normalized in this manner. The bar chart in Figure 10 helps visually assess the relative performance of the sampling methods according to our normalized figure of merit. The shorter the bar, the better the particular method ranks on balance across all the statistical quantities calculated. We see that Hammersley sampling ranked overall best on this series of test problems, then LCVT, CVT, LHS, SRS, and finally Halton.

Table 10. Sampling Method accuracy rankings for various calculated statistics of response

Sampling Method	response mean	response standard deviation	exceedence probability (0.2 threshold)	exceedence probability (0.5 threshold)	normalized weighted average
SRS	6	2	4.5	6	0.21
LHS	4	6	4.5	1	0.20
CVT	3	5	3	3	0.17
LCVT	1	3	2	5	0.12
Halton	5	4	6	4	0.22
Hammersley	2	1	1	2	0.07

A second bar chart corresponding to the investigation in [19] is plotted in Figure 10. The investigation was similar to the one in this paper, but compared calculated statistics based on a uniform JPDF, and did not include the LCVT sampling method. Since there is no mapping here from uniform sets to nonuniform JPDFs, only volumetric uniformity matters here and discrepancy properties are immaterial. Since pure CVT is more volumetrically uniform than LCVT, and for that matter, more volumetrically uniform than all the other sampling methods we've tested, CVT would be expected to generally rank best. This is the case shown in Figure 10 for the set of test problems investigated in [19].

4. Discussion and Conclusion

According to our weighted figure of merit, Hammersley sampling strongly ranked overall best on the set of bi-normal JPDF test problems in this paper, then LCVT, CVT, LHS, SRS, and

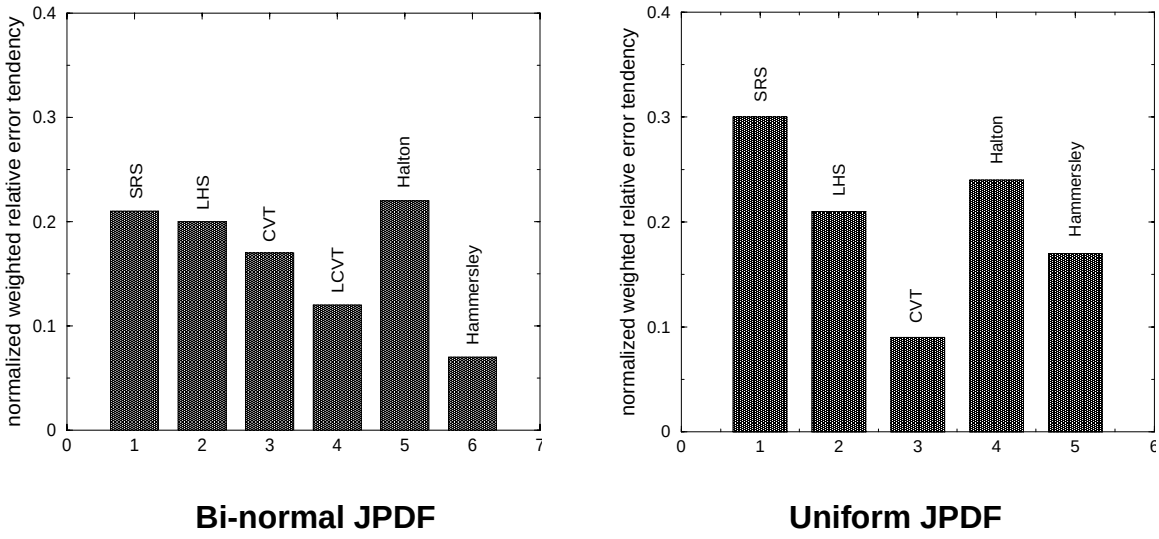


Figure 10. Normalized weighted measure of sampling method relative error tendency in calculated mean, standard deviation, and r_T = 0.2 and 0.5 exceedence probabilities for uniform and bi-normal joint probability densities in 2D test problems.

finally Halton. Furthermore, in [11], Hammersley was found to be significantly more efficient than SRS and LHS for resolving mean and standard deviation of response over a large set of test problems. For resolving response probabilities, Hammersley and modified-Halton were found in [15] to perform roughly the same as LHS on balance over several test problems.

Hence, Hammersley is consistently the best performer or among the top performers in these empirical studies. Hammersley is also the only one of the top contenders in these studies that allows incremental addition of samples to the parameter space (as little as one at a time), which enables it to have better efficiency prospects in both error estimation and control. Given these apparent advantages, it seems that Hammersley sampling might be a superior choice in many circumstances. However, when the number of random inputs grows beyond 10 or so dimensions and/or the sampling density in the hypercube becomes high, Hammersley might suffer from the spurious correlation effects that plague other sub-random sequence methods. This is shown, e.g., for standard Halton sequences in 16-D (ref. [12]) and 40-D (ref. [15]). This is something the authors need to further inquire about; the answer may already exist in the literature.

Furthermore, we cannot yet dismiss the competitive potential of CVT or LCVT based on the single limited investigation conducted in this paper. In particular, more than three instantiations of SRS, LHS, CVT, and LCVT point sets are needed to more reliably reflect the true performance tendencies of these methods on our test problems. Also, sample sets of much larger size than 100 would be valuable particularly to get another significant digit of resolution in the calculated exceedence probabilities in the study. Moreover, our results are somewhat tied to the specific figure of merit employed in this study. This figure of merit has the advantage that it allows comparison of merit across different types of statistics calculated and different problem

sets, but other better measures may exist for our purposes. Certainly, our weighted metric does not reveal method performance in the individual categories of response mean, standard deviation, and exceedence probability (but these can be found in Table 10 for the problems in this paper).

Finally, empirical studies are only point glimpses of the relative accuracy tendencies of one method over another under a very specific set of conditions. Certainly, much more empirical work needs to be performed to assess the performance of CVT and LCVT versus other sampling methods over a diverse problem space, but even more valuable would be more theoretical work to ascertain which method might be expected to perform best under given conditions (the characteristics of the function involved; the number of input random variables/dimensions; character of the JPDF, etc.).

This being said, we have early empirical indications of the promise of CVT in uniform JPDF problems. In [19], CVT strongly ranked overall best as expected, then Hammersley, LHS, Halton, and finally SRS. In particular, for statistical integration of functions, which involves uniform sampling over the integration domain, CVT appears to be the natural best choice theoretically, as corroborated by findings in [19]. Also, in point placement for response-surfaces, CVT appears very promising relative to other structured and unstructured sampling methods (see [18]). Already, for irregular (non-hypercube) interpolation and integration domains, the uniformity of CVT sampling over the domain gives it a well recognized status in the application of 2-D and 3-D meshless finite-element methods.

Hence, when volumetric sampling uniformity is desirable, early indications are that CVT performs very well versus other sampling methods. However, to reiterate, much more empirical and theoretical work remains to be done to broadly assess and characterize the potential of CVT and LCVT for various sampling tasks.

REFERENCES

[1] Beachkofski, B.K., and R.V. Grandhi, "Improved Distributed Hypercube Sampling," paper AIAA-2002-1274, 43rd Structures, Structural Dynamics, and Materials Conference, April 22-25, 2002, Denver, CO.

[2] Burkardt, J., M. Gunzburger, H. Nguyen, J. Peterson, and Y. Saka, "A new methodology for point sampling I: Uniform sampling in hypercubes" in preparation

[3] Burkardt, J., M. Gunzburger, H. Nguyen, J. Peterson, and Y. Saka, "A new methodology for point sampling II: Nonuniform, anisotropic sampling in general regions" in preparation

[4] Burkardt, J.V., M.D. Gunzburger, J.S. Peterson, research sampling software, http://www.csit.fsu.edu/~burkardt/f_src/cvt_dataset/cvt_dataset.html

[5] Conover, W.M., "On a Better Method for Selecting Input Variables," 1975 unpublished Los Alamos National Laboratories manuscript, reproduced as Appendix A of "Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems" by J.C. Helton and F.J. Davis, Sandia National Laboratories report SAND2001-0417, printed November 2002.

[6] Du, Q., V. Faber, M. Gunzburger, "Centroidal Voronoi tessellations: applications and algorithms," *SIAM Review.* Vol. 41, 1999, pp. 637-676.

[7] Hanson, K.M., “Quasi-Monte Carlo: halftoning in high dimensions?,” to appear in Proceedings of SPIE 5016 (2003), *Computational Imaging*, C.A. Bouman and R.L. Stevenson, eds.

[8] Iman, R.L., and Conover, W.J., “A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables,” *Communications in Statistics*, B11(3) 1982a, pp. 311-334.

[9] Iman, R.L., and Shortencarier, M.J., “A FORTRAN77 Program and User’s Guide for the Generation of Latin Hypercube and Random Samples to Use with Computer Models,” Sandia National Laboratories report SAND83-2365 (RG), printed March 1984.

[10] Ju, L., Q. Du., M. Gunzburger, “Probabilistic algorithms for centroidal Voronoi tessellations and their parallel implementation,” *Parallel Computing*, Vol. 28, 2002, pp. 1477-1500.

[11] Kalagnanam, J.R., and Diwekar, U.M., "An Efficient Sampling Technique for Off-line Quality Control," *Technometrics*, August 1997, Vol.39, No.3.

[12] Manteufel, R.D., “Distributed Hypercube Sampling Algorithm,” paper AIAA-2001-1673, 42nd Structures, Structural Dynamics, and Materials Conference, April 16-19, 2001, Seattle, WA.

[13] Okabe, A., B. Boots, K. Sugihara, S. Chui, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd Edition, Wiley, Chichester, 2000.

[14] Press, W.H., S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd Edition, Cambridge University Press, 1992.

[15] Robinson, D., and Atcitty, D., “Comparison of Quasi- and Pseudo- Monte Carlo Sampling for Reliability and Uncertainty Analysis,” paper AIAA-1999-1589, 40th Structures, Structural Dynamics, and Materials Conference, April 12-15, St. Louis, MO.

[16] Romero, V.J., “Effect of Initial Seed and Number of Samples on Simple-Random and Latin-Hypercube Monte Carlo Probabilities –Confidence Interval Considerations,” paper PMC2000-176, 8th ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability, Univ. of Notre Dame, IN, July 24-26, 2000.

[17] Romero, V.J., and S.D. Bankston, “Progressive-Lattice-Sampling Methodology and Application to Benchmark Probability Quantification Problems,” Sandia National Laboratories report SAND98-0567, printed March 1998.

[18] Romero, V.J., J.S. Burkardt, M.D. Gunzburger, J.S. Peterson, T. Krishnamurthy, “Initial Application and Evaluation of a Promising New Sampling Method for Response Surface Generation: Centroidal Voronoi Tessellation,” paper AIAA-2003-2008, 44th Structures, Structural Dynamics, and Materials Conference (5th AIAA Non-Deterministic Approaches Forum), Apr. 7-10, 2003, Norfolk, VA.

[19] Romero, V.J., J.S. Burkardt, M.D. Gunzburger, J.S. Peterson, “Initial Evaluation of Centroidal Voronoi Tessellation Method for Statistical Sampling and Function Integration,” Sandia National Laboratories library archive SAND2003-3672C, an extended version (with comparison of exceedence probabilities) of the paper in the proceedings of the 4th International Symposium on Uncertainty Modeling and Analysis (ISUMA’03), Sept. 21-24, 2003, University of Maryland, College Park, MA.

[20] Romero, V.J., and A.L. Mellen, 2002, preliminary research findings not yet published.

[21] Ye, K.Q., W. Li, A. Sudjianto, “Algorithmic Construction of Optimal Symmetric Latin Hypercube Designs,” *Journal of Statistical Planning and Inference*, Vol. 90 (2000), pp. 145-159.

MCMC-based Sensitivity Estimations

Pérez C, Martín J, and Rufo M. J.

Department of Mathematics, University of Extremadura, Avda. Universidad s/n, 10071 Cáceres, Spain
E-mail: carper@unex.es, jrmartin@unex.es, mrufo@unex.es

Abstract: The problem of calculating local parametric sensitivities is addressed. We propose a computationally low-cost method to estimate local sensitivities in Bayesian models. The proposed general method introduces a great flexibility because it can be applied to complex models that need to be solved by MCMC methods, and it allows to estimate the sensitivity measures and their errors with no additional random sampling. This sensitivity analysis method is easy to apply in practice as we show with an illustrative example.

Keywords: Sensitivity, Simulation, MCMC, Bayesian Inference, Bayesian Decision Theory.

1. INTRODUCTION

Many problems in statistics and operational research involve making decisions under uncertainty. Bayesian statistical methods provide a complete paradigm for both statistical inference and decision making under uncertainty. This methodology allows to combine information derived from observations with information elicited from experts. The range of its potential applicability is very wide. It is particularly useful for highly reliable components and systems where failures in test and field operations are very rare, requiring the use of all other engineering information. This methodology has become more popular due to the appearance of Markov Chain Monte Carlo (MCMC) methods (see Brooks [1] for a review). The application of these simulation techniques allows to obtain a numerical solution of problems based on really complex models. Sometimes, MCMC methods are the only computationally efficient alternative.

In addition to the solution, we need some description of its sensitivity with respect to reasonable changes and uncertainties in the specification of the inputs. Sensitivity analysis seeks to find out how the output of a model changes with variations in the inputs (see Saltelli et al. [2]). Such knowledge is important for (a) evaluating the applicability of the model, (b) determining parameters for which it is important to have more accurate values, and (c) understanding the behavior of the system being modeled. The output needs to be interpreted carefully whenever it changes significantly for input variations that are within the bounds of possible error. There are two kinds of sensitivity analysis: local and global. Local sensitivity studies parameter variations over neighborhoods around what are believed to be appropriate values, while global sensitivity considers parameter changes over the whole domain.

Further author information: (Corresponding author: C. Pérez)

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

Sensitivity analysis is required in many applications, for example, in those arising in engineering, medicine, archeology, or environment. It is particularly useful in reliability of hardware systems, space systems probabilistic risk analysis, nuclear power risk analysis or information security risk analysis. Sensitivity studies are demanded by several authors to be applied in models solved by MCMC methods (see, for example, Ríos and Ruggeri [3]). Some authors, like Hall et al. [4] and Halekoh and Vach [5], study parametric sensitivity by solving the model for some values of the parameters. The main disadvantage of this procedure is that they have to re-run the Markov chain, i.e, they have to generate new samples for the different parameter values. Therefore, it would be convenient to develop a general sensitivity method that can be applied to estimate local parametric sensitivities in Bayesian models solved by MCMC techniques. We address that issue in this paper.

The outline is as follows. In Section 2, a computationally low-cost method to estimate local parametric sensitivities is proposed. In order to show how the proposed method is easily applied in practice, an illustrative example is presented in Section 3. Finally, the conclusion is presented in Section 4.

2. LOCAL PARAMETRIC SENSITIVITY ESTIMATIONS

Suppose we are interested in the estimation of a quantity $\mathcal{I}$ that can be expressed as an integral of a function f over a multiple dimension domain with respect to a density g, i.e:

$$\mathcal{I} = \int_{\Theta} f(\theta)\, g(\theta)\, d\theta. \tag{1}$$

When g is the posterior distribution for θ, i.e, $g(\theta|x)$, this quantity could be, for example, the posterior mean. Note that $g(\theta)$ $(f(\theta))$ could depend on parameters, so a more convenient notation is $g_\lambda(\theta)$ $(f_\lambda(\theta))$ where λ represents a possibly multidimensional parameter in the space Λ. Firstly, we study the problem considering imprecision in g_λ, later we present a similar study for f_λ. In the former case, expression (1) becomes:

$$\mathcal{I}_\lambda = \int_{\Theta} f(\theta)\, g_\lambda(\theta)\, d\theta, \tag{2}$$

where Θ is independent of λ.

Suppose that sampling directly from $g_\lambda(\theta)$ is so complex that we need to use MCMC methods. Note that this is the case for most of the real problems. Let $\theta_1, \theta_2, \ldots, \theta_n$ be a sample generated from $g_{\lambda^0}(\theta)$ by MCMC methods, where λ^0 is a fixed quantity interior to Λ. Then, an estimate of $\mathcal{I}_{\lambda^0}$ is given by:

$$\widehat{\mathcal{I}_{\lambda^0}} = \frac{1}{n} \sum_{i=1}^{n} f(\theta_i). \tag{3}$$

Now, our interest is focused on evaluating the impact of changes on $\mathcal{I}_\lambda$ when λ varies in an infinitesimal neighborhood of λ^0, i.e, we want to make a local sensitivity analysis. The choice of a sensitivity analysis method depends on a great extent on (a) the sensitivity

measures employed, (b) the accuracy in the estimates of the sensitivity measures, and (c) the computational cost involved. All these topics are studied in this section.

The first step is to define a local sensitivity measure. This measure must be easily interpretable and efficiently computed. Sometimes sensitivity is characterized through gradients or partial derivatives at the target point (see Turányi and Rabitz [6] and references therein). Suppose that all the partial derivatives exist. As a local sensitivity measure, we consider the gradient vector evaluated at λ^0 , i.e:

$$\nabla \mathcal{I}_{\lambda^0} = (\partial_{\lambda_1} \mathcal{I}_{\lambda^0}, \partial_{\lambda_2} \mathcal{I}_{\lambda^0}, \ldots, \partial_{\lambda_m} \mathcal{I}_{\lambda^0}) \,. \tag{4}$$

Components in (4), i.e. the partial derivatives with respect to each λ_j evaluated at λ^0, indicate how rapidly $\mathcal{I}_\lambda$ is changing around an infinitesimal neighborhood of λ^0 along that axis. Therefore, they can be used as rates of change with respect to the parameter components. Then $\nabla \mathcal{I}_{\lambda^0}$ can be considered as a local sensitivity measure for the parameter λ at λ^0. The gradient vector represents the precise direction which has maximum increase of $\mathcal{I}_\lambda$ at λ^0. Furthermore, it indicates which component has the largest influence on the output.

In this context, the main problem is to calculate the gradient vector. We present a computationally low-cost method to estimate the components of (4). Under mild conditions, each component of $\nabla \mathcal{I}_{\lambda^0}$ can be expressed as:

$$\begin{aligned} \partial_{\lambda_j} \mathcal{I}_{\lambda^0} &= \int_\Theta \partial_{\lambda_j} \left(f(\theta)\, g_{\lambda^0}(\theta) \right) d\theta = \int_\Theta f(\theta) \partial_{\lambda_j} g_{\lambda^0}(\theta) d\theta = \\ &= \int_\Theta \frac{f(\theta) \partial_{\lambda_j} g_{\lambda^0}(\theta)}{g_{\lambda^0}(\theta)} g_{\lambda^0}(\theta)\, d\theta = E_{g_{\lambda^0}} \left(\frac{f(\theta)\, \partial_{\lambda_j} g_{\lambda^0}(\theta)}{g_{\lambda^0}(\theta)} \right) \end{aligned} \tag{5}$$

and estimated by:

$$\widehat{\partial_{\lambda_j} \mathcal{I}_{\lambda^0}} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\theta_i)\, \partial_{\lambda_j} g_{\lambda^0}(\theta_i)}{g_{\lambda^0}(\theta_i)}. \tag{6}$$

Also, we can estimate the error committed when estimating (5) by using (6). For each j, the estimate given in (6) is unbiased, so its error can be measured by its standard error (see e.g. Tanner [7]). The estimation of the error can be easily obtained from the generated sample that has been used to estimate $\mathcal{I}_{\lambda^0}$ and $\partial_{\lambda_j} \mathcal{I}_{\lambda^0}$.

The advantages of this local sensitivity analysis procedure are mainly two. First, it can be applied to complex models that need MCMC methods to sample from the objective densities. Second, the computations generally represent a very low additional cost because no further sampling is required. The same MCMC outputs obtained to estimate $\mathcal{I}_{\lambda^0}$ are used to estimate its sensitivity and the errors in the estimations. However, this approach can only be applied when we know a closed expression for g_{λ^0} and we can calculate its partial derivatives, what is not always possible. In fact, for complex models the explicit form for g_{λ^0} is usually analytically intractable. Nevertheless, we can obtain some results studying the practical implementation when g_{λ^0} is the posterior distribution. The following two cases are considered.

1. Prior sensitivity. Suppose that the prior distribution $\pi_\lambda(\theta)$ depends on a parameter λ, and let λ^0 be interior to Λ, then:

$$\mathcal{I}_{\lambda^0} = \int_\Theta f(\theta)\, p_{\lambda^0}(\theta|x)\, d\theta = \frac{\int_\Theta f(\theta)\, l(x|\theta)\, \pi_{\lambda^0}(\theta)\, d\theta}{\int_\Theta l(x|\theta)\, \pi_{\lambda^0}(\theta)\, d\theta}. \tag{7}$$

Under mild conditions that allow a derivative-integral interchange (see Spall [8]), we find that each component of $\nabla\mathcal{I}_{\lambda^0}$ can be expressed as:

$$\partial_{\lambda_j}\mathcal{I}_{\lambda^0} = \int_\Theta (f(\theta) - \mathcal{I}_{\lambda^0})\, \frac{\partial_{\lambda_j}\pi_{\lambda^0}(\theta)}{\pi_{\lambda^0}(\theta)}\, p_{\lambda^0}(\theta|x)\, d\theta.$$

The proof is mainly based on the derivative-integral interchange. The posterior steps are basic manipulations addressed to get the integral of a function with respect to the posterior distribution.

If $\theta_1, \theta_2, \ldots, \theta_n$ is generated from the posterior distribution $p_{\lambda^0}(\theta|x)$ (mainly by MCMC methods), then the estimate of $\partial_{\lambda_j}\mathcal{I}_{\lambda^0}$ is given by:

$$\widehat{\partial_{\lambda_j}\mathcal{I}_{\lambda^0}} = \frac{1}{n}\sum_{i=1}^{n} \left(f(\theta_i) - \widehat{\mathcal{I}_{\lambda^0}}\right) \frac{\partial_{\lambda_j}\pi_{\lambda^0}(\theta_i)}{\pi_{\lambda^0}(\theta_i)}, \tag{8}$$

where $\widehat{\mathcal{I}_{\lambda^0}} = \frac{1}{n}\sum_{i=1}^{n} f(\theta_i)$ is the estimate of $\mathcal{I}_{\lambda^0}$. The Monte Carlo standard error estimate of (8) is given by:

$$\widehat{SE}(\widehat{\partial_{\lambda_j}\mathcal{I}_{\lambda^0}}) = \sqrt{\frac{1}{n\,(n-1)}\sum_{i=1}^{n}\left(\frac{(f(\theta_i) - \widehat{\mathcal{I}_{\lambda^0}})\partial_{\lambda_j}\pi_{\lambda^0}(\theta_i)}{\pi_{\lambda^0}(\theta_i)} - \widehat{\partial_{\lambda_j}\mathcal{I}_{\lambda^0}}\right)^2} \tag{9}$$

Note that this case is more tractable because we need the partial derivatives for the prior distribution instead of the partial derivatives for the posterior distribution.

2. Function f sensitivity. If we consider that f belongs to a parametric class of functions, $\mathcal{F}_\lambda = \{f_\lambda, \lambda \in \Lambda\}$, then:

$$\mathcal{I}_{\lambda^0} = \int_\Theta f_{\lambda^0}(\theta)\, p(\theta|x)\, d\theta = \frac{\int_\Theta f_{\lambda^0}(\theta)\, l(x|\theta)\, \pi(\theta)\, d\theta}{\int_\Theta l(x|\theta)\, \pi(\theta)\, d\theta}.$$

Under the mild conditions analogous to the previous case, for each j we have:

$$\partial_{\lambda_j}\mathcal{I}_{\lambda^0} = \int_\Theta \partial_{\lambda_j} f_{\lambda^0}(\theta)\, p(\theta|x)\, d\theta,$$

and its estimate is given by:

$$\widehat{\partial_{\lambda_j}\mathcal{I}_{\lambda^0}} = \frac{1}{n}\sum_{i=1}^{n} \partial_{\lambda_j} f_{\lambda^0}(\theta_i), \tag{10}$$

where $\theta_1, \theta_2, \ldots, \theta_n \sim p(\theta|x)$. Now, the Monte Carlo standard error estimate of (10) is given by:

$$\widehat{SE}(\widehat{\partial_{\lambda_j}\mathcal{I}_{\lambda^0}}) = \sqrt{\frac{1}{n\,(n-1)}\sum_{i=1}^{n}\left(\partial_{\lambda_j} f_{\lambda^0}(\theta_i) - \widehat{\partial_{\lambda_j}\mathcal{I}_{\lambda^0}}\right)^2} \tag{11}$$

Note that if in any problem, the functions g and f depend on the same parameter λ, then the sensitivity measure proposed in this section can be estimated in the same sense. The estimate of each component can be expressed as the sum of analogous quantities to (8) and (10).

As a particular case, we can study the practical implementation of the proposed sensitivity measure in the context of Bayesian decision theory (among the many fine reviews are, for example, Berger [9] and French and Ríos [10]). Bayesian decision theory and inference describe a decision problem by a set of possible actions $a \in \Delta$, a set of states, or parameters, $\theta \in \Theta$, a prior distribution $\pi(\theta)$, a likelihood, $l(x|\theta)$ for the observed data x, and a loss (utility) function $l(a,\theta)$ ($u(a,\theta)$). The actions are ranked by the expected loss (utility). The optimal decision a^* is the action that minimizes (maximizes) the posterior expected loss (utility):

$$\begin{aligned} a^* &= \arg\min_{a\in\Delta} L(a), \\ L(a) &= \int l(a,\theta)\, p(\theta|x)\, d\theta = \frac{\int l(a,\theta)\, l(x|\theta)\, \pi(\theta)\, d\theta}{\int l(x|\theta)\, \pi(\theta)\, d\theta}. \end{aligned}$$

Practical implementation is hindered by the fact that $L(a)$ and hence the minimum a^* could be sensitive to the chosen prior $\pi(\cdot)$, likelihood $l(\cdot|\cdot)$ and/or loss function $l(\cdot)$. A skeptical decision maker will require, in addition to the optimal solution, some description of the sensitivity of a^* with respect to reasonable changes and uncertainties in the specification of the inputs. This type of sensitivity is known as functional sensitivity because the inputs are functions. Excellent summaries of Bayesian literature in this area are provided by Berger [11] and Ríos and Ruggeri [3].

In this context, we can investigate the local parametric sensitivity of $L_\lambda(a^*)$ where λ is a possibly multidimensional parameter that models the loss function and/or the prior distribution. Now, $f_\lambda(\theta) = l_\lambda(a^*, \theta)$ and the quantity of interest $\mathcal{I}_{\lambda^0}$ is $L_{\lambda^0}(a^*)$. Note that we refer to expected loss sensitivity instead of decision sensitivity (see Kadane and Srinivasan [12] for a distinction).

In the next section, we show how the proposed computationally low-cost sensitivity estimations and their errors can be easily calculated in practice.

3. APPLICATION

We consider an illustrative example relating to 10 power plant pumps. George et al. [13] provided a complete Bayesian hierarchical analysis of the pump failure data previously studied by Gaver and O'Muircheartaigh [14]. For the power plant pump i, the failure rate

i	1	2	3	4	5	6	7	8	9	10
t_i	94.32	15.72	62.88	12.76	5.24	31.44	1.05	1.05	2.09	10.48
x_i	5	1	5	14	3	19	1	1	4	22

Table 1. Pump failure data.

is denoted by θ_i and the length of operation time (in thousands of hours) is denoted by t_i. The data are given in Table 1.

Conditional on θ_i, the number of failures X_i is assumed to follow a Poisson distribution, $X_i|\theta_i \sim \text{Poisson}(\eta_i)$, $i = 1, \ldots, 10$, where $\eta_i = \theta_i\, t_i$ and X_i is independent of X_j for $i \neq j$. Conditional on α and β, independent gamma prior distributions are adopted for the failure rates, $\theta_i|\alpha, \beta \sim \text{Gamma}(\alpha, \beta)$. We assume the following prior specification for α and β :

$$\begin{aligned} \alpha &\sim \text{Exp}(\lambda_1), \\ \beta &\sim \text{Gamma}(\lambda_2, \lambda_3), \end{aligned}$$

where $\lambda_1 = 1$, $\lambda_2 = 0.1$, and $\lambda_3 = 1$. The model is graphically represented in Figure 1. This graph has been obtained by using DoodleBUGS that has been developed to specify graphical models in Bayesian context. This tool is included in WinBUGS PACKAGE (Spiegelhalter [15]).

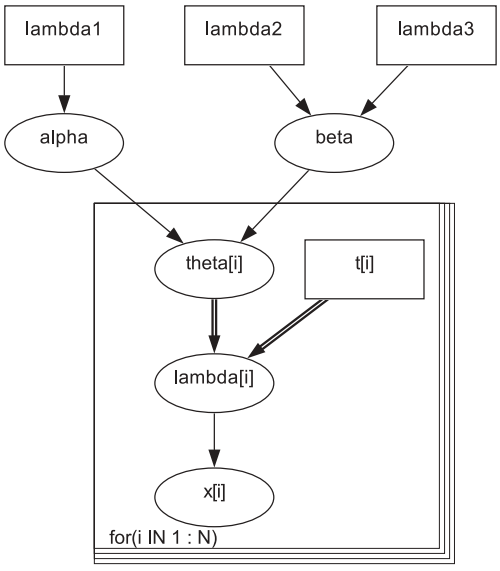


Figure 1. Graphical model.

We carry out a sensitivity analysis in the terms described in the previous section. We focus our interest on the posterior mean for the parameters θ_i, $i = 1, 2, \ldots, 10$. Those

quantities represent the means of the failure rates after the Bayes update has been done. We study if the posterior means of the parameters are sensitive to the initial values of the prior specification, i.e. we study local sensitivity with respect to the parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ in the neighborhood of $\boldsymbol{\lambda}^0 = (1, 0.1, 1)$. In this case, the quantities of interest are $E_{\boldsymbol{\lambda}^0}[\theta_i|x]$, and, in order to simplify, they will be denoted by $\mathcal{E}_{\boldsymbol{\lambda}^0}(i)$, $i = 1, 2, \ldots, 10$.

By using WinBUBS, we can generate MCMC samples from the posterior distributions for all parameters. After we consider that the convergence has been achieved, we generate a sample of size $n = 10000$. The estimations of $\mathcal{E}_{\boldsymbol{\lambda}^0}(i)$, are given in Table 2.

Parameters	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
$\widehat{\mathcal{E}_{\boldsymbol{\lambda}^0}(i)}$	0.059	0.102	0.089	0.116	0.604	0.609	0.893	0.881	1.584	1.992

Table 2. Estimations of the posterior means.

Table 3 shows the estimations of the partial derivatives $\partial_{\lambda_j}\mathcal{E}_{\boldsymbol{\lambda}^0}(i)$, $j = 1, 2, 3$, $i = 1, 2, \ldots, 10$.

	λ_1	λ_2	λ_3
θ_1	$1.34 \cdot 10^{-4}$	$-7.11 \cdot 10^{-5}$	$7.25 \cdot 10^{-5}$
θ_2	$4.34 \cdot 10^{-4}$	$3.46 \cdot 10^{-4}$	$-4.37 \cdot 10^{-4}$
θ_3	$4.58 \cdot 10^{-5}$	$2.25 \cdot 10^{-4}$	$-2.09 \cdot 10^{-4}$
θ_4	$1.13 \cdot 10^{-5}$	$2.69 \cdot 10^{-4}$	$-2.28 \cdot 10^{-4}$
θ_5	$1.45 \cdot 10^{-4}$	$-1.41 \cdot 10^{-3}$	$7.66 \cdot 10^{-4}$
θ_6	$1.13 \cdot 10^{-4}$	$3.69 \cdot 10^{-4}$	$-3.68 \cdot 10^{-5}$
θ_7	$-2.27 \cdot 10^{-3}$	$-3.09 \cdot 10^{-3}$	$5.72 \cdot 10^{-3}$
θ_8	$2.41 \cdot 10^{-3}$	$7.69 \cdot 10^{-3}$	$-7.33 \cdot 10^{-3}$
θ_9	$3.83 \cdot 10^{-3}$	$-1.19 \cdot 10^{-3}$	$5.34 \cdot 10^{-4}$
θ_{10}	$7.58 \cdot 10^{-4}$	$-8.09 \cdot 10^{-2}$	$6.83 \cdot 10^{-3}$

Table 3. Estimations of the partial derivatives.

We consider that the rate of change for λ_1, λ_2 and λ_3 are within the reasonable limits with respect to the values of $\widehat{\mathcal{E}_{\boldsymbol{\lambda}^0}(i)}$, $i = 1, 2, \ldots, 10$. So the components of $\widehat{\nabla\mathcal{E}_{\boldsymbol{\lambda}^0}(i)}$ indicate that we can consider $\boldsymbol{\lambda^0} = (1, 0.1, 1)$ as a robust value for the parameter $\boldsymbol{\lambda}$ in this model.

4. CONCLUSION

In Bayesian decision theory and inference the proposed local parametric sensitivity procedure can be very useful because it is a general technique applicable to complex models that need to be solved by MCMC methods. Besides, the MCMC simulations can be reused to estimate the sensitivity measures and their errors, avoiding the need of further sampling. This computationally low-cost method is easy to apply in practice and it is specially recommended to study sensitivities in reliability models.

ADKNOWLEDGEMENTS

Comments from James C. Spall are gratefully acknowledged. This research was partially supported by Junta de Extremadura, Spain (Project IPR00A075).

REFERENCES

1. S. P. Brooks. Markov chain Monte Carlo method and its application. *The Statistician*, 47:69–100, 1998.
2. A. Saltelli, K. Chan, and E.M. Scott. *Sensitivity Analysis*. Wiley, 2000.
3. D. Ríos Insua and F. Ruggeri. *Robust Bayesian Analysis*. Springer, 2000.
4. C. B. Hall, J. Y. Lynn Kuo, and R. B. Lipton. Bayesian and profile likelihood change point methods for modeling cognitive function over time. *Computational Statistics and Data Analysis*, 42(1-2):91–109, 2003.
5. U. Halekoh and W. Vach. A Bayesian approach to seriation problems in archaeology. *Computational Statistics and Data Analysis.*, 2003. In press, corrected proof, available online.
6. T. Turányi and H. Rabitz. *Sensitivity Analysis*, chapter Local Methods, pages 81–99. Wiley, 2003.
7. M.A. Tanner. *Tools for Statistical Inference*. Springer-Verlag, 1996.
8. J. C. Spall. *Introduction to Stochastic Search and Optimization. Estimation, Simulation and Control*. Wiley, 2003.
9. J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.
10. S. French and D. Ríos Insua. *Statistical Decision Theory*. Arnold, 2000.
11. J. O. Berger. An overview of robust Bayesian analysis. *Test*, 3:5–124, 1994.
12. J. Kadane and C. Srinivasan. Comment on "An Overview of Robust Bayesian Analysis" by Berger. *Test*, 3:93–95, 1994.
13. U. George, E. I. Makov and A. F. M. Smith. Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, 20:147–156, 1993.
14. D. P. Gaver and I. G. O′Muircheartaigh. Robust empirical Bayes analysis of event rates. *Technometrics*, 29:1–15, 1987.
15. D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. *WinBUGS User Manual, Version 1.4*. MRC Biostatistics Unit, Cambridge, 2003.

Sensitivity Analysis in Conjunction with Evidence Theory Representations of Epistemic Uncertainty

J.C. Helton, J.D. Johnson, W.L. Oberkampf

Sandia National Laboratories, Albuquerque, NM 87185-0779
E-mail: jchelto@sandia.gov, jdjohns@sandia.gov, wloberk@sandia.gov

Abstract: Three applications of sampling-based sensitivity analysis in conjunction with evidence theory representations for epistemic uncertainty in model inputs are described: (i) an initial exploratory analysis to assess model behavior and provide insights for additional analysis, (ii) a stepwise analysis showing the incremental effects of uncertain variables on complementary cumulative belief functions and complementary cumulative plausibility functions, and (iii) a summary analysis showing a spectrum of variance-based sensitivity analysis results that derive from probability spaces that are consistent with the evidence space under consideration.

Keywords: Epistemic uncertainty, evidence theory, sensitivity analysis, uncertainty analysis

1. INTRODUCTION

Uncertainty analysis and sensitivity analysis should be important components of any analysis of a complex system, with (i) uncertainty analysis providing a representation of the uncertainty present in the estimates of analysis outcomes and (ii) sensitivity analysis identifying the contributions of individual analysis inputs to the uncertainty in analysis outcomes[1]. Probability theory provides the mathematical structure traditionally used in the representation of epistemic (i.e., state of knowledge) uncertainty, with the uncertainty in analysis outcomes represented with probability distributions and typically summarized as cumulative distribution functions (CDFs) or complementary cumulative distribution functions (CCDFs) [2-4]. A variety of sensitivity analysis procedures have been developed for use in conjunction with probabilistic representations of uncertainty, including differential analysis [5, 6], the Fourier amplitude sensitivity test (FAST) and related variance decomposition procedures[7-11], regression-based techniques [12, 13], and searches for nonrandom patterns [14].

Although probabilistic representations of uncertainty have been successfully employed in many analyses, such representations have been criticized for inducing an appearance of more refined knowledge with respect to the existing uncertainty than is really present [15, 16]. Much of this criticism derives from the use of uniform distributions to characterize uncertainty in the presence of little or no knowledge with respect to where the appropriate value to use for a parameter is located within a set of possible values. As a result, a number of alternative mathematical structures for the representation of epistemic uncertainty have been proposed, including evidence theory, possibility theory, and fuzzy set theory [17].

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

Evidence theory provides a promising alternative to probability theory that allows for a fuller representation of the implications of uncertainty than is the case in a probabilistic representation of uncertainty. In particular, evidence theory involves two representations of the uncertainty associated with a set of possible analysis inputs or results: (i) a belief, which provides a measure of the extent to which the available information implies that the true value is contained in the set under consideration, and (ii) a plausibility, which provides a measure of the extent to which the available information implies that the true value might be contained in the set under consideration. One interpretation of the belief and plausibility associated with a set is that (i) the belief is the smallest possible probability for the set that is consistent with all available information and (ii) the plausibility is the largest possible probability for the set that is consistent with all available information. An alternative interpretation is that evidence theory is an internally consistent mathematical structure for the representation of uncertainty without any explicit conceptual link to probability theory. The mathematical operations associated with evidence theory are the same for both interpretations. Just as probability theory uses CDFs and CCDFs to summarize uncertainty, evidence theory uses cumulative belief functions (CBFs), cumulative plausibility functions (CPFs), complementary cumulative belief functions (CCBFs), and complementary cumulative plausibility functions (CCPFs) to summarize uncertainty.

Although evidence theory is beginning to be used in the representation of uncertainty in applied analyses, the authors are unaware of any attempts to develop sensitivity analysis procedures for use in conjunction with evidence theory. Due to the importance of sensitivity analysis in any decision-aiding analysis, the potential usefulness of evidence theory will be enhanced if meaningful and practicable sensitivity analysis procedures are available for use in analyses that employ evidence theory in the representation of uncertainty. As a result, the focus of this presentation is on the development of sensitivity analysis procedures for use in conjunction with evidence theory representations of uncertainty.

After a brief overview of evidence theory (Sect. 2), the following topics are considered: (i) exploratory sensitivity analysis (Sect. 3), (ii) use of sensitivity analysis results in the stepwise construction of CCBFs and CCPFs (Sect. 4), (iii) analysis of evidence theory representations of uncertainty (Sect. 5), and (iv) concluding summary (Sect. 6).

2. EVIDENCE THEORY

Evidence theory is based on the specification of a triple ($\mathcal{S}$, $\mathcal{A}$, m), where (i) $\mathcal{S}$ is the set that contains everything that could occur in the particular universe under consideration, (ii) $\mathcal{A}$ is a countable collection of subsets of $\mathcal{S}$, and (iii) m is a function defined on subsets of $\mathcal{S}$ such that $m(\mathcal{E}) > 0$ if $\mathcal{E} \in \mathcal{A}$, $m(\mathcal{E}) = 0$ if $\mathcal{E} \subset \mathcal{S}$ and $\mathcal{S} \notin \mathcal{A}$, and $\Sigma_{\mathcal{E} \in \mathcal{A}} m(\mathcal{E}) = 1$. For a subset $\mathcal{E}$ of $\mathcal{S}$, $m(\mathcal{E})$ characterizes the amount of "likelihood" that can be assigned to $\mathcal{E}$ but to no proper subset of $\mathcal{E}$. In the terminology of evidence theory, (i) $\mathcal{S}$ is the sample space or universal set, (ii) $\mathcal{A}$ is the set of focal elements for $\mathcal{S}$ and m, and (iii) $m(\mathcal{E})$ is the basic probability assignment (BPA) associated with a subset $\mathcal{E}$ of $\mathcal{S}$. The elements of $\mathcal{S}$ are often vectors $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, where each element x_i of $\mathbf{x}$ is a variable with its own evidence space ($\mathcal{S}_i$, $\mathcal{A}_i$, m_i). When the x_i's are assumed to be independent, (i) $m(\mathcal{E})$

$= \Pi_i m_i(\mathcal{E}_i)$ if $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2 \times \ldots \times \mathcal{E}_n$ and $\mathcal{E}_i \in \mathcal{S}_i$ for $i = 1, 2, \ldots, n$ and (ii) $m(\mathcal{E}) = 0$ otherwise. An evidence space $(\mathcal{S}, \mathcal{S}, m)$ plays the same role in evidence theory that a probability space $(\mathcal{P}, \mathcal{P}, p)$ plays in probability theory, where $\mathcal{P}$ is the sample space, $\mathcal{P}$ is a suitably restricted set of subsets of $\mathcal{P}$ (i.e., a σ-algebra), and p is the function (i.e., probability measure) that assigns probabilities to elements of $\mathcal{P}$.

The belief, $Bel(\mathcal{E})$, and plausibility, $Pl(\mathcal{E})$, for a subset $\mathcal{E}$ of $\mathcal{S}$ are defined by

$$Bel(\mathcal{E}) = \sum_{\mathcal{U} \subset \mathcal{E}} m(\mathcal{U}) \text{ and } Pl(\mathcal{E}) = \sum_{\mathcal{U} \cap \mathcal{E} \neq \phi} m(\mathcal{U}). \tag{2.1}$$

In concept, $Bel(\mathcal{E})$ is the amount of "likelihood" that must be assigned to $\mathcal{E}$, and $Pl(\mathcal{E})$ is the maximum amount of "likelihood" that could possibly be assigned to $\mathcal{E}$. When the elements of $\mathcal{S}$ are real valued, a CCBF and a CCPF provide a convenient summary of an evidence space $(\mathcal{S}, \mathcal{S}, m)$ and correspond to plots of the points

$$CC\mathcal{BF} = \left\{\left[v, Bel(\mathcal{S}_v)\right], v \in \mathcal{S}\right\} \text{ and } CC\mathcal{PF} = \left\{\left[v, Pl(\mathcal{S}_v)\right], v \in \mathcal{S}\right\}, \tag{2.2}$$

where $\mathcal{S}_v = \{x: x \in \mathcal{S} \text{ and } x > v\}$.

An important situation in the application of evidence theory is the consideration of a variable $y = f(\mathbf{x})$, where f is a function defined for elements $\mathbf{x}$ of the sample space $\mathcal{X}$ associated with an evidence space $(\mathcal{X}, \mathcal{X}, m_X)$ and $\mathbf{x}$ is represented as a vector because this is the case in most real analyses. The properties of f and $(\mathcal{X}, \mathcal{X}, m_X)$ induce an evidence space $(\mathcal{Y}, \mathcal{Y}, m_Y)$ on y, which provides a characterization of the uncertainty associated with y. In turn, this uncertainty can be summarized with a CCBF and a CCPF defined by

$$CC\mathcal{BF} = \left\{\left[v, Bel_X\left\{f^{-1}(\mathcal{Y}_v)\right\}\right], v \in \mathcal{Y}\right\} \text{ and } CC\mathcal{PF} = \left\{\left[v, Pl_X\left\{f^{-1}(\mathcal{Y}_v)\right\}\right], v \in \mathcal{Y}\right\}, \tag{2.3}$$

where Bel_X and Pl_X denote belief and plausibility defined with respect to $(\mathcal{X}, \mathcal{X}, m_X)$ and $\mathcal{Y}_v = \{y: y \in \mathcal{Y} \text{ and } y > v\}$. The generation and analysis of CCBFs and CCPFs of the preceding form are fundamental parts of the use of evidence theory to characterize the uncertainty in model predictions.

3. EXPLORATORY SENSITIVITY ANALYSIS

An initial exploratory sensitivity analysis plays an important role in helping to guide any study that involves uncertain inputs. This is particularly true in uncertainty analyses based on evidence theory as the uncertainties are likely to be large and an appropriate understanding of these uncertainties and their implications can provide insights that facilitate the computational estimation of beliefs and plausibilities.

Given that large uncertainties in many variables are likely to be present, a sampling-based approach to sensitivity analysis with Latin hypercube sampling [18, 19] is a broadly applicable procedure for an exploratory analysis in conjunction with an evidence theory representation for uncertainty. Use of this approach requires the specification of distributions for the uncertain variables for sampling purposes. This specification should provide for an adequate exploration of the range of each uncertain variable and be consis-

tent, in some sense, with the evidence theory specification of the uncertainty associated with individual analysis inputs.

A distribution that meets the preceding criteria can be obtained by sampling each focal element associated with a variable in consistency with its BPA and then sampling uniformly within that focal element. With the assumption that each focal element for a variable x_i with an evidence space ($\mathcal{X}_i$, $\mathscr{X}_i$, m_i) is an interval, this corresponds to defining a sampling distribution with a density function d_i given by

$$d_i(v) = \sum_{j=1}^{C(\mathscr{X}_i)} \delta_{ij}(v)\, m_i(\mathcal{E}_{ij}) \big/ (b_{ij} - a_{ij}), \tag{3.1}$$

where (i) $v \in \mathcal{X}_i$, (ii) $C(\mathscr{X}_i)$ is the cardinality of $\mathscr{X}_i$, (iii) $\mathcal{E}_{ij} = [a_{ij}, b_{ij}]$, $j = 1, 2, \ldots, C(\mathscr{X}_i)$, are the focal elements associated with x_i (i.e., the elements of $\mathscr{X}_i$), and (iv) $\delta_{ij}(v) = 1$ if $v \in \mathcal{E}_{ij}$ and 0 otherwise. Appropriate modifications can be made to the preceding definition to handle focal elements with a finite number of elements and focal elements that are unions of disjoint intervals.

Given that a relationship of the form $y = f(\mathbf{x})$, $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, is under consideration, sampling according to the distributions indicated in Eq. (3.1) generates a mapping $y_k = f(\mathbf{x}_k)$ from uncertain analysis inputs to uncertain analysis results, where $\mathbf{x}_k$, $k = 1, 2, \ldots, nS$, are the sampled values for $\mathbf{x}$. As previously indicated, Latin hypercube sampling is a likely candidate for the sampling procedure because of its efficient stratification properties. Once this mapping is generated, it can be explored with various sensitivity analysis procedures to develop an understanding of the relationship between y and the individual elements of $\mathbf{x}$.

A variety of techniques are available for use in sampling-based sensitivity analyses [13, 20]. However, given that the analysis problem is based on evidence theory, sensitivity analysis procedures that do not place excessive reliance on the sampling distributions indicated in Eq. (3.1) are desirable. Of course, no approach can fully divorce itself from these distributions because they ultimately give rise to the raw material of the sensitivity analysis (i.e., the mapping $[\mathbf{x}_k, y_k]$, $k = 1, 2, \ldots, nS$); however, this is an unavoidable situation when the sample space associated with $\mathbf{x}$ is infinite as no approach can consider all values of $\mathbf{x}$ and so a subset of the values for $\mathbf{x}$ must be selected in some manner. The examination of scatterplots is a natural initial procedure. Then, rank-based procedures (e.g., rank repression, partial rank correlation, squared rank differences) are natural techniques to employ because they reduce the effects of both nonlinearities and the original sampling distributions [13, 21, 22].

If carried out successfully, an initial exploratory sensitivity analysis should provide important insights with respect to the relationship between y and the elements of $\mathbf{x}$. Often, only a few of the elements of $\mathbf{x}$ will have significant effects on y. This is information that can be productively used in the estimation of the evidence theory structure associated with y.

4. STEPWISE CONSTRUCTION OF CCBFs AND CCPFs

For most models, the determination of beliefs and plausibilities for model predictions in general, and CCBFs and CCPFs in particular, is a demanding numerical challenge due to the need to determine the inverse of the model (i.e., function) involved. Sampling-based (i.e., Monte Carlo) procedures provide one way to carry out such determinations. With this approach, a sample $\mathbf{x}_k$, $k = 1, 2, \ldots, nS$, is generated from $\mathcal{X}$ (e.g., with distributions for the elements of $\mathbf{x}$ of the form indicated in Eq. (3.1)), and y is evaluated for each $\mathbf{x}_k$ to create the mapping $[\mathbf{x}_k, y_k]$, $k = 1, 2, \ldots, nS$, from $\mathcal{X}$ to $\mathcal{Y}$. Then, the CCBF and CCPF for y can be estimated by

$$\mathcal{CCBF} \cong \left\{\left[y, 1 - Pl_X\left(\{\mathbf{x}_k : y_k \le y\}\right)\right], y \in \mathcal{Y}\right\} \tag{4.1}$$

and

$$\mathcal{CCPF} \cong \left\{\left[y, Pl_X\left(\{\mathbf{x}_k : y_k > y\}\right)\right], y \in \mathcal{Y}\right\}, \tag{4.2}$$

respectively. The approximation to $\mathcal{CCBF}$ for y in Eq. (4.1) is based on the equality $Bel(\mathcal{E}) + Pl(\mathcal{E}^c) = 1$ and the fact that the subset criterion in the definition of belief (see Eq. (2.1)) does not allow for the direct estimate of belief with a finite sample when sets with infinite numbers of elements are under consideration. In general, the same approach can be used to estimate the belief $Bel_Y(\mathcal{E})$ and plausibility $Pl_Y(\mathcal{E})$ for any subset $\mathcal{E}$ of $\mathcal{Y}$.

The problem with the preceding approach is that it can be prohibitively expensive computationally when the cardinality $C(\mathcal{X})$ of $\mathcal{X}$ is high, which is usually the case in real analyses. Specifically, $C(\mathcal{X}) = \Pi_i\, C(\mathcal{X}_i)$, where $C(\mathcal{X}_i)$ is the cardinality of $\mathcal{X}_i$. For example, if $n = 8$ and $C(\mathcal{X}_i) = 10$, then $C(\mathcal{X}) = 10^8$; and as a result, a very large sample would be required to converge the approximations to the CCBF and CCPF in Eqs. (4.1) and (4.2).

The results of the exploratory sensitivity analysis described in Sect. 3 provide a basis for a potential path forward in developing the CCBF and CCPF approximations in Eqs. (4.1) and (4.2). The uncertainty in most analysis outcomes is significantly affected by the uncertainty in only a small number of analysis inputs (e.g., 3-5). Of course, this does not have to be the case but it does seem usually to be the case. In this situation, the approximations in Eqs. (4.1) and (4.2) can be determined by only considering the uncertainty (i.e., the evidence spaces $(\mathcal{X}_i, \mathcal{X}_i, m_i)$) associated with the x_i that significantly affect y. The remaining x_i (i.e., those that do not have a significant effect on y) can be assigned degenerate evidence spaces (i.e., spaces $(\mathcal{X}_i, \mathcal{X}_i, m_i)$ for which $m_i(\mathcal{X}_i) = 1$) for use in evaluating the approximations in Eqs. (4.1) and (4.2).

Increasing the resolution in the evidence spaces assigned to individual x_i (i.e., by subdividing elements of $\mathcal{X}_i$ and then apportioning the BPA for an original element of $\mathcal{X}_i$ over the subsets into which it is subdivided) tends to decrease, and can never increase, the uncertainty associated with evidence space for y. Specifically, beliefs tend to increase (and can never decrease) and plausibilities tend to decrease (and can never increase); or put another way, beliefs and plausibilities for subsets of $\mathcal{Y}$ move closer together as the resolution in the characterization of the uncertainties associated with the x_i is increased.

The preceding observations provide a basis for the use of sensitivity analysis results to guide a stepwise procedure for the construction of the CCBF and CCPF approximations in Eqs. (4.1) and (4.2). At Step 1, the approximations in Eqs. (4.1) and (4.2) are determined with the most important variable affecting the uncertainty in y assigned its original evidence space and all other variables assigned evidence spaces in which their original sample spaces are assigned a BPA of 1. At Step 2, the approximations in Eqs. (4.1) and (4.2) are determined with the two most important variables affecting the uncertainty in y assigned their original evidence spaces and all other variables assigned evidence spaces in which their original sample spaces are assigned a BPA of 1. Analogous steps follow for additional important variables determined in the sensitivity analysis until substantive changes in the CCBF and CCPF approximations in Eqs. (4.1) and (4.2) no longer occur, at which point the approximation procedure stops. This approach can produce substantial computational savings over what would be incurred if the approximations in Eqs. (4.1) and (4.2) were evaluated with the original evidence spaces assigned to all the x_i.

The construction procedure just outlined can also can be viewed as a sensitivity analysis in the context of evidence theory. The changes in the location of the CCBF and CCPF as additional variables are added in the preceding procedure provides an indication of the importance of individual variables with respect to the uncertainty in y characterized by ($\mathcal{Y}$, $\mathscr{Y}$, m_Y). At an intuitive level, this approach is analogous to the use of stepwise regression analysis in traditional sensitivity analyses.

5. SUMMARY SENSITIVITY ANALYSIS

Together, a CCBF and CCPF for y provide bounds on all possible CCDFs for y that could derive from different distributions for the x_i that are consistent with their specified evidence spaces ($\mathcal{X}_i$, $\mathscr{X}_i$, m_i). In particular, if ($\mathcal{P}_i$, $\mathscr{P}_i$, p_i) is a probability space for x_i that is consistent with the evidence space ($\mathcal{X}_i$, $\mathscr{X}_i$, m_i) for i = 1, 2, …, n, then these probability spaces give rise to corresponding probability spaces ($\mathcal{P}_X$, $\mathscr{P}_X$, p_X) and ($\mathcal{P}_Y$, $\mathscr{P}_Y$, p_Y) for **x** and y with the CCDF associated with ($\mathcal{P}_Y$, $\mathscr{P}_Y$, p_Y) falling somewhere between the CCBF and CCPF for y. Traditional sensitivity analysis methods can be used to investigate the relationships between the uncertainty in the x_i characterized by the probability spaces ($\mathcal{P}_i$, $\mathscr{P}_i$, p_i) and the uncertainty in y characterized by the probability space ($\mathcal{P}_Y$, $\mathscr{P}_Y$, p_Y). A possible approach is a variance decomposition for y that partitions the variance for y into the contributions to this variance from the individual x_i [8-10]. However, unlike a traditional sensitivity analysis in which the probability spaces ($\mathcal{P}_i$, $\mathscr{P}_i$, p_i) are uniquely specified, there are many possibilities for the spaces ($\mathcal{P}_i$, $\mathscr{P}_i$, p_i) in an evidence theory context and thus many possible variance decompositions for y. In variance-based sensitivity analysis, the variance $V(y)$ of y is expressed as

$$V(y) = \sum_{i=1}^{n} V_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} V_{ij} + \cdots + V_{12\cdots n}, \tag{5.1}$$

where V_i is the contribution of x_i to $V(y)$, V_{ij} is the contribution of the interaction of x_i and x_j to $V(y)$, and so on up to $V_{12\ldots n}$ which is the contribution of the interaction of x_1, x_2, …, x_n to $V(y)$. Possible sensitivity measures are provided by

$$s_i = V_i / V(y) \text{ and } s_{iT} = \left(V_i + \sum_{j \neq i} V_{ij} + \cdots + V_{12\cdots n} \right) \Big/ V(y), \tag{5.2}$$

where s_i the fraction of $V(y)$ contributed by x_i alone and s_{iT} is the fraction of $V(y)$ contributed by x_i and interactions of x_i with other variables. The term V_i is defined by iterated integrals involving the probability spaces for the individual variables. For example, when $n = 3$,

$$V_1 = \int_{\mathcal{P}_1} \left[\int_{\mathcal{P}_2} \int_{\mathcal{P}_3} f(x_1, x_2, x_3)\, d_3(x_3)\, d_2(x_2)\, dx_3 dx_2 \right]^2 d_1(x_1)\, dx_1 - E^2(y), \tag{5.3}$$

where d_i denotes the density function associated with ($\mathcal{P}_i$, $\mathscr{P}_i$, p_i) and $E(y)$ denotes the expected value of y; similar defining integrals hold for V_2 and V_3, and related, but more complicated, integrals define V_{12}, V_{13}, V_{23} and V_{123}. Analogous relationships hold for $n > 3$. By suitably orchestrating an analysis, V_i and s_i for $i = 1, 2, \ldots, n$ can be estimated with two independent random or Latin hypercube samples; further, s_i and s_{iT} for $i = 1, 2, \ldots, n$ can be estimated with a total of $n + 2$ suitably defined samples.

Three questions arise with respect to the implementation of a variance-based sensitivity analysis in the context of evidence theory: (i) How to select an appropriate spectrum of distributions for each x_i from the infinite number of distributions that are consistent with ($\mathscr{X}_i$, $\mathscr{X}_i$, m_i)?, (ii) How to implement the analysis in a computationally practicable manner for multiple distributions (i.e., multiple probability spaces ($\mathcal{P}_i$, $\mathscr{P}_i$, p_i)) for each x_i?, and (iii) How to display the results of the sensitivity analyses for multiple distributions of the x_i and hence multiple distributions for $\mathbf{x}$ and y?

The first question arises because there is no inherent structure associated with the infinite number of distributions for x_i that are consistent with ($\mathscr{X}_i$, $\mathscr{X}_i$, m_i). The situation is analogous to that encountered in an interval analysis for a real-valued quantity except that the uncertain quantity is now a probability space rather than a number. As there is no way to consider all probability spaces consistent with ($\mathscr{X}_i$, $\mathscr{X}_i$, p_i) and also no specific structure to guide the selection of individual probability spaces, some type of ad hoc procedure is needed to select representative probability spaces that are consistent with ($\mathscr{X}_i$, $\mathscr{X}_i$, p_i). Further, the number of selected distributions for each x_i must be relatively small; otherwise, the total number of combinations of selected distributions for all n variables will be too large to be computationally practicable.

An exploratory approach that should provide valuable information for many situations is to select three distributions for each x_i, with (i) one distribution emphasizing the smaller values associated with each focal element, (ii) one distribution uniform over the range of each focal element, and (iii) one distribution emphasizing the larger values associated with each focal element. The distributions indicated in (i) and (iii) could be left and right triangular or left and right quadratic. Left and right triangular distributions are actually quite similar to uniform distributions and thus may not be good choices. For fo-

cal element $\mathcal{E}_{ij} = [a_{ij}, b_{ij}]$ associated with x_i, the corresponding density functions d_{lij}, d_{uij} and d_{rij} for left quadratic, uniform, and right quadratic distributions, respectively, over $\mathcal{E}_{ij}$ are

$$d_{lij}(v) = \frac{3(b_{ij} - v)^2}{(b_{ij} - a_{ij})^3},\ d_{uij}(v) = \frac{1}{(b_{ij} - a_{ij})},\ \text{and}\ d_{rij}(v) = \frac{3(v - a_{ij})^2}{(b_{ij} - a_{ij})^3} \tag{5.4}$$

if $v \in \mathcal{E}_{ij}$ and $d_{lij}(v) = d_{uij}(v) = d_{rij}(v) = 0$ otherwise. In turn, the left quadratic, uniform and right quadratic distribution functions d_{li}, d_{ui} and d_{ri} for x_i are given by

$$d_{ci}(v) = \sum_{j=1}^{C(\mathcal{X}_i)} m_i(\mathcal{E}_{ij})\, d_{cij}(v) \tag{5.5}$$

for $v \in \mathcal{X}_i$ and $c = l, u, r$.

The second question arises because computational cost can easily become unreasonable unless the analysis is carefully planned. As a first step, only those variables that actually affect y need to be considered. The preliminary sensitivity analysis described in Sect. 3 should, in most analyses, identify the four or five variables that have significant effects on y. It is only those variables that require consideration of their original evidence spaces as indicated in Eq. (5.5); the remaining variables can be assigned a uniform or some other convenient distribution. For example, if four x_i affect y and the three distributions defined in Eq. (5.5) are considered for each of these x_i, then $3^4 = 81$ different probability spaces result for $\mathbf{x}$ and hence for y. As a second step, the analysis can be designed to use the same samples in the evaluation of s_i and s_{iT} for all probability spaces defined for $\mathbf{x}$ (e.g., the 81 spaces indicated above). For example, if Latin hypercube sampling is used, it is necessary to actually evaluate f for samples from only one of the probability spaces for $\mathbf{x}$; after these evaluations for f are performed, results for the other probability spaces for $\mathbf{x}$ under consideration (e.g., the remaining 80 probability spaces in the example above) can be obtained by reweighting the results obtained for the individual sample elements on the basis of the changed distributions for the x_i's [19, 23]. A similar reweighting procedure is also available for random sampling [24].

The third question arises because of the difficulty of displaying the results of multiple sensitivity analyses for y in a reasonably compact and understandable format. Presenting the sensitivity analyses individually is unlikely to be adequate because of the large number of analyses involved and the resultant difficulty of observing trends in variable importance across analyses. A promising presentation format to employ for this representation is a cobweb plot, which provides a representation for a multidimensional distribution in a two-dimensional plot [25]. For example, if nPS probability spaces ($\mathcal{P}_{Xj}$, $\mathcal{P}_{Xj}$, p_{Xj}) for $\mathbf{x}$ are under consideration and 4 uncertain variables have been identified for analysis, the results of the sensitivity analyses for y might be of the form

$$\mathbf{s}_j = \left[e_j, v_j, s_{1j}, s_{2j}, s_{3j}, s_{4j} \right],\ j = 1, 2, \ldots, nPS, \tag{5.6}$$

where e_j and v_j are the expected value and variance for y that derive from the probability space ($\mathcal{P}_{Xj}$, $\wp_{Xj}$, p_{Xj}) for $\mathbf{x}$ and s_{ij}, $i = 1, 2, 3, 4$, are the fractional contributions to v_j as defined in the first equality in Eq. (5.2) for the 4 uncertain variables under consideration.

With a cobweb plot, the *nPS* vectors in Eq. (5.6) can be presented in a single plot frame. Specifically, the individual elements of $\mathbf{s}_j$ are designated by locations on the horizontal axis and their values correspond to locations on the vertical axis. In general, it may be necessary to use multiple axis scales for the vertical axis or to plot quantiles for the elements of $\mathbf{s}_j$ rather than their actual values. Each $\mathbf{s}_j$ results in a single point in each of the vertical columns associated with its elements. The identity of $\mathbf{s}_j$ is maintained by a line that connects the values of its elements. As desired, the cobweb plot allows the presentation of all sensitivity analysis results in a single plot frame and also facilitates the recognition of interactions between variables.

In summary, the approach presented in this section to the performance and presentation of a sensitivity analysis for a function defined on an evidence space has three components: (i) Definition of representative probability spaces for the analysis input $\mathbf{x}$ that are consistent with the evidence space for $\mathbf{x}$, (ii) Use of efficient sampling-based numerical procedures to decompose the variance of the analysis outcome y for each probability space for $\mathbf{x}$, and (iii) Use of cobweb plots to summarize the results of the sensitivity analyses for y carried out for the individual probability spaces for $\mathbf{x}$. Thus, rather than having a single set of sensitivity analysis results for y, a spectrum of sensitivity analysis results for y is obtained that is consistent with the evidence space that characterizes the uncertainty in $\mathbf{x}$.

6. SUMMARY

Three applications of sampling-based sensitivity analysis in conjunction with evidence theory representations for epistemic uncertainty in model inputs have been described: (i) an initial exploratory analysis to assess model behavior and provide insights for additional analysis, (ii) a stepwise analysis showing the incremental effects of uncertain variables on CCBFs and CCPFs, and (iii) a summary analysis showing a spectrum of variance-based sensitivity analysis results that derive from probability spaces that are consistence with the evidence space under consideration. It is hoped that the ideas associated with these approaches will provide a start towards the development of effective sensitivity analysis procedures for use in conjunction with evidence theory representations for epistemic uncertainty.

ACKNOWLEDGMENTS

Work performed for Sandia National Laboratories, which is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Security Administration under contract DE-AC04-94AL-85000. Editorial support provided by F. Puffer and J. Ripple of Tech Reps, a division of Ktech Corporation.

REFERENCES

[1] A. Saltelli, K. Chan, and E.M. Scott, *Sensitivity Analysis*. New York: Wiley, 2000.

[2] G. W. Parry and P. W. Winter, "Characterization and Evaluation of Uncertainty in Probabilistic Risk Analysis," *Nuclear Safety*, vol. 22, pp. 28-42, 1981.

[3] G. Apostolakis, "The Concept of Probability in Safety Assessments of Technological Systems," *Science*, vol. 250, pp. 1359-1364, 1990.

[4] J. C. Helton, "Uncertainty and Sensitivity Analysis in the Presence of Stochastic and Subjective Uncertainty," *Journal of Statistical Computation and Simulation*, vol. 57, pp. 3-76, 1997.

[5] T. Turányi, "Sensitivity Analysis of Complex Kinetic Systems. Tools and Applications," *Journal of Mathematical Chemistry*, vol. 5, pp. 203-248, 1990.

[6] D. G. Cacuci, *Sensitivity and Uncertainty Analysis, Vol. 1: Theory*. Boca Raton, FL: CRC Press, 2003.

[7] R. I. Cukier, H. B. Levine, and K. E. Shuler, "Nonlinear Sensitivity Analysis of Multiparameter Model Systems," *Journal of Computational Physics*, vol. 26, pp. 1-42, 1978.

[8] I. M. Sobol', "Sensitivity Analysis for Nonlinear Mathematical Models," *Mathematical Modeling and Computational Experiment*, vol. 1, pp. 407-414, 1993.

[9] H. Rabitz, O. F. Alis, J. Shorter, and K. Shim, "Efficient Input-Output Model Representations," *Computer Physics Communications*, vol. 117, pp. 11-20, 1999.

[10] A. Saltelli, S. Tarantola, and K. P.-S. Chan, "A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output," *Technometrics*, vol. 41, pp. 39-56, 1999.

[11] M. D. McKay, J. D. Morrison, and S. C. Upton, "Evaluating Prediction Uncertainty in Simulation Models," *Computer Physics Communications*, vol. 117, pp. 44-51, 1999.

[12] R. L. Iman, "Uncertainty and Sensitivity Analysis for Computer Modeling Applications," presented at *Reliability Technology - 1992, The Winter Annual Meeting of the American Society of Mechanical Engineers, Anaheim, California, November 8-13, 1992*, Vol. 28, pp. 153-168. New York, NY, 1992.

[13] J. C. Helton and F. J. Davis, "Sampling-Based Methods," in *Sensitivity Analysis*, A. Saltelli, K. Chan, and E.M. Scott, Eds. New York, NY: Wiley. pp. 101-153, 2000.

[14] J. P. C. Kleijnen and J. C. Helton, "Statistical Analyses of Scatterplots to Identify Important Factors in Large-Scale Simulations, 1: Review and Comparison of Techniques," *Reliability Engineering and System Safety*, vol. 65, pp. 147-185, 1999.

[15] G. J. Klir, "Is there more to uncertainty than some probability theorists might have us believe?," *International Journal of General Systems*, vol. 15, pp. 347-378, 1989.

[16] G. J. Klir, "On the Alleged Superiority of Probability Representation of Uncertainty," *IEEE Transactions on Fuzzy Systems*, vol. 2, pp. 27-31, 1994.

[17] G. J. Klir and M. J. Wierman, "Uncertainty-Based Information," Physica-Verlag, New York, NY 1999.

[18] M. D. McKay, R. J. Beckman, and W. J. Conover, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, vol. 21, pp. 239-245, 1979.

[19] J. C. Helton and F. J. Davis, "Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems," *Reliability Engineering and System Safety*, vol. 81, pp. 23-69, 2003.

[20] J. C. Helton and F. J. Davis, "Illustration of Sampling-Based Methods for Uncertainty and Sensitivity Analysis," *Risk Analysis*, vol. 22, pp. 691-622, 2002.

[21] R. L. Iman and W. J. Conover, "The Use of the Rank Transform in Regression," *Technometrics*, vol. 21, pp. 499-509, 1979.

[22] S. C. Hora and J. C. Helton, "A Distribution-Free Test for the Relationship Between Model Input and Output when Using Latin Hypercube Sampling," *Reliability Engineering and System Safety*, vol. 79, pp. 333-339, 2003.

[23] R. L. Iman and W. J. Conover, "Small Sample Sensitivity Analysis Techniques for Computer Models, with an Application to Risk Assessment," *Communications in Statistics: Theory and Methods*, vol. A9, pp. 1749-1842, 1980.

[24] R. J. Beckman and M. D. McKay, "Monte-Carlo Estimation Under Different Distributions Using the Same Simulation," *Technometrics*, vol. 29, pp. 153-160, 1987.

[25] R. M. Cooke and J.M. van Noortwijk, "Graphical Methods," in *Sensitivity Analysis*, A. Saltelli, K. Chan, and E.M. Scott, Eds. Wiley: New York, NY, 2000, pp. 245-264.

Combined Array Experiment Design

L. M. Moore, M. D. McKay, and K. S. Campbell

Los Alamos National Laboratory, Box 1663, Los Alamos, NM 87545

E-mail: lmoore@lanl.gov, mdm@lanl.gov, ksc@lanl.gov

LA-UR-04-0374

Abstract: Experiment plans formed by combining two or more designs, such as orthogonal arrays, primarily with 2- and 3-level factors, creating multi-level arrays with subsets of different strength are proposed for computer experiments. The specific illustrations are designs for 5-level factors with fewer runs than generally required for 5-level orthogonal arrays of strength 2 or more. At least five levels for each input are desired to allow for runs at a nominal value, two values either side of nominal but within a normal, anticipated range, and two more extreme values either side of nominal. This number of levels allows for a broader range of input combinations to test the input combinations where a simulation code operates. Five-level factors also allow the possibility of up to 4th order polynomial models for fitting simulation results, at least in one dimension. By having subsets of runs with more than strength 2, interaction effects may also be considered. Also, the resulting designs have a "checker-board" pattern in lower-dimensional projections, in contrast to the grid projection that occurs with orthogonal arrays.

Keywords: Computer experiments, experiment design, fractional factorial design, orthogonal arrays, correlation coefficient

1. INTRODUCTION

The context for this paper is planning runs of a non-stochastic computer code for the purpose of assessing important inputs from among p inputs. As in McKay (1995), "important" input(s) are identified based on comparison of R^2, an estimate of the correlation coefficient associated with the goodness of fit to the simulated output Y of an analysis of variance model based on a subset of inputs X_s. The following is a formula for R^2 based on a subset of inputs X_s:

$$R^2(X_s) = \frac{\sum_{i \in X_s} \sum_j (y_{i\cdot} - y_{\cdot\cdot})^2}{\sum_{i \in X_s} \sum_j (y_{ij} - y_{\cdot\cdot})^2}$$

where the subscript i varies over distinct values of the s inputs identified in X_s, the subscript j varies over "replicate" experiments corresponding to a fixed value of the inputs X_s, and the "dot" subscript indicates the standard average. "Replicate" is in quotes since no true replicates are done. The computer simulation output is non-stochastic in that the output is fully determined by specification of the input with no variation in output for repeated runs of

the code for identical input. Variation in the output is induced solely by variation in the inputs. The (p-s) inputs identified by $X-X_s$ may differ while X_s is fixed and this is the basis of pseudo-replicate, or "replicate" runs for fixed values of X_s. The value $y_{i\cdot}$ will be identically y_{ij} if there are no pseudo-replicate runs. If this is the case for every value of the inputs identified by X_s, then R^2 will have a value identically 1. Otherwise, R^2 is between 0 and 1. This reasoning leads to considering experiment designs such that, for subsets of inputs of a specified size $s<p$, a sampling of values for that subset of inputs is required such that "replicates" determined by a sample of values for the remaining inputs occur, for at least one of the values of the subset of inputs. This is a property of factorial experiment designs, or orthogonal arrays, which naturally suit this analysis approach, per Moore and McKay (2002). However, in order to obtain non-degenerate values of R^2 for subsets of 2 or more inputs, orthogonal arrays of strength 2 or more are dictated.

The specific illustrations of experiment designs are for 5-level factors with fewer runs than generally required for 5-level orthogonal arrays of strength 2 or more. In statistical experiment design, particularly as used in industrial physical experiments, factorial experiments with 2 or 3 level factors are common. Here, at least five levels for each input are desired to allow for runs at a nominal value, two values either side of nominal but within a normal, anticipated range, and two more extreme values either side of nominal. This number of levels allows for a broader range of input combinations to test the input combinations where a simulation code operates. Five-level factors also allow the possibility of up to 4^{th} order polynomial models for fitting the simulation results, at least in one dimension.

The requirement for strength 2 or more arrays, in addition to requiring factors to have 5 levels, leads to orthogonal arrays with unacceptably large numbers of runs in some situations. Moore and McKay (2002) present a 625 run orthogonal array for up to 26 5-level factors that is strength 3. In fact, for 625 runs the maximum number of 5-level factors for which a strength 2 orthogonal array exists is 156. The maximum number for which a strength 3 array exists in 625 runs is 26 5-level factors, and the maximum number for a strength 4 array is 6 5-level factors. For 125 runs, the maximum strength for a 5-level orthogonal array is 3 and inequalities in Hedayat, et al (1999) show that the maximum number of 5-level factors that could be accommodated by a strength 3 orthogonal array in 125 runs is 5. Although it is conceivable that in computer experiments hundreds of runs might be achievable, for the problem at hand less than, or on the order of 100 runs of the computer code are acceptable. Additionally, often computer codes have at least tens of inputs and for the illustrations here no fewer than 7 inputs are considered.

As a result of these requirements, experiment plans formed by combining two or more designs, such as orthogonal arrays primarily with 2- and 3-level factors, creating multi-level arrays with subsets of different strength are proposed for computer experiments. Experiments constructed in this way will be referred to as combined array experiments, or combined arrays. Construction of combined arrays is illustrated in Section 2, specifically including investigation of 2-level and 3-level orthogonal arrays used to construct 5-level combined arrays. Additional analysis considerations, optimal experiment design properties and space-filling properties are discussed in Section 3 for combined arrays. Conclusions are in Section 4.

2. COMBINED ARRAYS

In the following, combined array experiments are constructed by combining 2- and 3-level fractional factorial experiments, or orthogonal arrays, creating 5-level arrays with subsets of different strength. The resulting combined array is not orthogonal although, obviously, subsets of runs are orthogonal arrays. While 5-levels are formed and the underlying arrays are orthogonal, clearly the concepts can be extended to form any number of levels for the factors and to combine arrays that are not orthogonal although the arrays should have some specified, desirable properties.

Factorial experiments are experiments for inputs, called factors, with a finite number of discrete values, referred to as levels, so if each input has K levels and there are p inputs then there are K^P possible distinct runs referred to as the K^P factorial design space. The K levels could be associated with K equal probability content intervals for a continuous input. If the experiment plan consisted of the entire K^P factorial design space, then for each pair of inputs (subsets of size 2) there are K^2 values (levels) with K^{P-2} "replicates" for each value. Obviously this extends to subsets of inputs of size s in the obvious way. For relatively moderate K and even small sizes for p the full product space of possible experiment runs quickly becomes unmanageably large, even given the ability to run the simulation code thousands of times. As stated previously, inputs with at least 5 levels are desired and only 5-level factors are considered in the following.

Orthogonal array experiment designs are subsets of full factorial designs, also referred to as fractional factorial designs, with reduced runs obtained by relaxing the property that for any subset of inputs there are "replicate" inputs for each value of the subset. Wu and Hamada (2000) and Hedayat, et al (1999) are good references on orthogonal arrays, in addition to several older texts on statistical experiment design and fractional factorial experiments by John (1971) and Raktoe, et al (1981). For K levels identified by elements in the set $L=\{0,1,2,\ldots,k-1\}$, an $N \times p$ array X with entries from L is an orthogonal array with K levels, strength t ($0 \le t \le p$) and index λ if every $N \times t$ sub-array of X contains each t-tuple based on L exactly λ times as a row. An array with parameters N (number of runs), p (number of factors), k (number of levels for each factor), and t (strength) is denoted OA(N,p,k,t). From this definition, a strength t orthogonal array with index λ is a set of p-dimensional factorial design points such that if one considers any t-dimensional projection then every point in the K^t factorial design space is replicated λ times. Likewise, any projection of dimension smaller than t, say $s<t$, consist of $\lambda * K^{(t-s)}$ replicates of the K^S factorial design space. A full K^P factorial design space is itself an OA(K^P,p,K,p) with index unity, that is $\lambda=1$. To reduce the number of runs from the full factorial design, a compromise is made on strength in orthogonal arrays. In a strict sense, fractional factorial designs may be any subset of the full factorial design space but often this terminology, or the term regular fractional factorial, is reserved for subsets that form an orthogonal array. For K prime, fractions of resolution III, IV and V defined in John (1971) or Raktoe, et al (1981) correspond to orthogonal arrays of strength 2, 3, and 4 respectively for which "replicate" runs occur for X_s including all values in the K^S grid, where $s \le t$ and, respectively, t=2, 3, and 4 is the strength of the array.

Again, experiment design options for 5-level factors are desired. The number of levels is required to be 5: a nominal value (coded as 2), two values either side of nominal (referred to as inner limits, coded as 1 and 3) but with values that might be reasonably expected, and two

values either side set a little further out (referred to as outer limits, coded as 0 and 4). The potential exists for failed runs at some of the extreme values. Less than 100 runs, or on the order of 100, could be done. Strength 3, at a minimum, is also desirable but that requires too many runs for a fully orthogonal array, on the order of 5^4=625 at a minimum, for 7 to 10 5-level factors. In reality, strength 3 is probably not absolutely required, that is the ability to assess a possibly unique effect for all 3 variable combinations of 5-level variables. Instead, this strength requirement reflects the experimenter's suspicion that there are potential interaction effects and the experimenter's desire to obtain some information about interactions from the experiment.

To obtain 5-level factors, 2-level and 3-level experiments designs are combined associating the levels of these two designs with 5-levels. The 2-levels are assigned the reasonable values either side of nominal (inner limits) and 3-levels assigned to nominal and the two extreme values (outer limits) either side of nominal. With this construct in mind, it is clear all that is required are desirable (high strength, allowing for run size limitations) 2-level and 3-level experiment designs. It is expected that a good (high strength) 2-level factorial design would yield main effects assessments independent of (at least pair-wise) cross factor interactions while a riskier (lower strength) 3-level factorial design would give somewhat more limited information on code functioning at nominal and extreme values of the factors. One would not run as much risk of losing information if code runs at extreme values fail since results on a good 2-level design would be obtained. However, there is potential for additional information over the limited 2-level factorial experiment, such as departure from linearity assessable with runs at the nominal values of factors as well as code performance at extremes. In the following, combined arrays are denoted CA(N,p,k,"i"t,"o"t) with parameters N (number of runs), p (number of factors), k (number of levels for each factor, here k=5), strength t labeled "i"t corresponding to the orthogonal array associated with the inner limits , and strength t labeled "o"t corresponding to the orthogonal array associated with the outer limits.

Substantial research and continuing development exists for constructing 2- and 3-level fractional factorial designs and the variety of methods and results in the literature are not surveyed here. Specific arrays are used to illustrate the construction of combined arrays. Hedayat, et al (1999) is a source of most constructions of these designs, and Tables 12.6 (c-e) on pages 326-327 of this text index constructed (fixed-level) orthogonal arrays for 2-level arrays with strength at least 3 and 3-level arrays with strength at least 2. Electronic data-bases containing these, and other, arrays can be found at the website:

www.research.att.com/~njas/oadir.

For seven factors, there is an OA(16,8,2,3), a 16-run orthogonal array for eight 2-level factors that is strength 3, and an OA(18,7,3,2), an 18 run orthogonal array for seven 3-level factors. Using only 7 of the 8 factors from an OA(16,8,2,3) combined with the OA(18,7,3,2), a CA(34,7,5,i3,o2) combined array is constructed. An OA(16,8,2,3), 2-level array is defined by columns $x_1, \ldots, x_8$ such that the first four columns are the full 2^4 array and the remaining columns are defined by the following equations (with modulus 2 addition):

$x_5 = x_1 + x_2 + x_3,$

$x_6 = x_1 + x_2 + x_4,$

$x_7 = x_1 + x_3 + x_4$,

$x_8 = x_2 + x_3 + x_4$.

Table 1 lists the 16 design points in this OA(16,8,2,3) with levels coded as 0 and 1 and then recoded to the values either side of nominal (inner limits) coded as 1 and 3 for the 5-level factors denoted $f_1, \ldots, f_8$:

Table 1: OA(16,8,2,3) and associated points in CA(34,7,5,i3,o2)

OA(16,8,2,3) Run/Input Coded {0,1}	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	CA(34,7,5,i3,o2) Run/Input Coded {1,3}	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
2	1	0	0	0	1	1	1	0	2	3	1	1	1	3	3	3	1
3	0	1	0	0	1	1	0	1	3	1	3	1	1	3	3	1	3
4	1	1	0	0	0	0	1	1	4	3	3	1	1	1	1	3	3
5	0	0	1	0	1	0	1	1	5	1	1	3	1	3	1	3	3
6	1	0	1	0	0	1	0	1	6	3	1	3	1	1	3	1	3
7	0	1	1	0	0	1	1	0	7	1	3	3	1	1	3	3	1
8	1	1	1	0	1	0	0	0	8	3	3	3	1	3	1	1	1
9	0	0	0	1	0	1	1	1	9	1	1	1	3	1	3	3	3
10	1	0	0	1	1	0	0	1	10	3	1	1	3	3	1	1	3
11	0	1	0	1	1	0	1	0	11	1	3	1	3	3	1	3	1
12	1	1	0	1	0	1	0	0	12	3	3	1	3	1	3	1	1
13	0	0	1	1	1	1	0	0	13	1	1	3	3	3	3	1	1
14	1	0	1	1	0	0	1	0	14	3	1	3	3	1	1	3	1
15	0	1	1	1	0	0	0	1	15	1	3	3	3	1	1	1	3
16	1	1	1	1	1	1	1	1	16	3	3	3	3	3	3	3	3

Hedayat, et al (1999) lists an OA(18,7,3,2) on page 20 and discusses construction in Chapter 3. The reader is referred to the text for construction and the design is listed here in Table 2 with standard {0,1,2} coding followed by coding for the nominal and extreme values (outer limits) for $f_1, \ldots, f_7$:

Table 2: OA(18,7,3,2)) and associated points in CA(34,7,5,i3,o2)

OA(18,7,3,2) Run/Input Coded{0,1,2}	x_1	x_2	x_3	x_4	x_5	x_6	x_7	CA(34,7,5,i3,o2) Run/Input Coded {0,2,4}	f_1	f_2	f_3	f_4	f_5	f_6	f_7
1	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0
2	1	1	1	1	1	1	0	18	2	2	2	2	2	2	2
3	2	2	2	2	2	2	0	19	4	4	4	4	4	4	4
4	0	0	1	2	1	2	0	20	0	0	2	4	2	4	0
5	1	1	2	0	2	0	0	21	2	2	4	0	4	0	0
6	2	2	0	1	0	1	0	22	4	4	0	2	0	2	0
7	0	1	0	2	2	1	1	23	0	2	0	4	4	2	2
8	1	2	1	0	0	2	1	24	2	4	2	0	0	4	2
9	2	0	2	1	1	0	1	25	4	0	4	2	2	0	2
10	0	2	2	0	1	1	1	26	0	4	4	0	2	2	2
11	1	0	0	1	2	2	1	27	2	0	0	2	4	4	2
12	2	1	1	2	0	0	1	28	4	2	2	4	0	0	2
13	0	1	2	1	0	2	2	29	0	2	4	2	0	4	4
14	1	2	0	2	1	0	2	30	2	4	0	4	2	0	4
15	2	0	1	0	2	1	2	31	4	0	2	0	4	2	4
16	0	2	1	1	2	0	2	32	0	4	2	2	4	0	4
17	1	0	2	2	0	1	2	33	2	0	4	4	0	2	4
18	2	1	0	0	1	2	2	34	4	2	0	0	2	4	4

Table 3 lists additional examples of combined arrays that could be formed in a like fashion to CA(34,7,5,i3,o2) based on arrays that are indexed in Hedayat, et al (1999).

Table 3: CA formed from binary and ternary OA

Binary OA	Ternary OA	CA
OA(16,8,2,3)	OA(18,7,3,2)	CA(34,7,5,i3,o2)
OA(24,12,2,3)	OA(27,13,3,2)	CA(51,12,5,i3,o2)
OA(32,16,2,3)	OA(27,13,3,2)	CA(59,13,5,i3,o2)
OA(64,14,2,3)	OA(27,13,3,2)	CA(91,13,5,i3,o2)
OA(64,14,2,3)	OA(54,25,3,2)	CA(118,14,5,i3,i2)
OA(128,15,2,4)	OA(54,25,3,2)	CA(182,15,5,i4,i2)

3. STATISTICAL ANALYSIS AND SPACE-FILLING FEATURES FOR COMBINED ARRAYS

Examining CA(34,7,5,i3,o2) in a similar way as an orthogonal array is evaluated, lower dimensional projections may be considered or, equivalently, multi-way tables of the counts of values of the factors that occur in the experiment design. For any two columns of the CA(34,7,5,i3,o2) experiment, the two-way table (Table 4) of values that occur in the design is:

Table 4 : Incidence of values for any two columns in CA(34,7,5,i3,o2)

"replicates"	fj=	0	1	2	3	4	totals
fi=							
0		2	0	2	0	2	6
1		0	4	0	4	0	8
2		2	0	2	0	2	6
3		0	4	0	4	0	8
4		2	0	2	0	2	6
Totals		6	8	6	8	6	34 runs

For a strength 2 orthogonal array this table would have the same values in every cell. For combined orthogonal arrays such as CA(34,7,5,i3,o2), there is a "checkerboard" pattern for the cells with non-zero and zero counts and the cells with non-zero counts may not have the same counts.

Considering any three factors in CA(34,7,5,i3,o2), the tables of values that occur are variants of one of the three tables labeled below as Table 5 for f_1, f_2, and f_3, Table 6 for f_1, f_2, and f_7, or Table 7 for f_3, f_5, and f_7. The variations that occur are that the rows that correspond to the even values of a factor may be permuted, although the marginal count values stay the same. There are 28 triples of factors, which have a 3-way table like Table 5, 6 triples correspond to Table 6, and factors f_3, f_5, and f_6 are the only ones with the pattern in Table 7.

Table 5: Values of f_1, f_2, and f_3 in the design CA(34,7,5,i3,o2).

"reps"	f_3=	0					1					2					3					4					total
	f_2=	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	
f_1=	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	6
	1	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	8
	2	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	6
	3	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	8
	4	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	6
Total		2	0	2	0	2	0	4	0	4	0	2	0	2	0	2	0	4	0	4	0	2	0	2	0	2	34

Table 6: Values of f_1, f_2, and f_7 in the design CA(34,7,5,i3,o2).

"reps"	f_7=	0					1					2					3					4					total
	f_2=	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	
f_1=	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	6
	1	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	8
	2	0	0	2	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	6
	3	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	8
	4	0	0	0	0	2	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	6
Total		2	0	2	0	2	0	4	0	4	0	2	0	2	0	2	0	4	0	4	0	2	0	2	0	2	34

Table 7: Values of f_3, f_5, and f_7 in the design CA(34,7,5,i3,o2).

"reps"	f_7=	0					1					2					3					4					total
	f_5=	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	
f_3=	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	6
	1	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	8
	2	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6
	3	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	8
	4	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	6
Total		2	0	2	0	2	0	4	0	4	0	2	0	2	0	2	0	4	0	4	0	2	0	2	0	2	34

A strength 3 orthogonal array would dictate that every cell in the 3-way tables has the same non-zero count. There are 125 cells, so, obviously, with only 34 runs not every cell can have a non-zero count. The trade-off with fewer runs than cells is to have non-zero count in as many cells as possible and have these cells "spread" around as much as possible. Visually, this is best achieved in Table 5 which is the associated table for 28 of the 35 possible triples of factors. Based on this observation, combined array designs do a good job of space-filling in lower dimensional projections that correspond to the strengths of the combined arrays. Specifically, CA(34,7,5,i3,o2) is a good space-filling design in its 2- and 3-dimensional projections.

Since the combined arrays have underlying structure of orthogonal arrays on subsets of runs, analyses investigating main effects and interactions are possible and there are "replicates" required for the comparison of R^2 as in McKay (1995) for identifying "important" input(s). In statistical experiments, 2- and 3-level experiments are common and relate to the fitting of polynomial regression models with degree 1 or 2, respectively. For 2-level factors, at most a first order, or linear, polynomial in a single factor can be modeled. For 3-level factors, a second order polynomial model can be fit. In the analysis of variance paradigm, 2-level factors allow fitting of linear main effects only while 3-level factors coincide with fitting linear and quadratic main effects. The requirement of strength 2 or 3 orthogonal arrays is

associated with fitting of polynomial regression models without or with cross factor terms, respectively. In an analysis of variance interpretation, strength 2 corresponds to the ability to fit main effects only where at least some main effects are biased by possibly significant two-factor interactions. In the experiment design literature this type of experiment is referred to as a resolution III design. Strength 3 corresponds to a resolution IV design where only a main effects model is estimable but the main effects estimates are not biased by any two-factor interactions, although bias due to any higher order interactions exists. Strength 4 corresponds to a resolution V design where main effects and two-factor interactions are estimable, although again biased by any potentially non-negligible higher order interactions. The capacity of an experiment to evaluate assorted polynomial trends does not necessarily indicate that the polynomial is in any sense the replacement model, but as for analysis based on comparison of R^2 for different sets of inputs, it provides a means for identifying inputs that are most influential subject to the limits of the experiment design.

4. CONCLUSIONS

Combining 2- and 3-level orthogonal arrays leads to designs with 5-level factors but with full orthogonality compromised. The resulting array is not orthogonal but high strength is achieved with respect to some level combinations or a subset of runs and as a result there is the capacity to make assessment of important effects based on comparison of R^2 for different input sets as in McKay (1995). These properties are achieved with fewer runs than would be required for an orthogonal design for 5-level factors.

REFERENCES

1. Hedayat, A. S., N. J. A. Sloane, and John Stufken (1999). **Orthogonal Arrays**. Springer-Verlag New York, Inc.

2. John, P. W. M. (1971). **Statistical Design and Analysis of Experiments**. The Macmillan Company, New York.

3. McKay, M. D., R. J. Beckman, and W. J. Conover (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, 21, No. 2, pp. 239-245.

4. McKay, M. D. (1995). Evaluating Prediction Uncertainty. NUREG/CR-6311, LA-12915-MS.

5. Moore, L. M., and M. D. McKay (2002). Orthogonal Arrays for Computer Experiments to Assess Important Inputs. LA-UR-02-1434. *Proceedings of PSAM6, 6th International Conference on Probabilistic Safety Assessment and Management*, (San Juan, Puerto Rico, USA, 23-28 June 2002).

6. Raktoe, B. L., A. Hedayat, and W. T. Federer (1981). **Factorial Designs**. John Wiley & Sons, Inc.

7. Wu, C. F. Jeff, and Michael Hamada (2000). **Experiments: Planning, Analysis, and Parameter Design Optimization**. John Wiley & Sons, Inc.

Halftoning and Quasi-Monte Carlo

Kenneth M. Hanson

Los Alamos National Laboratory, MS-D413, Los Alamos, New Mexico 87545
E-mail: kmh@lanl.gov

Abstract: The goal in Quasi-Monte Carlo (QMC) is to improve the accuracy of integrals estimated by the Monte Carlo technique through a suitable specification of the sample point set. Indeed, the errors from N samples typically drop as N^{-1} with QMC, which is much better than the $N^{-1/2}$ dependence obtained with Monte Carlo estimates based on random point sets. The heuristic reasoning behind selecting QMC point sets is similar to that in halftoning, that is, to spread the points out as evenly as possible, consistent with the desired point density. I will outline the parallels between QMC and halftoning, and describe an halftoning-inspired algorithm for generating a sample set with uniform density, which yields smaller integration errors than standard QMC algorithms in two dimensions.

Keywords: Quasi-Monte Carlo, Monte Carlo integration, low-discrepancy sequences, Halton sequence, Sobel sequence, halftoning, direct binary search, minimum visual discrepancy, Voronoi analysis

1. INTRODUCTION

The goal of the standard Monte Carlo (MC) technique[1] is to estimate the integral of a function over a specified M-dimensional domain from evaluations of the function at points that are randomly chosen within that domain. The objective in Quasi-Monte Carlo[2] (QMC) is to improve those estimates through a suitable specification of the sample point set. It has been shown that the errors from N samples for a fixed number of dimensions typically fall off as N^{-1} with QMC, much more quickly than with MC, namely, $N^{-1/2}$.

Digital halftoning is the process of creating a pattern of black dots on a white background to create the illusion of a gray-scale image.[3, 4] One of the principal goals in halftoning is avoid introducing undesirable texture into the rendered image, which is typically caused by clumping of the dots, or uneven dot placement that accompanies random dot distributions. In a sense, QMC has the same goal, whether it is implicitly or explicitly stated. The clumpiness in random point distributions also exists in standard Monte Carlo, and lead to lower sampling efficiency than more uniformly distributed point distributions. One observes that in regions of uniform low density, halftoned images seem to have characteristics deemed desirable in QMC.

The heuristic reasoning behind selecting QMC point sets is similar to that in halftoning, that is, to avoid clumping of the points (dots). The visual similarities between the patterns generated in halftoning and QMC lead one to speculate whether halftoning techniques might provide some useful lessons for quasi-Monte Carlo, or visa versa? I will

Author's web URL: http://public.lanl.gov/kmh/

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

demonstrate that a method for generating point sets, which is inspired by basic concepts used in halftoning, yields more accurate estimates of 2D integrals than is obtained with some standard QMC sequences. Of course, halftoning is conducted in only two dimensions. It must be kept in mind that the same algorithms that work for 2D may not work for higher dimensions. I will discuss the implications for higher dimensions and other potential approaches to enhanced QMC methods.

The objective of the present study is to find improved QMC techniques to reduce the number of function evaluations needed to achieve a specified accuracy in the estimate of an integral. For example, the function to be integrated may depend on a simulation of a complex physical process that might take several hours, or even several days, to calculate on the fastest computers available. In such a situation, the time required to generate an appropriate set of samples is inconsequential. Incidentally, intelligently selected point sets can be used for purposes other than integration, for example, for performing sensitivity analysis of computer models.[5, 6]

This paper summarizes the results presented in Ref. 7. Further details may be found there.

2. HALFTONING

The goal of the halftoning process is to render a gray-scale image, subject to whatever limitations are present in the printing or display process. Given the wide variety of constraints in printing technologies, a similarly wide variety of halftoning techniques exist[4]. For the purpose of the present study, I will focus on a single type of rendering, referred to as digital halftoning, in which the printing process is only capable of putting black dots on a white surface.[3, 8] The constraints on this process can include the dot size, the minimum distance between dots, etc. Because halftoning is used in commercial printers, which need to print pages rapidly, a large fraction of the published work on halftoning is devoted to finding ways to speed up the halftoning process, through use of look-up tables, for example. The trade-off between speed and rendition quality becomes a critical design issue.

Figure 1 shows an excellent example of a high-quality halftoned image.[9] This figure was produced using the direct-binary-search (DBS) technique, which will be described in the following section.[8, 9] As in many halftoning techniques, DBS is based on minimizing the perceived difference between the halftone image and the original gray-scale image that it is supposed to represent. Because the judge of the quality of the halftoned image is a human observer, halftoning algorithms are often are based on properties of the human visual system (HVS).

In one simple description of the HVS[3, 4], it is assumed that the effective modulation transfer function (MTF) for the eye is proportional to an exponential of $-c\,|f|$, where f is the radial spatial frequency on the observed page, and the factor c is related to the distance of the observer from the page. The 2D inverse Fourier transform of this MTF yields a blur function of the form

$$h(r) \propto (w^2 + r^2)^{-3/2}\,, \tag{1}$$

Figure 1. An example of a digital halftone image generated with the direct-binary-search algorithm taken from Ref. 9, which shows the high quality that is achievable with state-of-the-art halftone rendering of gray-scale images. (©IEEE)

where r is the radial distance in the printed image. The width of this 2D Cauchy distribution is characterized by the parameter w. The full-width at half maximum (FWHM) of the radial profile of this distribution is $1.553\,w$. An important feature of $h(r)$ is that it has a long tail, behaving like r^{-3} for large r.

Assuming a position-invariant blur function $h(x, y)$, the difference between the two perceived images is the convolution:

$$e(x, y) = h * [d - g] = \int h(x - x', y - y')\,[d(x', y') - g(x', y')]\,dx'\,dy'\,, \tag{2}$$

where $*$ denotes the convolution operation, $d(x, y)$ is the dot image, and $g(x, y)$ is the original gray-scale image to be rendered. Because the convolution redistributes intensities, it is necessary to specify what to use outside of the domains of images d and g. This topic is not often mentioned in discussions of halftoning but will be dealt with in Sect. 4. To quantify the perceived discrepancy between the halftone image and the actual gray-scale image, the most-often-used cost function is the total power in the error image

$$\varphi = \int_R |e(x, y)|^2\,dx\,dy\,. \tag{3}$$

A variety of simplifying assumptions go into this formula, but it seems to be adequate for producing halftone images of high visual quality.

The DBS algorithm[8, 9] is a specific approach to minimizing the HVS-motivated cost function given by Eq. (3). The following is a simplified description of the DBS algorithm. It is assumed that the halftone dot pattern is represented in terms of a discretized image in which each pixel represents a dot and has one of two values, either black or white. An initial pattern of dots is produced in which the density of the dots is approximately proportional to the gray-scale image being rendered. Any one of several methods may be used to generate this initial image, for example, thresholding of a set of random numbers.

In this iterative algorithm, each pixel in the image is considered one at a time. The change in φ produced by swapping the pixel's value with each of its eight nearest neighbors is calculated. The effect on φ of toggling, or reversing, the pixel's value is also calculated. If any of these options results in a decrease in φ, the change that reduces φ the most is kept. One pass through all pixels in the halftone image is counted as one iteration. The number iterations can vary from a half dozen to many times that, depending on the characteristics of the initial image and the stopping criterion. Although the DBS algorithm yields halftoned images of excellent quality, it requires intensive calculation.[9]

3. QUASI-MONTE CARLO

In standard Monte Carlo techniques[1], one evaluates integrals on the basis of a set of point samples. The integral of a function $f()$ of the parameter vector $\boldsymbol{x}$ is estimated as

$$\int_R f(\boldsymbol{x})\, d\boldsymbol{x} = \frac{V_R}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i)\,, \tag{4}$$

where R indicates the domain of integration in M dimensions, V_R is the volume of R, and the N samples $\boldsymbol{x}_i$ are randomly drawn from a uniform probability density function defined over R.

The objective of the quasi-Monte Carlo technique is to reduce the number of function evaluations needed to obtain a given accuracy in a Monte Carlo type of integration, and to accelerate its convergence as N increases, a goal that is typically achieved[10]. One useful feature of QMC is that any number of samples can be generated. Furthermore, an arbitrary number of additional samples can be added to an existing set of samples. The subject of space-filling or uniform point distributions has been extensively studied in the field of statistics [11, 12].

Figure 2 shows four different sets of points that cover the unit square in 2D. Panel (a) shows a set of random numbers that represents the type of point distribution that would be used in classical MC. The points in panel (c) are taken from the low-density portion of the sky near the top of Fig. 1. This dot pattern is observed to be somewhat similar to the patterns seen in 2D point sets that used in quasi-Monte Carlo. It is known[13] that by using quadrature methods, which typically rely on uniformly spaced sampling on a Cartesian square grid, shown in panel (d), the rms error drops as N^{-1}. For classical MC, the rms error drops more slowly, as $N^{-1/2}$.

It is interesting to compare the four patterns in Fig. 2, and consider the fact that the accuracy for integrating func2, defined in Eq. (8), using these point sets is (a) 2.5%, (b) 0.5%, (c) 0.14%*, and (d) 0.09%. It seems that the more uniformly distributed the points, the better they are for MC-style integration.

With most standard QMC sequences, the point sequences are the same from one time to the next. The algorithms for generating the points are deterministic, and typically

*This rms error is actually for the point set produced by the MVD algorithm, and shown in Fig. 4a, which is virtually indistinguishable in its general character from Fig. 2c.

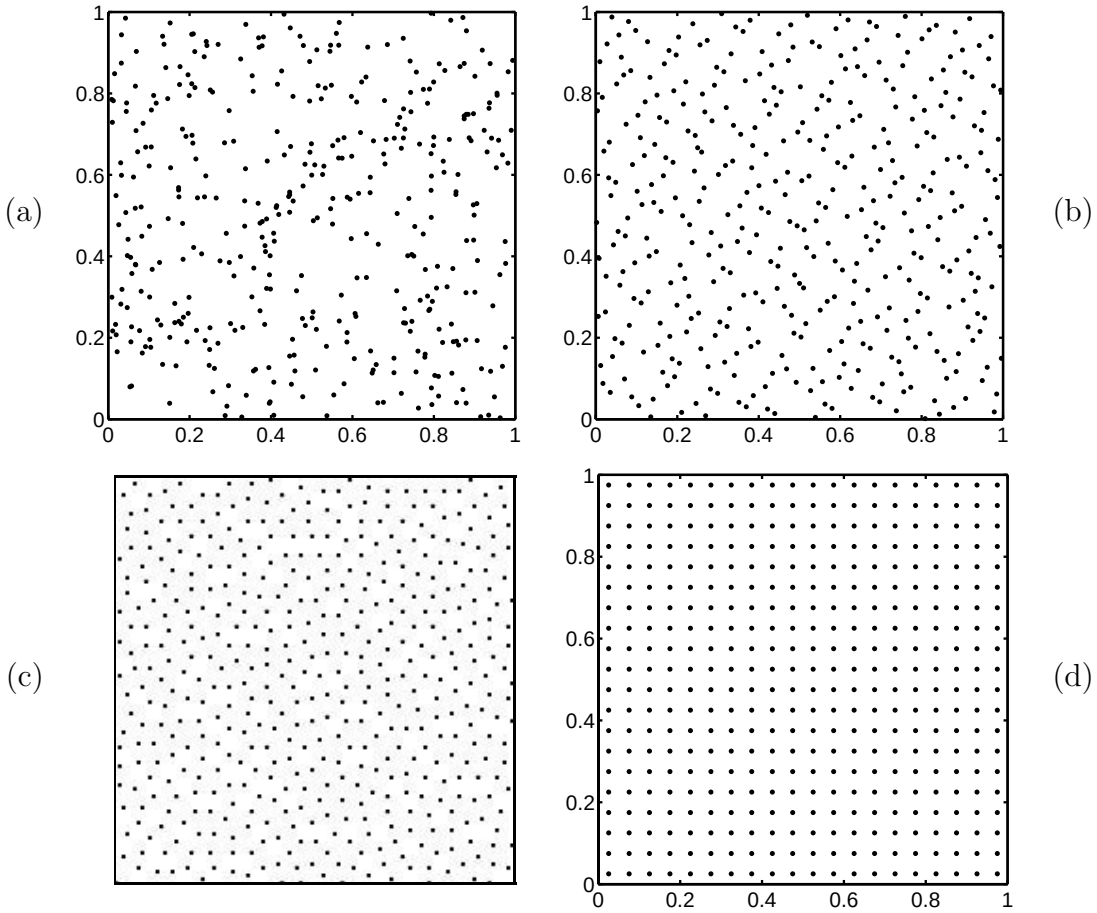


Figure 2. Four examples of different kinds of sample sets, each consisting of 400 points, which can be used to estimate an integral over the unit square using Eq. (4). (a) A random point set, each point in which is obtained by randomly drawing values for x and y from a uniform distribution from 0 to 1. (b) The first 400 points from the Halton sequence [2,3]. (c) Subsample taken from the sky region in the halftoning example shown in Fig. 1. (d) A regular array of points arranged on a square grid.

depend on the prime numbers, which are typically small primes. This situation is quite different than that for the generation of pseudo-random numbers, where the sequence is usually different (and independent) each time it is requested. Of course, by setting the 'seed' to the same value, the same sequence can be obtained again, but this is under the user's control.

A metric that is often used in the QMC field is the local discrepancy, the L_2 norm of which is, in 2D

$$D_2 = \left\{ \int_U \left[\frac{n(x,y)}{N} - A_0(x,y) \right]^2 dx\, dy \right\}^{\frac{1}{2}} , \tag{5}$$

where $A_0(x,y)$ is the area of the rectangle with one corner anchored at the origin and its opposite corner at (x,y), and $n(x,y)$ is the number of samples that lie inside the same rectangle out of a total of N samples. The integration is over the unit square, designated

by U. This quantity is clearly a measure of the uniformity of the distribution of the sample points. D_2, and other similarly defined discrepancy measures, are useful because they form the basis for several upper bounds on the integration error for specified classes of functions[2]. The definition of D_2 may be made more robust by employing an average over rectangles anchored at all four corners of the unit square[14]. One thing that seems odd about this metric is that it is based on rectangles with horizontal and vertical edges, which seems to emphasize the separable characteristics of point sets and may be not well suited for nonseparable functions. A metric with more isotropic properties might be more appropriate.

QMC sequences are typically obtained by combining low-discrepancy one-dimensional sequences[2]. Numerous QMC sequences have been proposed. One of the simplest is the Halton sequence, which is based on the prime numbers 2 and 3. For example, the Halton sequence based on the prime number 3 is 1/3, 2/3, 1/9, 4/9 7/9, 2/9, 5/9, 8/9, 1/27, 10/27, 19/27 etc. Other primes would yield different patterns. However, when the same primes are used, the sequences are perfectly repeatable and the same 2D pattern are obtained. The Sobel sequence is based on primitive polynomials.[13]

One argument for using MC instead of sampling on a regular grid (quadrature method) is to avoid a coherent interaction between the sample set and the unknown function being integrated.[13] Consider for example, the case when the integrand is a periodic function with the same period as the spacing between sample points. Then the estimated value of the integral depends critically on the position of the sample points relative to the structure in the integrand. A regular point pattern will yield the same result every time, if the same pattern is used. A random sample pattern used in Monte Carlo will yield different results every time, so one is less likely to be fooled. Another well-known reason for using MC in high dimensions is that the number of points required by quadrature methods becomes too large.[13]

Although there are similarities between QMC and halftoning patterns, there seems to have been little intersection of the two fields of research. One possible exception is the work by Ohbuchi and Aono[15] in which they used QMC sequences to improve rendering of surfaces and shadows in a 3D scene. It should be mentioned that many techniques other than QMC and MC exist for conducting efficient sensitivity studies on computer models, the intended application of the present work. Some of these are Latin hypercube sampling[11], stratified sampling,[1] and orthogonal arrays[6].

4. MINIMUM VISUAL DISCREPANCY FOR POINT SETS

Taking a cue from the DBS algorithm, I now describe an algorithm for minimizing the visual discrepancy (MVD) between a set of points and an image with uniform density. Starting with some arbitrary point pattern, the MVD algorithm considers each point in the set in a randomly permuted order. Instead of comparing the blurred dot image to the blurred gray-scale image, as in Eq. (3), it is better to reference the blurred dot image to its own mean value by computing its variance:

$$\psi = \int_R |e(x,y)|^2 \, dx \, dy \; - \left(\int_R e(x,y) \, dx \, dy \right)^2 , \tag{6}$$

where e is the convolution of d with h, as in Eq. (2), but with $g = 0$. This self referencing is possible because the number of points is fixed. To minimize ψ, the following iterative algorithm is employed. Each point is considered, one at a time, in a randomly ordered way. The value of ψ is calculated for fixed-length steps in eight different directions, along the axes and along the 45° diagonals. If any of these steps results in a decrease in ψ, that step yielding the smallest ψ is accepted. After all points in the point set have been considered, which is called one iteration, the step length is decreased by a factor and the process is repeated until a predetermined number of iterations K have been performed. The total number of iterations required to achieve satisfactory point distributions ranges from five to about 15.

The step length for the first iteration is chosen on the basis of how evenly distributed the points are in the initial pattern. It is specified in terms a characteristic distance between points for the final, uniformly distributed point set, taken in two dimensions as $a = \sqrt{N}$, for N points. In the present study, the initial step length is chosen between $0.2\,a$ to $0.4\,a$, dependent on the uniformity of the initial point pattern. The larger value is used when the initial pattern is a random point set, the smaller when starting with the Halton sequence, for example. In practice, the step length is adjusted so that approximately half the points considered get moved to lower ψ values. In the last iteration of the prescribed number of iterations K, the final step length is normally about $0.1\,a$.

An important aspect of the MVD algorithm is the specification and calculation of ψ, the function to be minimized. Through experimentation, I have found that the radial FWHM of the blur function Eq. (1) is best chosen to be on the order of the expected distance between points a. A value of $w \approx a/2$ seems to work well. Larger values do not keep points away from each other strongly enough. Another detail in the calculation of ψ concerns the convolution. The convolution operation redistributes intensities from inside the unit square to the outside, and visa versa. It therefore becomes necessary to specify the values of the image, not only inside the unit square, but also outside. In this study, the image is extended with a value equal to the average point density inside the unit square. Auxiliary runs show that this choice minimizes the error in the integration tests described in the next section. The convolution operations are performed using standard Fast Fourier Transform methods. The image sizes used in this study are typically 200 × 200 pixels to cover the unit square, and are extended to 256 × 256 pixels for the FFT.

The result of the blurring operation of the point (dot) image, stated in Eq. (6), is shown in Fig. 3. This point distribution is obtained by using the above optimization algorithm starting with the 100-point Halton sequence [2,3]. In this example, $a = 0.1$, and $w = 0.05$, which results in a radial profile with a FWHM of 0.0767. I have found that it is helpful to start with a point pattern that has relatively good coarse-scale uniformity. For example, starting with a Halton sequence of points or a stratified sample set[13] makes it easier to minimize ψ compared to starting with a random distribution. Similar observations have been made about the operation of centroidal Voronoi tessellation algorithms[16]. Other examples of point sets obtained with MVD are presented in Fig. 4. Figure 4a looks very similar to Fig. 2c, the DBS example.

Minimizing ψ amounts to minimizing the variance in Fig. 3b. It is fairly clear that to obtain the lowest variance in the blurred image, the points must be distributed so no

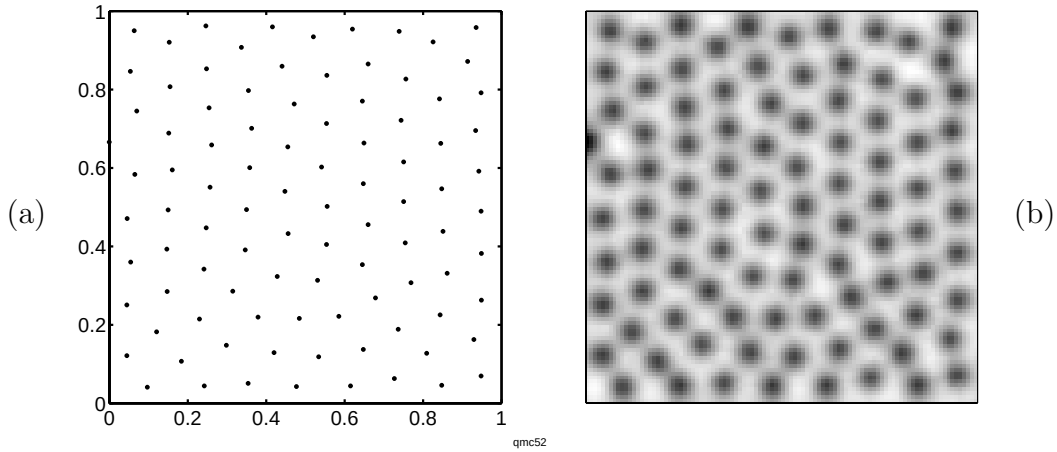


Figure 3. (a) This distribution of 100 points is obtained by minimizing the visual discrepancy from a uniform density level, and (b) the visualized point image, that is, the point image on the left blurred by the human visual system. The MVD algorithm is based on minimizing the variance in this blurred image. The rms deviation for the blurred image, relative to its mean value, is 1.1%.

gaps exist in the pattern. Consequently, for the optimal pattern the points will be spread out as far apart and as uniformly as possible. One might expect regular patterns to minimize ψ. In fact, it is possible to identify regular patterns in the point sets produced by the MVD algorithm, as for example, in Figs. 3a and 4. The local patterns tend to be principally hexagonal, but do not possess a long-range order. The situation is similar to that in solid-state physics in which carefully fabricated materials produce single crystals with well-defined lattice structures. While the MVD and the DBS halftoning algorithm do not produce such regular arrays, there is definitely a tendency to do so. The MVD structures have defects similar to those seen in crystals, e.g., dislocations and interstitial defects. Actually, the patterns produced by the MVD and BDS algorithms more closely resemble glassy structures, obtained by quickly freezing a molten material, than crystalline structures.

It should be kept in mind that the MVD algorithm tends to find a local minimum in ψ, not the global minimum. Single crystals represent the global minimum in energy. It may be possible to show that similar regular structures occur at the global minimum in ψ. Auxiliary tests indicate that regular square and hexagonal arrays of points tend to have the lowest observed values of ψ. However, initializing MVD with slightly perturbed versions these patterns does not generally result in the same regular pattern being regained. It seems that irregularities in the point distribution create barriers that the simple downhill optimization algorithm presently used does not overcome. Other choices for the blur kernel or smaller values for w may make it easier to come closer to the global minimum. On the other hand, it is not clear that reaching the global minimum is desirable, as commented on in the Discussion section.

The optimization approach in MVD moves each point a limited distance. As with many iterative approaches to optimization, the high-frequencies in the image (corresponding to

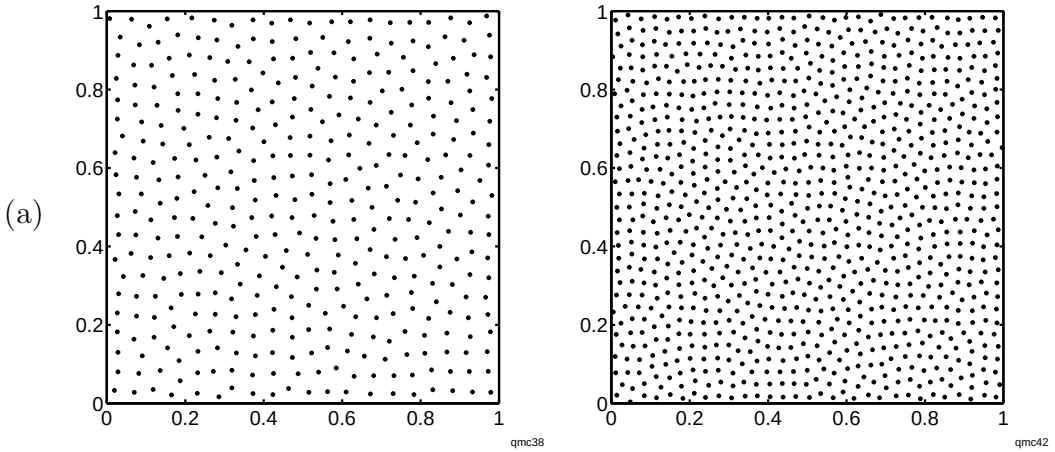


Figure 4. Point sets obtained with the MVD algorithm containing (a) 400 points and (b) 1000 points.

small scales) tend to be optimized at the beginning of the optimization procedure. The low frequencies (coarse scales) often take many iterations to be optimized. Thus, one might think that a multiscale approach to this optimization procedure could help speed up long-range adjustment of the point patterns to reach a suitable local minimum. For example, in the early iterations, one might use a relatively large value of w in the blur function (1) and move groups of neighboring points instead single points. An alternative would be to start with point patterns that possess coarse-scale uniformity, for example, a QMC sequence or stratified sampling. Even a regular pattern, such as the square-grid pattern shown in Fig. 2d or a hexagonal pattern might be used to initialize MVD, provided that the positions of each point is randomly budged a suitable, but small, distance.

5. INTEGRATION TESTS

The performance of the various kinds of points sets for MC-like integration may be tested for selected functions. The following two separable functions are used in this study

$$\text{func1}(\boldsymbol{x}) = \prod_i |4(x_i - x_i^0)|\,, \tag{7}$$

$$\text{func2}(\boldsymbol{x}) = \prod_i \exp\{-2|x_i - x_i^0|\}\,, \tag{8}$$

where $\boldsymbol{x}$ represents a position vector, $\boldsymbol{x}^0$ is the center of the distribution, and $i = 1$, 2 for two dimensions. The mean-square error is determined for the integral estimated using Eq. (4) by averaging the squared error over $\boldsymbol{x}^0$ values that uniformly cover the unit square. This average is accomplished using the standard MC technique and is based on 1000 randomly chosen $\boldsymbol{x}^0$ positions. The first function (7) is an inverted pyramid that rises linearly from zero at $\boldsymbol{x}^0$. The second function (8) peaks at $\boldsymbol{x}^0$, and gradually drops toward zero with a FWHM of 0.693 in both the x_1 and x_2 directions.

Figure 5 summarizes the results of the integration test of the function func2 (8). The first and most important observation is that the random point set yields substantially

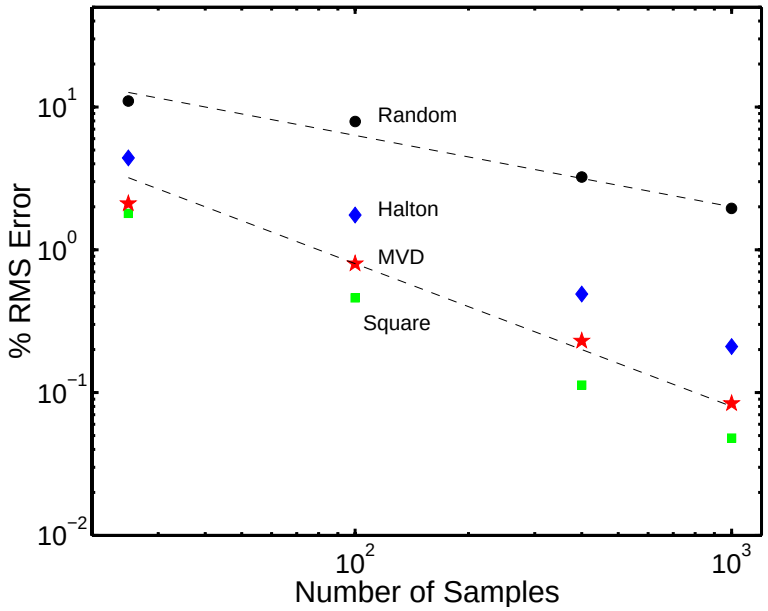


Figure 5. Plot of the per-cent rms error in Monte Carlo evaluations of integrals of the two-dimensional function func2, given in Eq. (8), for various kinds point sets versus the number of samples, N. The largest errors occur for the random sample set, shown as dots. The rms errors are smaller for the Halton sequence (diamonds), the minimum visual discrepancy (MVD) algorithm (stars), and the square grid (squares). The top line shows the convergence behavior of the rms error expected for standard MC, namely $N^{-1/2}$; the bottom line shows the behavior often achieved by quasi-MC sample sets, N^{-1}.

poorer integration accuracy than the other point sets. Its rms error drops like $N^{-1/2}$, indicated by the upper line, as is expected for Monte Carlo integration. The Halton sequence, a mainstay in the QMC field, performs much better. The MVD algorithm yields rms errors that are generally better than those for the Halton sequence by at least a factor of two. For 100 points, MVD provides approximately ten times lower error than random sampling, and at 1000 samples it is at least 20 times better. The general dependence of the error for these three kinds of points sets is N^{-1}. For expensive function evaluations, these improvements represent substantial savings.

The discrepancy D_2, given in Eq. (5), does not correlate very well with the integration accuracy. In particular, the values for D_2 for the square pattern are always somewhat greater than those of the Halton and MVD sample sets, and yet the square pattern almost always produced the smallest rms error. Furthermore, D_2 is often smaller for the Halton sequence than for the MVD pattern, even though MVD always has smaller rms error. One can conclude that the D_2 discrepancy for a point set is not a very good indicator of its integration accuracy, at least for the functions considered here.

One way to look at Eq. (4) for MC integration is that each function value in the sum is representative of a volume element surrounding it. A useful way to partition the domain of integration into neighborhoods surrounding each sample point is through Voronoi analysis.[16–18] A simple, if somewhat unorthodox way to perform Voronoi analysis, is based on a Monte Carlo technique. One randomly draws a set of K points $\{\boldsymbol{X}_k\}$ from

a uniform distribution defined over the domain of interest. The distance is computed between each point $\boldsymbol{X}_k$ and each point in the sample set $\{\boldsymbol{Z}_j\}$ being analyzed. It is assigned to the closest $\boldsymbol{Z}_j$. The set of points that are closest to $\boldsymbol{Z}_j$ belong to its Voronoi region, the number of which, divided by K, is an estimate of the fractional volume of that Voronoi region.

From the equal weighting of the function value at each sample point in Eq. (4), one would suppose that for a good sample set, the Voronoi areas should all be about equal. However, in further tests involving other types of point sets and other integration test functions, the correlation between these two quantities is not upheld. Thus, the uniformity of the Voronoi areas does not not seem to be uniquely related to the integration error. In addition, one might expect that the average second moment of the radius should be minimum for a point set with good sampling properties. This condition is attained with a centroidal Voronoi tessellation.[16, 17] Unfortunately, this quantity does not seem to be correlated with integration error either.

6. DISCUSSION

The ultimate aim of the present work is to improve on present techniques for analyzing the sensitivity of computer models outputs to numerous input parameters. The goal therefore is to develop point sets in high dimensions, which possess desirable convergence properties for integration estimates, comparable to, or better than QMC sequences. As argued earlier, it seems desirable for the sample points to be spatially distributed somewhat randomly. The sample sets should also be independent from one sequence to the next. Another desirable trait is for the sample sets to be easily augmentable, so that additional points may be easily generated, if deemed necessary.

The MVD algorithm, based on the ideas behind halftoning has been shown to achieve very good performance for integrating two simple functions. On the other hand, the MVD algorithm, per se, may be impossible to implement in high dimensions because it is based on determining the mean-square error, Eq. (6), of an M-dimensional image. This implies the necessity for storing a discretized image in M dimensions, which may be infeasible when M gets larger than four or five, even when coarsely discretized in each dimension. The convolution is not a problem, because with fast Fourier transforms, the cost grows only linearly with M.

Another approach to generating a suitable point set, which is closely related to MVD, is to draw an analogy between the point set and a collection of particles, which interact by means of a potential field. The potential field can be chosen so that the particles repel each other at close distances, but are less repulsive when they are sufficiently far apart. This type of action occurs in the MVD approach, although it is not explicit. Appropriate conditions need to be specified at the boundary of the region. The advantage of this potential-field approach is that an integral over the M-dimensional domain is not required to evaluate the cost function. To calculate the total energy of a specific configuration of the particles, the distances between each point and every other point needs to be calculated, an order-MN^2 calculation for N points in M dimensions, which is quite managable. On the

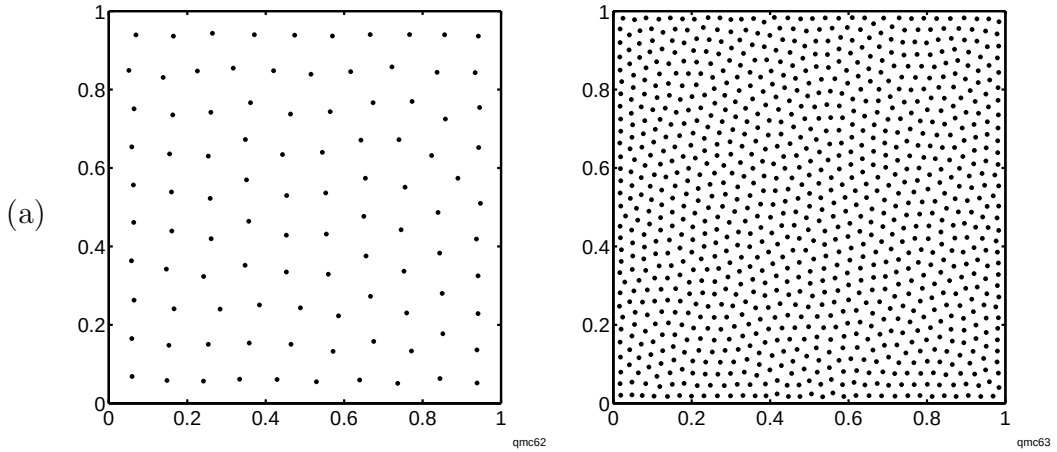


Figure 6. Examples obtained with the potential-field algorithm for (a) 100 points and (b) 1000 points.

other hand, to determine the change in energy for optimizing each point, the calculation is only of order MN.

Initial results indicate that this potential-field approach is promising. Figure 6 shows the results obtained with a potential with the same form used in the DBS algorithm, given in Eq. (1)

$$\zeta = \sum_{j=1}^{N} \sum_{i=j+1}^{N} (w^2 + |\boldsymbol{x}_i - \boldsymbol{x}_j|^2)^{-3/2}\,, \tag{9}$$

with suitable boundary conditions. The indices i and j label the points in the sequence. These patterns are visually very similar to those produced by MVD, shown in Fig. 3a and Fig. 4b. The rms errors for integrating func2 with these two point sets are 0.96% for 100 points and 0.17% for 1000 points, which is competitive with MVD. This potential-field approach is somewhat akin to the 'springs' idea of Atkins et al.[19]. Also, Idé et al. [20] have proposed a very similar approach for minimizing the L_2 discrepancy using an analogy to molecular dynamics.

ACKNOWLEDGMENTS

I appreciatively acknowledge the discussions I have had with Jan Allebach, Charlie Bouman, Tony Warnock, Paul Pederson, Luc Wille, Dave Robinson, Tony Giunta, Dave Higdon, Mike McKay, and Tom Booth.

REFERENCES

1. M. H. Kalos and P. A. Whitlock. *Monte Carlo Methods: Vol. 1, Basics*. John Wiley, New York, 1986.
2. H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia, 1992.
3. J. Sullivan, L. Ray, and R. Miller. Design of minimum visual modulation halftone patterns. *IEEE Trans. Sys., Man, Cyber.*, 21:33–38, 1991.

4. D. L. Lau and G. R. Arce. Digital halftoning. In S. K. Mitra and G. L. Sicuranza, editors, *Nonlinear Image Processing*, pages 375–401. Academic, San Diego, 2001.
5. M. D. McKay. Nonparameteric variance-based methods of assessing uncertainty importance. *Reliab. Eng. Syst. Safety*, 57:267–279, 1997.
6. A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity Analysis*. John Wiley, Chichester, 2000.
7. K. M. Hanson. Quasi-monte carlo: halftoning in high dimensions. In C. A. Bouman and R. L. Stevenson, editors, *Computational Imaging*, volume 5016 of *Proc. SPIE*, pages 161–172, 2003.
8. M. Analoui and J. P. Allebach. Model-based halftoning using direct binary search. In B. E. Rogowitz, editor, *Human Vision, Visual Processing, and Digital Display III*, volume 1666 of *Proc. SPIE*, pages 96–108, 1992.
9. P. Li and J. P. Allebach. Look-up-table-based halftoning algorithm. *IEEE Trans. Image Proc.*, 9:1593–1603, 2000.
10. H. Niederreiter. *Monte Carlo and Quasi-Monte Carlo Methods 1998*. Springer, Berlin, 1998.
11. M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods of selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245, 1979.
12. T.J. Santner, B. J. Williams, and W. I. Notz. *The Desigan and Analysis of Computer Experiments*. Springer, Berlin, 2003.
13. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C++*. Cambridge University, New York, 2002.
14. T. T. Warnock. Computational inverstigations of low-discrepancy point sets II. In H. Niederreiter, editor, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 354–361. Springer, New York, 1995.
15. R. Ohbuchi and M. Aono. Quasi-Monte Carlo rendering with adaptive sampling, 1996. On web: http://www.kki.yamanashi.ac.jp/∼ohbuchi/online_pubs/eg96_html/eg96.htm.
16. Q. Du, V. Faber, and M. Gunzburger. Centroidal Voronoi tesselations: applications and algorithms. *SIAM Rev.*, 41:637–676, 1999.
17. L. Ju, Q. Du, and M. Gunzburger. Probabilistic methods for centroidal Voronoi tesselations and their parallel implementations. *Parallel Comp.*, 28:1477–1500, 2002.
18. A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. *Spatial Tesselations: Concepts and Applications of Voronoi Diagrams*. John Wiley, Chichester, 2000.
19. C. B Atkins, J. P. Allebach, and C. A. Bouman. Halftone postprocessing for improved rendition of highlights and shadows. *J. Elec. Imaging*, 9:151–158, 2000.
20. T. Idé, H. Mizuta, H. Numata, Y. Tair, M. Suzuki M. Noguchi, and Y. Katsu. Dot pattern generation technique using molecular dynamics. *J. Opt. Soc. Amer.*, 20:248–255, 2003.

Model Calibration Methods for Phase Diagram Determination

Brian J. Reardon[a] *and Marius Stan*[b]

[a]Los Alamos National Laboratory, Engineering Sciences and Applications Division, Los Alamos, NM, USA
[b] Los Alamos National Laboratory Materials Science and Technology Division, Los Alamos, NM USA
E-mail: breardon@lanl.gov, mastan@lanl.gov

Abstract: A heuristic optimization methodology based on a Genetic Algorithm is presented with the goal to help researchers decide on the optimal set of thermodynamic data and models to use to accurately model phase diagrams and their associated uncertainty. This approach accounts for the errors associated with reported data and how reliable the researcher believes the model to be. Additionally, the results of the Genetic Algorithm provides guidance as to which experiments are needed to enhance the reliability of the dataset and is ideally suited for parameter optimization and sensitivity analysis. Applications include the UO_2-PuO_2 and UO_2-BeO systems.

Keywords: Genetic Algorithms, model calibration, uncertainty

1.0 INTRODUCTION

Finding an optimal model by fitting thermodynamic data is a difficult problem in materials science due to the large uncertainty associated with the experimental or calculated data sets that are used as input. This situation is most prevalent in the case of the calculation of phase diagrams [1] where the solidus and liquidus boundaries are highly uncertain [2] due to limitations in the accuracy of temperature measurements, limitations in determining the phase transition boundaries, and the potential for off-stoichiometric compositions at high temperature. In spite of all of these uncertainties or perhaps because of them, one rarely finds the uncertainty bounds reported with the phase diagram.

One could address this problem in several ways, each having a number of limitations. The first would be to simply accept a particular set of thermodynamic data as fact and use these values to calculate phase equilibrium curves. This approach ignores all the data available in other thermodynamic assessments as well as any phase diagram data. It also assumes that the selected data set is the best available. The second approach would be to take an average of all the known thermodynamic data sets. It assumes that all the thermodynamic data is equal in quality and thus only an average is necessary. This approach also ignores the available phase diagram data. Another option is to fit the solidus and liquidus equations to the known phase diagram data. The starting point of the optimization heavily influences this approach. Furthermore, it essentially ignores the experimental thermodynamic data once the optimization has initiated. The method proposed in this work uses a Genetic Algorithm to incorporate all the data and its associated uncertainties into an optimal fit of what is known.

The scientific literature is notably bereft of papers dedicated to the analysis of the uncertainties associated with equilibrium phase diagrams. A classical approach, based on the "spread of mistakes" formalism, is presented in [3] while in [4] a way of extracting the maximum information from a minimal set of experimental data is investigated. Bayesian based methods have been used to produce self-consistent thermodynamic data sets for binary

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

[5] or multi-component [6-8] systems. All papers emphasize the importance of starting with a reliable, self-consistent thermodynamic data set and draw the reader's attention to the propagation of the errors in the input parameters during the calculation process.

This work demonstrates that modern heuristic optimization techniques such as Genetic Algorithms offer a viable way of relating models to the data in the face of large uncertainties both on the model parameters and the training data. To illustrate the utility of this approach the solidus and liquidus boundaries of the UO_2-PuO_2 and UO_2-BeO systems have been selected. The parameter optimization was performed for each binary system given some information about the solidus and liquidus boundaries, the heats of melting, ΔH^M, and the melting temperatures, T^M.

1.1 Genetic Algorithms

Genetic Algorithms are heuristic optimization techniques that borrow heavily from the ideas of Darwinian evolution. Using evolution as an optimization tool was first proposed by Holland [9] and ever since has spurred a large amount of interest [10]. A genetic algorithm borrows the three main constructs of Darwinian evolution (selection, crossover, and mutation) to evolve a set of parameter vectors towards an optimal solution.

In the parlance of the GA community, the set of parameter vectors is a population. Each member of the population is evaluated to determine how well it solves the problem at hand, i.e. to determine their fitness. The most fit members are selected with a probability proportional to their fitness and allowed to exchange genetic information with other members and thus create the next population which usually has a higher average fitness than the parent population.

As a population based optimization method, the GA is ideally suited to handle the various forms of uncertainty found in this problem with a minimum of assumption about what the uncertainty should look like. The first form of uncertainty lays in the phase diagram data itself. Figure 1 shows the experimental solidus and liquidus data from three different sources. In this work the GA must fit a model through this data. The error bars on both the composition and temperature can be interpreted as nothing more than random intervals and thus there is no way to discern any type of uncertainty distribution on the intervals. Uncertainty in the experimental data relaxes the constraints on the optimal parameter values and forces us to identify a range of parameter values that provide a range of calculated values that lay within the experimental uncertainty. Thus, this problem is under specified. Previous work [11] has shown that a selection operator that uses a fuzzy logic-weighting scheme effectively handles optimization scenarios such as this where there are potentially a very large number of solutions all of equal fitness and plausibility.

A fuzzy logic-weighting scheme [12] looks at all the objective values of a particular member and rescales them to a value between 0 and 1. 0 if the value is the worst of the population. 1 if it falls within the experimental uncertainty. Once the objectives have been scaled, the average is taken over all objectives and that single number is the fitness for the member in the population.

Using the fuzzy logic weighting scheme, the GA is run until all the members of the population reach a fitness of 1 or at least reach a state of equilibrium where there is no more improvement. When this state is reached, the members of the final population are used to determine the uncertainty bounds on the model parameters. The population of final

parameters can then be used to bound output of the model and show where the model is most uncertain and in need of more data.

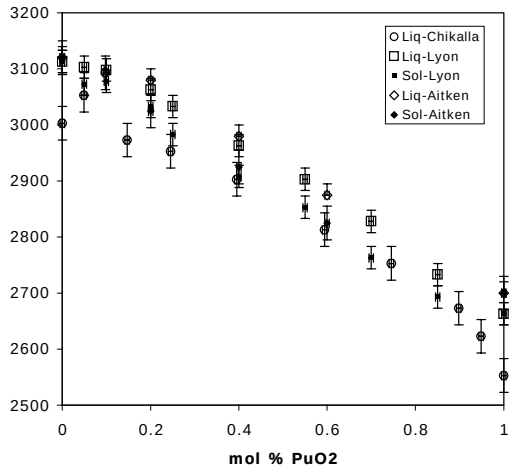


Figure 1. The experimental solidus and liquidus data for the UO_2-PuO_2 system from Chikalla, Lyon and Baily, and Aitken and Evans.

Another form of uncertainty lays in the search ranges for many of the input model parameters being optimized. For some models, the input parameters are in fact experimentally accessible values which themselves contain a large amount of uncertainty. Such parameters would be the melting points of the starting compositions (UO_2, PuO_2, BeO) as well as their heats of melting. Other models are purely empirical and were developed by various authors to fit their particular data sets. The empirical parameters have no physical meaning. If such a model is to be fit to a different set of data, the published parameters could be used as a starting point. Unfortunately, the parameters in question are rarely published with any sort of assessment of uncertainty and one must then make an educated guess as to what a reasonable search range would be.

There are a number of advantages to using a GA in this problem over other calibration approaches. First, a multi-objective GA tends to be robust enough to identify distributions of solutions. These distributions are often multi-modal and thus have shapes not easily captured by traditional calibration routines. Second, the GA requires no preconceived assumptions about the uncertainty distributions on the objective data or the parameter values. However, if desired, assumed distributions are easily incorporated. Third, and probably most importantly, the GA incorporates all the known data into its search. For example, the known phase diagram data defines the objectives and the spread of the known thermodynamic data defines the search space.

2. PROCEDURE

This work will address how well 5 different models fit and explain 3 different data sets and combinations thereof all of which are supposed to describe the solidus and liquidus curves of the UO_2-PuO_2 phase diagram as shown in Figure 1.

The first data set is that of Chikalla [13] which only shows a liquidus curve. The liquidus curve behaves as one would expect for an ideal solid solution except for the PuO_2 values of

of 5 and 10 wt%. The data also shows significant scatter in liquidus temperature at all compositions but especially at 75 wt% PuO_2. Note that there is no scatter displayed or reported in the composition (x axis). It was presumed by the authors that the composition was known exactly even though it is common knowledge that composition can drift due to changes in oxygen partial pressure and vaporization of the components. It is also worth nothing that Chikalla, admits later in his paper that the liquidus must be much lower in temperature throughout the entire PuO_2 composition range due to thermodynamic considerations and known melting points and heats of melting. However, other authors believe the liquidus to be much higher.

The second data set is from Lyon and Baily [14] and is generally considered to be a much more well-behaved and thus reliable experimental determination of both the solidus and liquidus of the UO_2-PuO_2 phase diagram. Lyon and Baily also compute the solidus and liquidus curves using the Ideal Solutions laws and generally show a much better fit than Chikalla did in his study. Like Chikalla, however, uncertainties were never properly accounted for in this study. For example, the composition is still assumed to be known exactly and in many cases, especially at very small and very large PuO_2 compositions, the temperature error bars of the solidus and curves overlap significantly. Both Chikalla and Lyon also use the generally accepted Ideal Solution model to calculate the solidus and liquidus curves. While this model is known to be quite accurate and extendable to other systems because it is based on first principles, it has a down fall when it comes to fitting phase diagram data. Namely, phase diagram data is usually collected by changing the composition (x) and measuring the temperature of the phase transitions (y) whereas the Ideal Solution model is the inverse. It assumes a temperature and calculates the composition of the solidus and liquidus. This inversion presents somewhat of a problem in the context of fitting a model to the data since the effect of experimental uncertainties cannot be directly propagated through the optimizer.

The third data set is that of Aitken and Evans [15, 16]. Aitken and Evans, like Lyon and Bailly, experimentally determine the solidus and liquidus of the UO_2-PuO_2 system by varying the composition and measuring the temperature at which the solidus and liquidus are observed. Aitken and Evans differ from Lyon and Baily and also Chikalla by attempting to fit the observed solidus and liquidus data with different polynomial forms. The advantage of this approach is that the polynomial forms, like the data, provide a measure of temperature as a function of composition and thus are more amenable to proper uncertainty propagation. The downside of this approach is obvious. Namely, the polynomials are only applicable to the phase diagrams at hand and their parameters cannot be used in a predictive fashion for other thermodynamic studies.

2.1 Model Descriptions

2.1.1 Model 1. Ideal Solid Solution Law

The UO_2-PuO_2 system shows complete solubility of the two components in the solid phase [17]. The liquidus (x^{Liq}) and solidus (x^{Sol}) mole fractions for each fixed temperature (T) can be approximated [18, 19] by:

$$x^{Liq}(T) = \frac{1 - \exp\left(\frac{\Delta H^{M}_{UO_2}}{R}\left(\frac{1}{T} - \frac{1}{T^{M}_{UO_2}}\right)\right)}{\exp\left(\frac{\Delta H^{M}_{PuO_2}}{R}\left(\frac{1}{T} - \frac{1}{T^{M}_{PuO_2}}\right)\right) - \exp\left(\frac{\Delta H^{M}_{UO_2}}{R}\left(\frac{1}{T} - \frac{1}{T^{M}_{UO_2}}\right)\right)} \quad (1)$$

and

$$x^{Sol}(T) = x^{Liq}(T) \cdot \exp\left(\frac{\Delta H^{M}_{PuO_2}}{R}\left(\frac{1}{T} - \frac{1}{T^{M}_{PuO_2}}\right)\right) \quad (2)$$

Where R is the gas constant (8.314 J/mol K).

The values of the input parameters for the UO_2-PuO_2 system are scattered [13, 15, 20-22], leading to large uncertainty bounds, as shown in Table I. The goal of our work is to further refine the values of the input parameters using the GA given known experimental data on the solidus and liquidus positions (Figure 1). Thus, the optimization proceeds as follows: First, the search range for each parameter is defined for the GA based on the accepted uncertainty or variation in the published parameter values. Second, the GA evolves the parameter values based on how well the values generate solidus and liquidus curves, which match the available experimental data – taking into account the uncertainties of the experimental solidus and liquidus curves. For this study the initial uncertainty in concentration was assumed to be 0.005 and the uncertainty in the liquidus and solidus temperatures was 55K and 35K respectively. Once the range of parameter values is optimized, values from this range are placed in the forward model. This results in fuzzy bands that define the position of the curves which are most self consistent given all data and the underlying model. Of particular importance to note is that the total uncertainty in model parameter values as well as the solidus and liquidus curves decreases by using this method approach.

Table I. The Upper and Lower Limits of the Variable Search Space [13, 15, 20-22].

Variable	Units	Lower value	Upper value
ΔH^{M}_{UO2}	kJ/mol	25	125
ΔH^{M}_{BeO}	kJ/mol	42	125
ΔH^{M}_{PuO2}	kJ/mol	25	100
T^{M}_{UO2}	K	3000	3200
T^{M}_{BeO}	K	2700	2900
T^{M}_{PuO2}	K	2600	2800

2.1.2 Model 2: polynomial in (x)

Adamson et al. (in Aitken) recommend the following model for the solidus and liquidus curves of UO_2-PuO_2

$$T_s(K) = a_s + b_s x + c_s x^2 + d_s x^3$$
$$T_l(K) = a_l + b_l x + c_l x^2 \quad (3)$$

where a_s = 3120, b_s=-655.3, c_s=336.4, d_s=-99.9 and a_l=3120, b_l=-388.1, c_l=-30.4. Note that the polynomial is in x as opposed to T as in Model 1.

2.1.3 Model 3; polynomial in (x)

Lyon and Baily recommend the following model for the solidus and liquidus curves of UO_2-PuO_2

$$T_s(K) = a_s + b_s x + c_s x^2$$
$$T_l(K) = a_l + b_l x + c_l x^2 \quad (4)$$

where a_s = 3113.15, b_s=-5.41395, c_s=7.4639e-3 and a_l=3113.15, b_l=-3.2166, c_l=-1.448518e-3.

2.1.4 Model 4: polynomial in (x)

Komatsu et al. (in Aitken) recommend the following model for the solidus and liquidus curves of UO_2-PuO_2

$$T_s(K) = T_{MUO_2} / \left(1 + b_s x + c_s x^2\right)$$
$$T_l(K) = T_{MUO_2} / \left(1 + b_l x + c_l x^2\right) \quad (5)$$

where b_s=0.1811, c_s=-0.011 and b_l=0.1068, c_l=0.06316.

2.1.5 Model 5: standard thermodynamic in T but extended to other phase diagram data

Another advantage of using a GA with a fuzzy logic selection method is that different types of objective functions can be easily combined. In this exercise the objective goals of Model 1 are combined with the objective goals and data of a thermodynamic model of the eutectic UO_2-BeO system. To optimize the UO_2-BeO system, a similar procedure as for the UO_2-PuO_2 system was employed. For this type of diagram the equilibrium lines are defined by:

$$x^{Liq}_{UO_2+Liq}(T) = 1 - \exp\left(\left(\frac{-\Delta H^M_{UO_2}}{RT}\right)\ln\left(\frac{T^M_{UO_2}}{T}\right)\right) \quad (6)$$

and

$$x^{Liq}_{Liq+BeO}(T) = \exp\left(\left(\frac{-\Delta H^M_{BeO}}{RT}\right)\ln\left(\frac{T^M_{BeO}}{T}\right)\right) \quad (7)$$

The values of the melting enthalpy and temperature were obtained from the literature and are displayed in Table I. Note that the both Model 1 and the eutectic model require thermodynamic values for UO_2. Thus, by incorporating this model into the optimization scheme, the potential thermodynamic values for UO_2 are constrained. Unfortunately, the thermodynamic values for BeO are also required. The reader should take heart, however, that this process of combining objectives, models and data allows one to obtain thermodynamically self consistent values for the basic properties of the constituent compositions – something that is typically very hard to do using other uncertainty propagation methods. The UO_2-BeO phase diagram [23] shows a eutectic point at T = 2450 K and BeO mole fraction x = 0.68. Namely x_{BeO} = 0.68±0.05 (Figure 2). For this study the uncertainty in

the liquidus concentration was again 0.05 and the temperature uncertainty was 40K. In the model, the eutectic composition is defined as that point in which the curves calculated from Eq. (6) and Eq. (7) intersect. This point also defines the calculated eutectic temperature. While the eutectic temperature is known experimentally, there is no information gain in comparing it to the calculated value since the calculated value is determined by the calculated value of the eutectic composition.

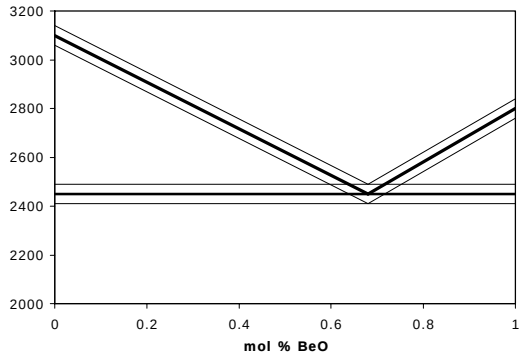


Figure 2. The experimental eutectic point for the UO_2-BeO system at 0.68 mol% BeO.

3. RESULTS

Table II shows the results of the optimization with each model described previously and when using different data sets. The table shows the model used, data set used, number of solutions found and the fitness of said solutions. The maximum possible fitness is 1.0 and would indicate that all the calculated values fall within the experimental uncertainty of the data.

3.1 Model 1

Tests 1-10 used model 1 with the listed data sets. The difference between tests 1-5 and 6-10 is the stated uncertainty of the composition values. In tests 1-5 the uncertainty is 0.005 whereas in tests 6-10 it is 0.05. Note that experimentally the uncertainty in composition is very small (0.005 is a reasonable number) whereas the uncertainty in temperature is much higher. Unfortunately, Model 1 is written as a function of temperature, not composition. Thus when the model optimized against the raw data, it rarely falls within the experimental uncertainty of x. The net result of this fact for tests 1-5 is that the apparent fitness of the optimal solutions is very low. To get around this problem and find a large set of solutions that actually pass through the known uncertainty bounds of the experimental data, the uncertainty in composition was expanded based on the degree to which the uncertainty in x would intersect uncertainties in temperature of neighboring compositions. The final assessment of this 'graphically driven' as opposed to data driven uncertainty was an interval of size ±0.05. Tests 6-10 show the results of using this 'graphically driven' uncertainty in x. Note that all of the fitnesses increase as would be expected but most notably, a total of 394 solutions were found that perfectly match the experimental data of Lyon and Baily.

These results indicate that Lyon and Baily's experimental data is most consistent with the Ideal Solution assumptions of Model 1. Further it correctly identifies Chikalla's data as being the most suspect.

Table II. The results of optimizing each model against the available data sets. C: Chikalla, L: Lyon and Baily, and A: Aitken and Evans.

Test	Model	Data Sets	# Solutions	Fitness
1	1a	L	1	0.949274
2	1a	A	1	0.976892
3	1a	C	1	0.84066
4	1a	L+A	1	0.928932
5	1a	L+A+C	1	0.790519
6	1b	L	394	1
7	1b	A	1	0.989953
8	1b	C	1	0.887064
9	1b	L+A	1	0.99024
10	1b	L+A+C	1	0.874315
11	2	L	5	0.998358
12	2	A	377	1
13	2	C	1	0.9591
14	2	L+A	14	0.995911
15	2	L+A+C	1	0.941159
16	3	L	145	0.998775
17	3	A	291	1
18	3	C	502	1
19	3	L+A	82	0.993154
20	3	L+A+C	2	0.930555
21	4	L	3	0.983283
22	4	A	1	0.993808
23	4	C	449	0.960122
24	4	L+A	1	0.982623
25	4	L+A+C	1	0.920988
26	5a	L	11	0.974211
27	5a	A	322	0.981022
28	5a	C	105	0.899736
29	5a	L+A	1	0.963978
30	5a	L+A+C	7	0.894648
31	5b	L	308	0.999815
32	5b	A	255	1
33	5b	C	395	0.933822
34	5b	L+A	43	0.995135
35	5b	L+A+C	127	0.930271

3.2 Models 2-4

Models 2, 3, and 4 are polynomial functions originally designed to fit specific data sets. From the results of Table 1I this fact is clear since some models find a large number of highly fit solutions for one set of data and not the others. It should also be pointed out that since the polynomials are functions of composition, they are able to more accurately account for the uncertainty in the temperature data. Thus, they give the illusion of being better fit models than the thermodynamically based Model 1. Unfortunately, though these models appear to fit

much of the data very well they are in no way extensible to other phase systems. In other words where the melting points and heats of melting optimized in Model 1 can then be used to estimate the behavior other phase systems, the parameters of Models 2-4 cannot.

3.3 Model 5

Like Model 1, two values were used for the uncertainties in composition. Namely, ±0.005 for Tests 26-30 and ±0.05 for Tests 31-35. Also, as was the case for Model 1, the larger uncertainties in composition faired better in the optimization. This is evidenced by the higher fitness values and number of solutions for all of the data sets optimized against. The most notable difference between the output of Models 1 and 5 is shown in Figure 3, however. Model 1 had to optimize the values for the melting point and heats of melting of UO_2 and PuO_2 whereas Model 5 also had to optimized the same values for BeO. Since the values for UO_2 were needed in both the Ideal solution model of UO_2-PuO_2 and the simple eutectic model of UO_2-BeO, the ultimate optimal values were much more constrained in Model 5. Figure 3 shows the optimal melting point and heat of formation for UO_2 from Models 1 (circle) and Model 5 (square). Note that since UO_2 was much more constrained in Model 5, only one viable solution was found.

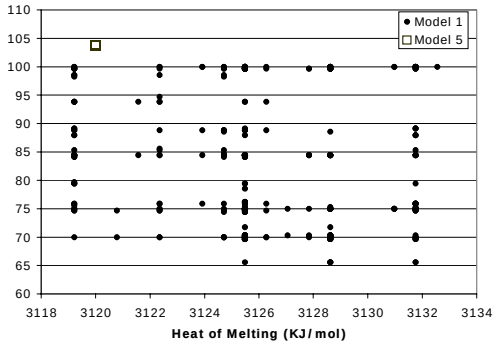


Figure 3. The final solution sets for the heats of melting and the melting points of UO_2 determined through the optimization of Model 1 (circle) and Model 5 (square).

4. CONCLUSIONS

The use of a Genetic Algorithm allows for incorporating uncertain data sets, both large and small, in an efficient and meaningful way. This process then leads to the optimization of the parameters of proposed models and the assessment of the overall predictive credibility of said models. Specifically, this work determines the degree of uncertainty on the phase boundaries of the UO_2-PuO_2 and UO_2-BeO systems by taking into account the available phase boundary data, the accepted models of the phase boundaries, and the thermodynamic data used in those models. The net result was an overall reduction in uncertainty of the values of the thermodynamic data as well as the phase boundary positions in a way that is internally self-consistent. The use of modern heuristic optimizers such as genetic algorithms was crucial to this work since they are both robust and require no assumptions about the forms of the uncertainty distributions.

5. REFERENCES

1. M. Stan, Phase Diagram Calculations in Materials Processing, *Control and Optimization in Minerals, Metals and Materials Processing*. 1999, Met. Soc.: Montreal.

2. M.I. Baskes, and M. Stan, An atomistic study of solid/liquid interfaces and phase equilibrium in binary systems. *Met. Trans. A*: Physical Metallurgy and Materials Science, 2003. **34**: p. 435-439.
3. D.V. Malakhov, Confidence intervals of calculated phase boundaries. *Calphad-Computer Coupling of Phase Diagrams and Thermochemistry*, 1997. **21**(3): p. 391-400.
4. E. Konigsberger, Improvement of Excess Parameters from Thermodynamic and Phase-Diagram Data by a Sequential Bayes Algorithm. *Calphad-Computer Coupling of Phase Diagrams and Thermochemistry*, 1991. **15**(1): p. 69-78.
5. J. Acker, and K. Bohmhammel, Optimization of thermodynamic data of the Ni-Si system. *Thermochimica Acta*, 1999. **337**(1-2): p. 187-193.
6. N.D. Chatterjee, et al., The Bayesian approach to an internally consistent thermodynamic database: theory, database, and generation of phase diagrams. *Contributions to Mineralogy and Petrology*, 1998. **133**(1-2): p. 149-168.
7. N.D. Chatterjee, K. Miller, and W. Olbricht, Bayes Estimation - a Novel-Approach to Derivation of Internally Consistent Thermodynamic Data for Minerals, Their Uncertainties, and Correlations .2. Application. *Physics and Chemistry of Minerals*, 1994. **21**(1-2): p. 50-62.
8. W. Olbricht, N.D. Chatterjee, and K. Miller, Bayes Estimation - a Novel-Approach to Derivation of Internally Consistent Thermodynamic Data for Minerals, Their Uncertainties, and Correlations .1. Theory. *Physics and Chemistry of Minerals*, 1994. **21**(1-2): p. 36-49.
9. J.H. Holland, Adaptation in Natural and Artificial systems. Complex adaptive systems. 1992, Cambridge: MIT Press. 211 p. : ill. ; 24 cm.
10. D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. 1989, Reading: Addison-Wesley Pub. Co. 412 p. : ill. ; 25 cm.
11. B.J. Reardon, Inversion of micromechanical powder consolidation and sintering models using Bayesian inference and genetic algorithms. *Modelling and Simulation in Materials Science and Engineering*, 1999. **7**(6): p. 1061-1081.
12. B.J. Reardon, Fuzzy logic versus niched Pareto multiobjective genetic algorithm optimization. *Modelling and Simulation in Materials Science and Engineering*, 1998. **6**(6): p. 717-734.
13. T.D. Chikalla, Melting Behavior in the System UO2-PuO2. *Journal of the American Ceramic Society*, 1963. **46**(7): p. 323-328.
14. W.L. Lyon, and W.E. Baily, Solid-Liquid Phase Diagram for UO2-PuO2 System. *Journal of Nuclear Materials*, 1967. **22**(3): p. 332.
15. E.A. Aitken, and S.K. Evans, A Thermodynamic Data Program Involving Plutonia and Urania at High Temperatures, in Quarterly Report. 1968, General Electric: Pleasanton, California.
16. M.G. Adamson, E.A. Aitken, and R.W. Caputi, Experimental and Thermodynamic Evaluation of the Melting Behavior of Irradiated Oxide Fuels. *Journal of Nuclear Materials*, 1985. **130**(EB): p. 349-365.
17. L.F. Epstein, Ideal Solution Behavior and Heats of Fusion from UO2-PuO2 Phase Diagram. *Journal of Nuclear Materials*, 1967. **22**(3): p. 340-349.
18. L.F. Epstein, and W.H. Howland, Binary Mixtures of UO2 and Other Oxides. *Journal of the American Ceramic Society*, 1953. **36**(10): p. 334-335.
19. H. Seltz, Thermodynamics of Solid Solutions: I. Perfect Solutions. *Journal of the American Chemical Society*, 1934. **56**: p. 307-311.
20. M.G. Adamson, E.A. Aitken, and R.W. Caputi, Experimental and Thermodynamic Evaluation of the Melting Behavior of Irradiated Oxide Fuels. *Journal of Nuclear Materials*, 1985. **130**(EB): p. 349-365.
21. M.H. Rand, et al., Thermodynamic Properties of the Urania Phase. *Revue Internationale Des Hautes Temperatures Et Des Refractaires*, 1978. **15**(4): p. 355-365.
22. W.L. Lyon, and W.E. Baily, Solid-Liquid Phase Diagram for UO2-PuO2 System. *Journal of Nuclear Materials*, 1967. **22**(3): p. 332.
23. P.P. Budnikov, S.G. Tresvyatski, and V.I. Kushakovsky. in Proc. 2nd U. N. Intern. Conf. Peaceful Uses At. Energy. 1958. Geneva.

Approximate Representations of Random Intervals for Hybrid Uncertainty Quantification in Engineering Modeling

Cliff Joslyn[a] *and Scott Ferson*[b]

[a]Computer and Computational Sciences, Los Alamos National Laboratory, MS B265, Los Alamos, NM, 87545, USA joslyn@lanl.gov
[b]Applied Biomathematics, 100 North Country Road, Setauket, New York, 11733, USA, scott@ramas.com

Abstract:

We review our approach to the representation and propagation of hybrid uncertainties through high-complexity models, based on quantities known as random intervals [15, 20, 21]. These structures have a variety of mathematical descriptions, for example as interval-valued random variables [4], statistical collections of intervals [17], or Dempster-Shafer bodies of evidence on the Borel field [19]. But methods which provide simpler, albeit approximate, representations of random intervals are highly desirable, including p-boxes and traces. Each random interval, through its cumulative belief and plausibility measures functions [36], generates a unique p-box whose constituent CDFs are all of those consistent with the random interval. In turn, each p-box generates an equivalence class of random intervals consistent with it. Then, each p-box necessarily generates a unique trace which stands as the fuzzy set representation of the p-box or random interval. In turn each trace generates an equivalence class of p-boxes. The heart of our approach is to try to understand the tradeoffs between error and simplicity introduced when p-boxes or traces are used to stand in for various random interval operations. For example, Joslyn [18] has argued that for elicitation and representation tasks, traces can be the most appropriate structure, and has proposed a method for the generation of canonical random intervals from elicited traces. But alternatively, models built as algebraic equations of uncertainty-valued variables (in our case, random-interval-valued) propagate uncertainty through convolution operations on basic algebraic expressions, and while convolution operations are defined on all three structures, we have observed that the results of only some of these operations are preserved as one moves through these three levels of specificity. We report on the status and progress of this modeling approach concerning the relations between these mathematical structures within this overall framework.

Keywords: Dempster-Shafer theory, random sets, random intervals, p-boxes, probability bounds, fuzzy arithmetic.

1. INTRODUCTION

Engineering modeling problems are frequently characterized by a large number of inputs with different forms and levels of uncertainty present on them. For example, it might be desirable in a given context to combine uncertainties characterized by coarse-grained probability distributions, strong or weak statistical data, interval data, possibility distributions, or linguistic information represented as fuzzy sets. Propagating such hybrid

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

uncertainties through high-complexity models (whether analytical or computational) is thereby especially challenging, as are elicitations and interpretations of both input and output uncertainties by domain experts and customers.

We have been developing an approach to the representation and propagation of hybrid uncertainties in engineering modeling applications based on quantities known as random intervals [15, 20, 21]. These structures have a variety of mathematical descriptions, for example as interval-valued random variables [4], statistical collections of intervals [17], or Dempster-Shafer bodies of evidence on the Borel field [19].

One of the advantages of random interval structures is their ability to generalize more specific kinds of uncertainty quantities with a relative minimum of computational and mathematical complexity. Nonetheless, random intervals are not especially simple structures to represent or manipulate, and therefore methods which provide simpler, albeit approximate, representations of them are highly desirable. In this paper we report on a framework we are developing to accomplish this. In our approach, random interval quantities can be represented in increasingly simplified and approximate forms through first p-box, and then trace, structures.

A p-box [8] is an ordered pair of monotonically increasing functions which together bound a collection of cumulative probability distribution functions. Each random interval, through its cumulative belief and plausibility measures functions [36], generates a unique p-box whose constituent CDFs are all of those consistent with the random interval. In turn, each p-box generates an equivalence class of random intervals consistent with it.

A trace [17] is defined in this context as a fuzzy quantity on the real line. Each p-box necessarily generates a unique trace which stands as the fuzzy set representation of the p-box or random interval. Under different conditions it can take on the properties of a probability distribution, possibility distribution, or so-called "fuzzy interval" quantity (used in fuzzy arithmetic). In turn each trace generates an equivalence class of p-boxes.

The heart of our approach is to try to understand the tradeoffs between error and simplicity introduced when p-boxes or traces are used to stand in for various random interval operations. For example, Joslyn [18] has argued that for elicitation and representation tasks, traces can be the most appropriate structure, and has proposed a method for the generation of canonical random intervals from elicited traces.

But alternatively, models built as algebraic equations of uncertainty-valued variables (in our case, random-interval-valued) propagate uncertainty through convolution operations on basic algebraic expressions. But while convolution operations are defined on all three structures (random intervals, p-boxes, and traces), we have observed that the results of only some of these operations are preserved as one moves through these three levels of specificity.

In this paper, we report on the status and progress of this modeling approach concerning the relations between these mathematical structures within this overall framework.

2. GENERALIZED UNCERTAINTY QUANTIFICATION FOR ENGINEERING MODELING

Consider the situation where we have a model, perhaps a large computer code, which acts as a function f mapping inputs X to outputs Y. This model f might be quite complex, with high run times, and more significantly multiple input parameters (expressed as the dimensionality of the space X), with different kinds of uncertainty represented on them. Given the necessity for many "gaps" between the information present in our simulations from those of reality (model incompleteness and error, and inherent system variability and imprecision), we wish to represent amounts, degrees, and kinds of these uncertainties in formal systems.

But information available on inputs may be rich or sparse, so-called "aleatory" (related to well-known, but chance, outcomes) or "epistemic" (related to a less-than-well-known outcome), and may be made known through objective measurements or through the subjective elicitation of experts. Mathematically, inputs might be represented as probability distributions, paramaterized classes of probability distributions (e.g. $N(\mu, \sigma)$), by a strong statistical collection of data points, by a sparse such collection, by simple intervals, statistical collections of such intervals, or even by non-quantified linguistic expressions.

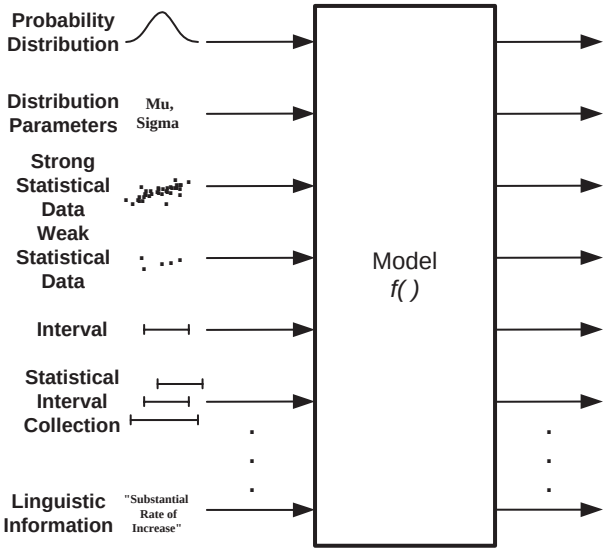


Figure 1. Hybrid uncertainty quantification for an engineering modeling problem.

So given a risk or reliability problem related to our model f as charicatured in Fig. 1, how can we quantify this uncertainty on the input space X, and furthermore propagate it through f to the output space Y? More to the point, how can we do so in a way which respects all the original uncertainty quantifications as provided, making no unnecessary assumptions? Paraphrasing Klir [25], how can we do such in a way which uses no less than, but also no more than, all available information; that is, uses *only*, but *all of*, what we are given?

So wherever possible, we should fit formalism to available information, and not *vice versa*.

Through the 20th century, uncertainty modeling has been dominated by the mathematics of probability, and since Shannon and Weaver [32], information has been defined as a statistical measure of a probability distribution. But also starting in the 1960s, alternative formalisms have arisen. Some of these were intended to stand in contrast to probability theory; others are deeply linked to probability theory, but depart from or elaborate on it in various ways. In the intervening time, there has been a proliferation of methodologies (including interval analysis [28], fuzzy systems [26], fuzzy and monotone measures [24], Dempster-Shafer evidence theory [13], random sets and intervals [22], possibility theory [7], probability bounds [8], rough sets [30], imprecise probabilities [34], and info-gap theory [2]) along with concomitant movements to synthesize and generalize them. Together, following Klir [19, 23], we call these **Generalized Information Theory** (GIT). These methods are increasingly accepted in engineering modeling [15, 33], and our approach is squarely centered here.

As a very simple example motivating our approach, consider that for one of the variables x in our space X, we're given only an interval, that x might be between two quantities a and b, so that $x \in [a, b]$. How do we represent the uncertainty $U(I)$ in I? A standard answer might be to use a uniform probability ditribution $U(I) := p(x)$ with

$$p(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases} ,$$

as shown by the horizontal line in Fig. 2. No doubt this answer is justified (by maximum entropy, insufficient reason, and related principles) *when it is necessary to use a single probability distribution*. But this was not specified in the problem. Indeed, one could argue that *any* probability distribution with support on $[a, b]$ can be justified, perhaps denoted $U(I) \in \mathcal{P}([a, b])$ (perhaps the truncated normal shown in Fig. 2); but better yet, why isn't our uncertainty *all* such distributions: $U(I) = \mathcal{P}([a, b])$ (the box bounded by the dashed lines in Fig. 2).

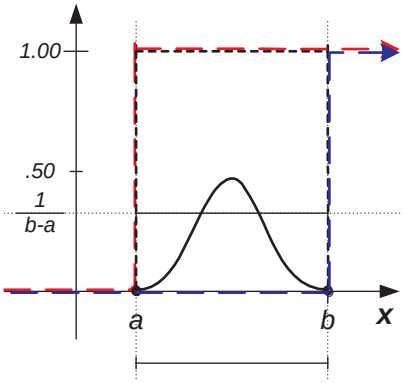


Figure 2. Representations of $x \in [a, b]$.

In its purest form, our answer should in fact be *none* of these, but rather that $U(I)$ is best represented by the information as provided us, that is, by the interval itself: $U(I) = [a, b]$.

However, when it is necessary to combine information for some variable $x \in X$ with another $y \in Y$, then these other forms may be vailable. We argue below that all of these answers are approached consistently within the proper GIT context.

3. NOTATION

Throughout the paper, assume a universe of discourse Ω, with $\omega \in \Omega$. Denote $A \perp B := A \cap B = \emptyset$. Given a class $\mathcal{C} = \{A\} \subseteq 2^{\Omega}$, define the core and support respectively as

$$\mathbf{C}(\mathcal{C}) := \bigcap_{A \in \mathcal{C}} A, \qquad \mathbf{U}(\mathcal{C}) := \bigcup_{A \in \mathcal{C}} A.$$

We begin considering $\Omega = \{\omega_i\}, 1 \leq i \leq n$ to be finite, but move to recognize $\Omega = \mathbb{R}$, and consider Borel sets (half-open interval subsets), elements of a interval Borel field.

Definition 3.1 (Interval Borel Field). *Let*

$$\mathcal{I} := \{[a, b) \subseteq \mathbb{R} : a, b \in \mathbb{R} \cup \{-\infty, \infty\}, a \leq b\},$$

where $\forall a \leq b \in \mathbb{R}, -\infty < a \leq b < \infty, [-\infty, b) := \lim_{a \longrightarrow -\infty} [a, b) \in \mathcal{I}, [a, \infty) := \lim_{b \longrightarrow \infty} [a, b) \in \mathcal{I}, [-\infty, \infty) = \mathbb{R} \in \mathcal{I}$, *and* $[-\infty, -\infty) = [\infty, \infty) := \emptyset \in \mathcal{I}$ *by convention.*

In general, let $I := [a, b) \in \mathcal{I}$.

A vector denoted $\vec{a} = \langle a_i \rangle = \langle a_1, a_2, \ldots, a_m \rangle$ is a structure of length $|\vec{a}| := m$ where each element a_i of the vector is an element of some set $a_i \in X$. The a_i are ordered and may include duplicates. Let an element $b \in X$ be said to be included in a vector $b \in \vec{a}$ if $\exists a_i, b = a_i$. Define subtraction of an element a_i from a vector $\vec{a}$ as a new vector

$$\vec{a} - a_i := \langle a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_m \rangle$$

so that $|\vec{a} - a_i| = m - 1$.

Since a vector may contain duplicate elements $a_{i_1}, a_{i_2} \in \vec{a}$ with $a_{i_1} = a_{i_2}$, therefore each vector $\vec{a}$ determines a unique non-empty set A constructed by including one instance of each element $a_i \in \vec{a}$, so that $b \in \vec{a} \leftrightarrow b \in A$, $1 \leq |A| \leq m$, and the quantity $|\vec{a}| - |A|$ is the number of elements of $\vec{a}$ which are duplicates.

Generalized convolution operators will be introduced, and denoted $\oplus \in \{+, -, \times, \div, \hat{}\}$ for addition, subtraction, multiplication, division, and exponentiation respectively. Let $\vee$ be the maximum and $\wedge$ the minimum operator.

4. RANDOM SETS, RANDOM INTERVALS, AND EVIDENCE THEORY

We now introduce the fundamental ideas of random sets and intervals.

Definition 4.1 (General Random Set). *Given a probability space* $\langle X, \Sigma, \Pr \rangle$, *then a function* $S : X \to 2^{\Omega} - \{\emptyset\}$, *where* $-$ *is set subtraction, is a random subset of* Ω *if* S *is* Pr*-measurable, so that* $\forall A \subseteq \Omega, A \neq \emptyset, S^{-1}(A) \in \Sigma$.

Random sets were originally developed as a branch of stochastic geometry, and their mathematics in general can be quite complex [1, 22]. But for our purposes, and especially in the finite case, they can be seen more simply as random variables taking values on subsets of Ω. Further, they are mathematically isomorphic to bodies of evidence in Dempster-Shafer evidence theory [3, 31]. We now reintroduce random sets in this context.

DEFINITION 4.2 (EVIDENCE FUNCTION, BASIC ASSIGNMENT). *A function* $m\colon 2^{\Omega} \to [0,1]$ *is an evidence function (basic assignment) when* $m(\emptyset) = 0$ *and* $\sum_{A \subseteq \Omega} m(A) = 1$.

DEFINITION 4.3 (FINITE RANDOM SET). *Given an evidence function* m*, then*

$$\mathcal{S} := \{\langle A_j, m_j \rangle : m_j > 0\}, \tag{1}$$

is a finite random set where $A_j \subseteq \Omega, m_j := m(A_j)$*, and* $1 \leq j \leq N := |\mathcal{S}| \leq 2^n - 1$*. Denote the focal set of* $\mathcal{S}$ *as the class* $\mathcal{F}(\mathcal{S}) := \{A_j : m_j > 0\} \subseteq 2^{\Omega}$.

NOTE 4.4. *Each finite random set* $\mathcal{S}$ *determines a unique general random set* $S\colon \mathcal{F}(\mathcal{S}) \to 2^{\Omega} - \emptyset$ *defined on the probability space* $\left\langle \mathcal{F}(\mathcal{S}), 2^{\mathcal{F}(\mathcal{S})}, \Pr \right\rangle$*, where* Pr *is the measure determined by* m *acting as its density function [16]. Moreover,* $\mathcal{S}$ *simply* is *a Dempster-Shafer body of evidence [13].*

We recognize random sets with the following special structures:

Consistent: The global intersection is non-empty: $\forall A_{j_1}, A_{j_2} \in \mathcal{F}(\mathcal{S}), A_{j_1} \not\perp A_{j_2} \leftrightarrow \mathbf{C}(\mathcal{F}(\mathcal{S})) \neq \emptyset$.

Consonant: Focal elements are all nested: $\forall A_{j_1}, A_{j_2} \in \mathcal{F}(\mathcal{S}), A_{j_1} \subseteq A_{j_2}$ or $A_{j_1} \subseteq A_{j_2}$.

Disjoint: No focal elements intersect: $\forall A_{j_1}, A_{j_2} \in \mathcal{F}(\mathcal{S}), A_{j_1} \perp A_{j_2}$.

Specific: All focal elements are singletons: $\forall A_j \in \mathcal{F}(\mathcal{S}), \exists! \omega \in \Omega, A_j = \{\omega\}$.

Note that consonance implies consistency, and specificity implies disjointness. Finally, disjointness implies a lack of consistency, and *vice versa*.

DEFINITION 4.5 (MONOTONE MEASURE, MONOTONE MEASURE TRACE). *[35] Assume a general universe of discourse* Ω*, a class of subsets* $\mathcal{C} \subseteq 2^{\Omega}$*, and a sequence of such sets* $\{A_1, A_2, \ldots\} \subseteq \mathcal{C}$*. Then* $\nu\colon \mathcal{C} \to [0,1]$ *is a monotone measure if*

1. $\nu(\emptyset) = 0$
2. **Monotonicity:**
$$\forall A, B \subseteq \Omega, \quad A \subseteq B \to \nu(A) \leq \nu(B) \tag{2}$$
3. **Continuity from Below:**
$$A_1 \subseteq A_2 \subseteq \ldots \text{ and } \mathbf{U}(\mathcal{C}) \in \mathcal{C} \to \lim_{i \longrightarrow \infty} \nu(A_i) = \nu\left(\bigcup_{i=1}^{\infty} A_i\right).$$

4. **Continuity from Above:**

$$A_1 \supseteq A_2 \supseteq \ldots \text{ and } \mathbf{C}(\mathcal{C}) \in \mathcal{C} \rightarrow \lim_{i \longrightarrow \infty} \nu(A_i) = \nu\left(\bigcap_{i=1}^{\infty} A_i\right).$$

ν is normal when $\nu(\Omega) = 1$. Define the trace of a monotone measure ν as its "one-point covering function" $q_\nu \colon \Omega \rightarrow [0,1]$, with $q_\nu(\omega) := \nu(\{\omega\})$.

Definition 4.6 (Evidence Measures). *The plausibility and belief measures on $\forall A \subseteq \Omega$ are*

$$\mathrm{Pl}(A) := \sum_{A_j \not\perp A} m(A_j), \qquad \mathrm{Bel}(A) := \sum_{A_j \subseteq A} m(A_j),$$

Pl and Bel are generally normal, non-additive monotone measures [35], and are dual, in that $\forall A \subseteq \Omega, \mathrm{Bel}(A) = 1 - \mathrm{Pl}(\overline{A})$.

Random intervals were introduced by Dempster [4].

Definition 4.7 (Finite Random Interval). *A finite random interval, denoted $\mathcal{A}$, is a finite random set on $\Omega = \mathbb{R}$ for which $\mathcal{F}(\mathcal{A}) = \{I_j\} \subseteq \mathcal{I}, 1 \leq j \leq N$.*

Thus a finite random interval is a finite random left-closed interval subset of $\mathbb{R}$.

Previously Ω had been postulated as a finite set, which leads to a great deal of mathematical simplicity. However, even though Ω is now uncountable, complications can still be avoided as long as $\mathcal{A}$ is finite, that is as long as only finitely many (N) focal elements are present. This is because each $I = [l, r) \subseteq \mathbb{R}$ is characterized completely by the two endpoints l and r. With each new focal element A_j, N grows by 1, and the total number of endpoints present in $\mathcal{F}(\mathcal{A})$ grows by at most 2. Thus the focal set of a finite random interval can be completely represented by the finite collection of these endpoints: $\mathcal{F}(\mathcal{A}) = \{I_j\} = \{[l_j, r_j)\}$. It is only these endpoints that need to be considered, and none of the properties of the continuum of points between them is significant.

On this basis we can describe the various components of a random interval. In general denote $I_j = [l_j, r_j)$. Then, denote the vector of all endpoints $\vec{L} := \langle l_1, r_1, l_2, l_2, \ldots, l_j, r_j, \ldots, l_N, r_N \rangle$, and let $L := \{x_k\}$ be the set derived from eliminating duplicates from $\vec{L}$, with $\forall x_k \in L, \exists x_j \in \vec{L}, x_k = x_j$ and $1 \leq k \leq Q := |L|, N + 1 \leq Q \leq 2N = \left|\vec{L}\right|$.

The elements of L determine a class $\Gamma = \{G_k\} \subseteq \mathcal{I}$, now with $1 \leq k \leq Q - 1$, which is the finest partition of the support $\mathbf{U}(\mathcal{A})$ induced by the total intersections of the I_j with each other and with all their intersections recursively. In practice, the G_k are determined simply by ordering the $x_k \in L$ and then traversing them from $\min x_k$ rightward, forming an interval for each point in turn.

An example is shown in Fig. 3, with $N = 4$, $\mathcal{F}(\mathcal{A}) = \{[3.5, 4), [1, 2), [3, 4), [2, 3.5)\}$, and m is as shown. Here $Q = 5$, with $\vec{L} = \langle 3.5, 4, 1, 2, 3, 4, 2, 3.5 \rangle, L = \{1, 2, 3, 3.5, 4\}$, and thus $\Gamma = \{[1, 2), [2, 3), [3, 3.5), [3.5, 4)\}$.

Our definition differs somewhat from others in the literature [9] who use fully closed intervals. But not only is the Borel field $\mathcal{I}$ more consistent with that of measure theory [14, 35], it also makes the algebraic manipulations of the I_j much easier, since e.g. for $x \leq y \leq z, [x, y) \cap [y, z) = \emptyset$.

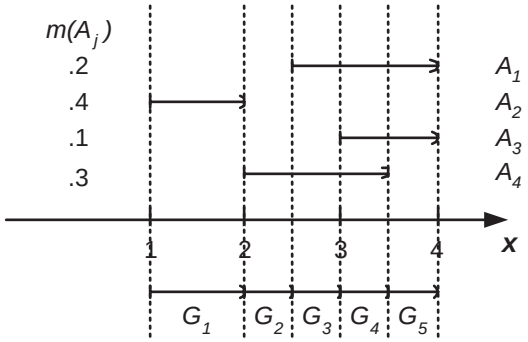


Figure 3. Example of a finite random interval.

In real problems, random intervals are derived from collections of observed intervals. In Joslyn's formalism for random interval measurement [17], the values $m(I_j)$ are derived by their relative frequency in this observation record. But depending on the application, it may or may not be likely that two identical intervals $I, I' = [l, r)$ will be observed, as distinct from another interval I' "very close" to I, for example $I' = [l + \epsilon, r - \delta)$ for some very small ϵ, δ. In this case, as N grows each I_j is observed once, but with increasing refinement among the endpoints $\{l_j, r_j\}$. Thus it is common in real applications to deal with random intervals where all the focal elements $I_j \in \mathcal{F}(\mathcal{A})$ are distinct with $Q \sim 2N$, and therefore each with frequency $m(I_j) = 1/N$.

Yager [36] introduced convolution operators on random intervals.

Definition 4.8 (Random Interval Convolution). *Assume two independent random intervals $\mathcal{A}_1 = \{\langle I_j, m_j \rangle\}, 1 \leq j \leq N_1, \mathcal{A}_2 = \{\langle I_k, m_k \rangle\}, 1 \leq k \leq N_2$, and a convolution operator $\oplus$. Then $\mathcal{A}_1 \oplus \mathcal{A}_2 := \{\langle I_l, m_l \rangle\}$ where:*

$$1 \leq l \leq N_1 N_2, \qquad I_l = \{z = x \oplus y, x \in I_j, y \in I_k\}, \qquad m_l = m_j m_k.$$

5. PROBABILITY BOXES

Random intervals can be difficult structures to elicit, represent, and manipulate. The first of the approximations we introduce are so-called **probability boxes**, or just **p-boxes**.

Definition 5.1 (Probability Box (P-Box)). *A p-box [8] is a structure $\mathcal{B} := \left\langle \underline{B}, \overline{B} \right\rangle$, where $\underline{B}, \overline{B} \colon \mathbb{R} \to [0, 1]$ with:*

1. $\lim_{x \longrightarrow -\infty} \underline{B}(x) \longrightarrow 0, \quad \lim_{x \longrightarrow \infty} \underline{B}(x) \longrightarrow 1$
2. $\lim_{x \longrightarrow -\infty} \overline{B}(x) \longrightarrow 0, \quad \lim_{x \longrightarrow \infty} \overline{B}(x) \longrightarrow 1$
3. *$\underline{B}(x), \overline{B}(x)$ are non-decreasing in x, and*
4. *$\underline{B} \leq \overline{B}$.*

$\underline{B}$ and $\overline{B}$ are interpreted as bounds on cumulative distribution functions (CDFs). In other words, $\mathcal{B} = \left\langle \underline{B}, \overline{B} \right\rangle$ can be identified with the set of all functions $\{F : \underline{B} \leq F \leq \overline{B}\}$ such that F is the CDF of some probability measures Pr on $\mathbb{R}$. For each such F, denote $F \in \mathcal{B}$. In this way, each p-box defines such a class of probability measures.

Definition 5.2 (P-Box Convolution). *Assume two p-boxes $\mathcal{B}_1, \mathcal{B}_2$, and a convolution operator $\oplus$. When $\mathcal{B}_1$ and $\mathcal{B}_2$ are independent, then define*

$$(\mathcal{B}_1 \oplus \mathcal{B}_2)(z) := \left\{ \int_{x \oplus y \leq z} d\Pi(F(x), G(y)) : F \in \mathcal{B}_1, G \in \mathcal{B}_2 \right\},$$

where $\Pi(u, v) = uv$ is the product copula [29].

Each random interval naturally generates a p-box.

Theorem 5.3. *Given a random interval $\mathcal{A}$, then $\mathcal{B}(\mathcal{A}) := \langle \mathrm{BEL}, \mathrm{PL} \rangle$ is a P-Box, where* BEL *and* PL *are the "cumulative belief and plausibility distributions"* $\mathrm{PL}, \mathrm{BEL} \colon \mathbb{R} \to [0, 1]$ *originally defined by Yager [36]*

$$\mathrm{BEL}(x) := \mathrm{Bel}([-\infty, x)), \quad \mathrm{PL}(x) := \mathrm{Pl}([-\infty, x)).$$

Proof. Assume a random interval $\mathcal{A}$. We need to show:

1. First,

   $$\lim_{x \longrightarrow -\infty} \mathrm{BEL}(x) = \lim_{x \longrightarrow -\infty} \mathrm{Bel}([-\infty, x)) = \mathrm{Bel}\left(\lim_{x \longrightarrow -\infty} [-\infty, x) \right) = \mathrm{Bel}([-\infty, -\infty)) = \mathrm{Bel}(\emptyset) = 0.$$

   Similarly,

   $$\lim_{x \longrightarrow \infty} \mathrm{BEL}(x) = \lim_{x \longrightarrow \infty} \mathrm{Bel}([-\infty, x)) = \mathrm{Bel}\left(\lim_{x \longrightarrow \infty} [-\infty, x) \right) = \mathrm{Bel}([-\infty, \infty)) = \mathrm{Bel}(\mathbb{R}) = 1.$$

   The results $\lim_{x \longrightarrow -\infty} \mathrm{PL}(x) = 0, \lim_{x \longrightarrow \infty} \mathrm{PL}(x) = 1$ follow identically.

2. Since $x \leq y \to [-\infty, x) \subseteq [-\infty, y)$, and since Bel and Pl are monotone measures, therefore from monotone measure monotonicity $x \leq y \to \mathrm{BEL}(x) \leq \mathrm{BEL}(y)$, therefore BEL(x) is monotone non-decreasing in x. And similarly for PL.

3. $\forall I \in \mathcal{I}, \mathrm{Bel}(I) \leq \mathrm{Pl}(I)$, and thus in particular $\forall x \in \mathbb{R}, \mathrm{Bel}([-\infty, x)) \leq \mathrm{Pl}([-\infty, x))$, and so $\forall x \in \mathbb{R}, \mathrm{BEL}(x) \leq \mathrm{PL}(x)$.

Therefore $\langle \mathrm{BEL}, \mathrm{PL} \rangle$ is a p-box.

The p-box generated from the example random interval is shown in Fig. 4. Since $\overline{B}$ and $\underline{B}$ partially overlap, the diagram is somewhat ambiguous on its far left and right portions, but note that

$$\overline{B}([-\infty, 1)) = 0, \quad \underline{B}([-\infty, 2,)) = 0, \quad \overline{B}([3, \infty)) = 1, \quad \underline{B}([3.5, \infty)) = 1.$$

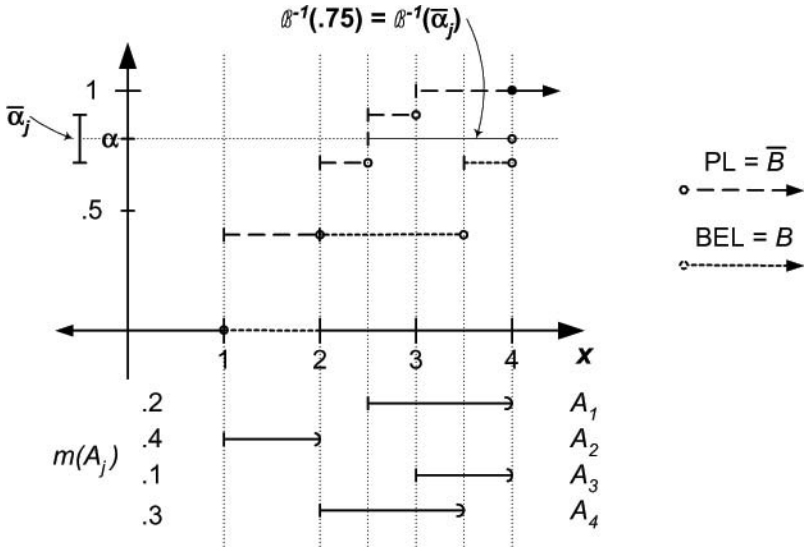


Figure 4. A finite random interval and its piecewise-constant p-box $\mathcal{B}(\mathcal{A})$.

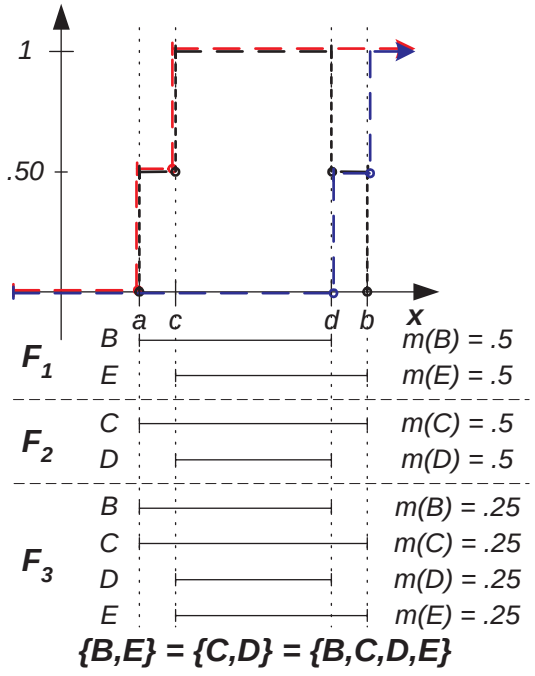


Figure 5. Three different random intervals and their common p-box and trace.

But for the converse, each p-box determines only an equivalence class of random intervals. Consider the example shown in Fig. 5 for $a < c < d < b \in \mathbb{R}, \mathcal{C} = \{B = [a, d), C = [a, b), D = [c, d), E = [c, b)\}$, and three different focal classes $\mathcal{F}_1 = \{B, E\}, \mathcal{F}_2 = \{C, D\}$, and $\mathcal{F}_3 = \{B, C, D, E\}$ with their respective m_1, m_2, and m_3 are shown. We have $\mathcal{B}(\mathcal{A}_1) = \mathcal{B}(\mathcal{A}_2) = \mathcal{B}(\mathcal{A}_3)$.

Thus for a given p-box $\mathcal{B}$, we can denote $\mathcal{A}(\mathcal{B})$ as the equivalence class of random intervals consistent with it: $\mathcal{A}(\mathcal{B}) := \{\mathcal{A} : \mathcal{B}(\mathcal{A}) = \mathcal{B}\}$.

$\underline{B}$ and $\overline{B}$ have inverses under reasonable conditions. Assume that $\underline{B}$ and $\overline{B}$ are piecewise continuous from the left. Then define the quasi-inverses

$$\underline{B}^{-1}(\alpha) := \operatorname*{argmin}_{x \in \mathbb{R}} |\alpha - \underline{B}(x)|, \qquad \overline{B}^{-1}(\alpha) := \operatorname*{argmin}_{x \in \mathbb{R}} |\alpha - \overline{B}(x)|,$$

for $\alpha \in [0, 1]$, and

Definition 5.4 (P-Box Inverse). *Given a p-box $\mathcal{B}$, let $\mathcal{B}^{-1}: [0, 1] \to \mathcal{I}$ where $\forall \alpha \in [0, 1]$*

$$\mathcal{B}^{-1}(\alpha) := \left\{ \left[\overline{B}^{-1}(\alpha), \underline{B}^{-1}(\alpha) \right) \right\}.$$

Condition 4 of (5.1) guarantees that for each $\alpha = [0, 1], \mathcal{B}^{-1}(\alpha)$ exists and is a member of $\mathcal{I}$. When $\underline{B}$ and $\overline{B}$ are piecewise-constant, $\mathcal{B}^{-1}$ naturally partitions $[0, 1]$ into disjoint intervals denoted $\bar{\alpha}_j$ over which $\forall \alpha, \alpha' \in \bar{\alpha}_j, \mathcal{B}^{-1}(\alpha) = \mathcal{B}^{-1}(\alpha')$. In practice, denote $\bar{\alpha}_j := \left[\alpha_j^l, \alpha_j^r\right]$, where

$$\alpha_j^l = \operatorname*{argmin}_{x \in \mathbb{R}} \overline{B}(x) \geq \alpha, \qquad \alpha_j^u = \operatorname*{argmax}_{x \in \mathbb{R}} \underline{B}(x) \leq \alpha.$$

This is shown in Fig. 4.

Given a piecewise-constant p-box, there is a canonical way to construct a random interval consistent with it.

Definition 5.5 (Canonical Random Interval from P-Box). *Assume a p-box $\mathcal{B}$. Then construct $\mathcal{A}^*(\mathcal{B}) := \{\langle \mathcal{B}^{-1}(\bar{\alpha}_j), m_j \rangle\}$, where $\mathcal{B}^{-1}(\bar{\alpha}_j) := \mathcal{B}^{-1}(\alpha_j^l) = \mathcal{B}^{-1}(\alpha_j^r)$ and $m_j = \alpha_j^l - \alpha_j^r$.*

Theorem 5.6. *$\mathcal{A}^*(\mathcal{B})$ is a random interval, and $\mathcal{A}^*(\mathcal{B}) \in \mathcal{A}(\mathcal{B})$.*

Proof. It is evident from the definitions (5.4) and (5.5) that each $\bar{\alpha}_j \in \mathcal{I}$. Also, since the $\hat{\alpha_j}$ partition $[0, 1]$, therefore

$$\sum_j m_j = \sum_j \left(\alpha_j^l - \alpha_j^r\right) = 1.$$

It is relatively easy to see in Fig. 4 that $\mathcal{A}^*(\mathcal{B}(\mathcal{A})) = \mathcal{A}$, although we know that this is not always so.

6. RANDOM INTERVAL TRACES

A fuzzy (sub)set of Ω, denoted $\widetilde{A} \widetilde{\subseteq} \Omega$, is determined by its membership function, which is any function of the form $\mu_{\widetilde{A}}: \Omega \to [0,1]$. Denote the core of a fuzzy set as $\mathbf{C}(\mu) := \{\omega \in \Omega : \mu(\omega) = 1\}$.

The value of $\mu_{\widetilde{F}}(\omega_i)$ indicates the degree or extent to which $\omega_i \in \Omega$. Fuzzy sets generalize classical (crisp) sets in that a subset $A \subseteq \Omega$ has a memberhsip function defined as the characteristic function $\mu_A := \chi_A$. In the sequel, let each fuzzy set be considered to be a fuzzy subset of the reals $\widetilde{A} \widetilde{\subseteq} \mathbb{R}$.

The trace of any monotone measure defined on $\mathbb{R}$ is a fuzzy set.

COROLLARY 6.1. *Given a monotone measure ν, then q_ν is a membership function.*

Proof. Follows trivially from the definition of trace (4.5).

Fuzzy sets also have convolutions.

DEFINITION 6.2 (FUZZY SET CONVOLUTION). *Assume two fuzzy intervals $\widetilde{A}_1, \widetilde{A}_2$, a convolution operator $\oplus$, and a T-norm $\sqcap$. Let $\widetilde{A}_3 = \widetilde{A}_1 \oplus \widetilde{A}_2$. Then*

$$\mu_{\widetilde{A}3}(z) := \bigvee_{x \oplus y = z} \mu_{\widetilde{A}1}(x) \sqcap \mu_{\widetilde{A}2}(y).$$

There are two special kinds of fuzzy subsets which are of particular interest to us.

DEFINITION 6.3 (FUZZY INTERVAL). *[5, 6] A fuzzy subset of the real line $\widetilde{F} \widetilde{\subseteq} \mathbb{R}$ is a* **fuzzy interval** *if $\widetilde{F}$ is maximally normalized and convex, so that*

$$\forall x, y \in \mathbb{R}, \quad \forall z \in [x, y], \quad \mu_{\widetilde{F}}(z) \geq \mu_{\widetilde{F}}(x) \wedge \mu_{\widetilde{F}}(y).$$

Note that convexity here implies unimodality in the weak sense that $\mathbf{C}(\widetilde{F})$ is a closed interval. This goes to a limit for fuzzy numbers.

DEFINITION 6.4 (FUZZY NUMBER). *A* **fuzzy number** *is a fuzzy interval $\widetilde{F}$ where $\exists x \in \mathbb{R}, \mathbf{C}(\widetilde{F}) = \{x\}$.*

So each random interval naturally generates a trace.

DEFINITION 6.5 (RANDOM INTERVAL TRACE). *Given a random interval $\mathcal{A}$, define the function $\rho_{\mathcal{A}}: \mathbb{R} \to [0,1]$ as the plausibilistic trace, or just trace, of $\mathcal{A}$, where $\rho_{\mathcal{A}} = q_{\mathrm{Pl}}$.* Therefore

$$\forall x \in \mathbb{R}, \quad \rho_{\mathcal{A}}(x) := \mathrm{Pl}(\{x\}) = \sum_{A_j \ni x} m_j. \tag{3}$$

An example is shown in Fig. 6, with $\mathcal{A}$ as before, and ρ shown in the top of the figure.

But for the converse, each fuzzy subset of $\mathbb{R}$ determines only an equivalence class of random intervals. Consider again the example shown in Fig. 5. Each of the three random intervals $\mathcal{A}_1, \mathcal{A}_2$, and $\mathcal{A}_3$ generates exactly the same trace, here shown in the bold, dashed, "step-pyramid" shaped curve.

So for a given fuzzy set $\widetilde{F}$, denote $\mathcal{A}\left(\widetilde{F}\right)$ as the equivalence class of random intervals consistent with it: $\mathcal{A}(\widetilde{F}) := \left\{\mathcal{A} : \rho(\mathcal{A}) = \widetilde{F}\right\}$. The structure of this equivalence class

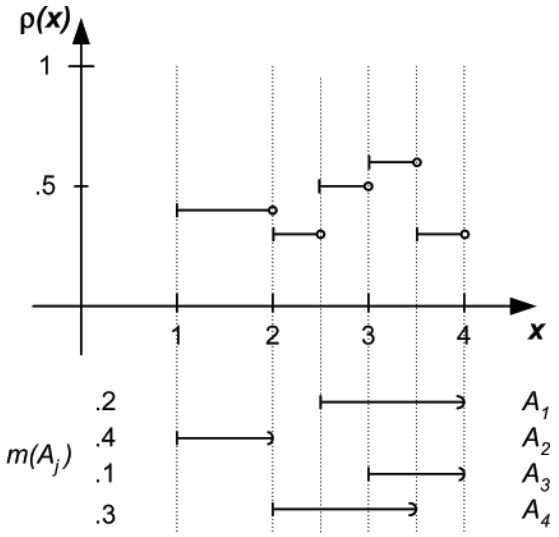


Figure 6. Example random interval with its trace and its components.

is not simple, and has been dealt with in depth by Goodman and his colleagues [10–12]. Furthermore, they have shown that operations on fuzzy sets are preserved when projected through the random set space.

Joslyn has shown the basis to derive fuzzy mathematics from (empirically derived) random intervals [17]. First, ρ is constant over each $G_k \subseteq \mathbb{IR}$. But moreoever:

THEOREM 6.6. *[17] The trace $\rho_{\mathcal{A}}$ of a random interval $\mathcal{A}$ is a fuzzy interval iff $\mathcal{A}$ is consistent.*

This is important because fuzzy intervals generalize crisp intervals as fuzzy sets generalize crisp sets. They are also the basis for "fuzzy arithmetic", since the set of fuzzy intervals is closed under convolution. In addition:

PROPOSITION 6.7. *Given two fuzzy intervals $\widetilde{F}_1, \widetilde{F}_2$, a convolution operator $\oplus$, and a T-norm $\sqcap$, then $\widetilde{F}_1 \oplus \widetilde{F}_2$ is not necessarily a fuzzy interval. However,*

$$\mathbf{C}\left(\widetilde{F}_1 \oplus \widetilde{F}_2\right) = \mathbf{C}\left(\widetilde{F}_1\right) \oplus \mathbf{C}\left(\widetilde{F}_2\right), \quad \mathbf{U}\left(\widetilde{F}_1 \oplus \widetilde{F}_2\right) = \mathbf{U}\left(\widetilde{F}_1\right) \oplus \mathbf{U}\left(\widetilde{F}_2\right).$$

7. P-BOXES AND TRACES

We now begin to explore the relations among the categories of random intervals and their trace and p-box representations. These are diagrammed in Fig. 7.

First, a given p-box determine a trace uniquely.

DEFINITION 7.1 (TRACE OF A P-BOX). *Assume a p-box $\mathcal{B}$. Then its trace, denoted $\rho(\mathcal{B})$, is determined by $\rho(\mathcal{B}) := \overline{B} - \underline{B}$.*

The trace determined in this way from the p-box of a random interval is the same as the trace of the random interval itself, as we will now show.

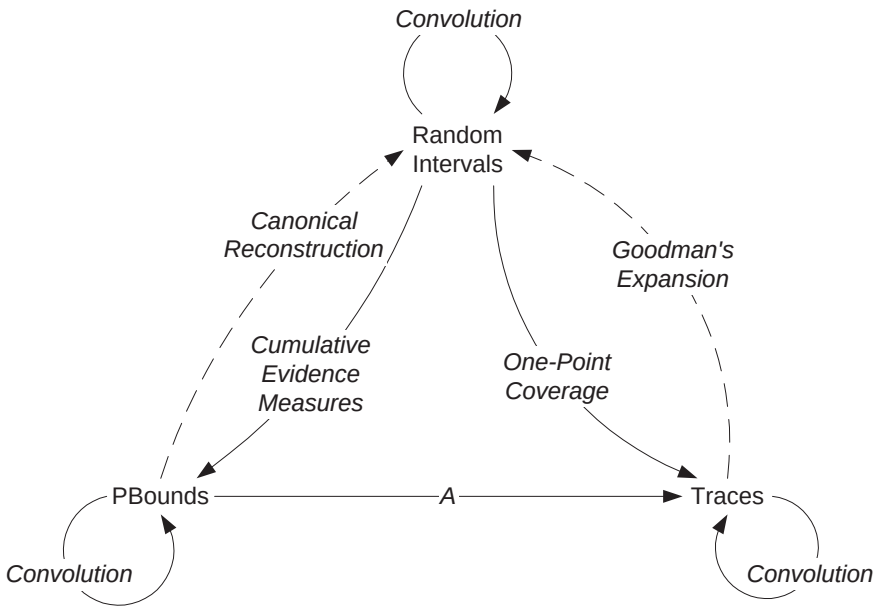


Figure 7. Relations among random intervals, p-boxes, and traces.

THEOREM 7.2. *For all random intervals* $\mathcal{A}$, $\rho(\mathcal{B}(\mathcal{A})) = \rho(\mathcal{A})$.

Proof. Assume a random interval $\mathcal{A}$. Fix a point $x \in \mathbb{R}$. Then

$$\mathrm{PL}(x) = \mathrm{Pl}((\infty, x)) = \sum_{I_j \not\subset (\infty, x]} m(I_j) = \sum_{l_j \leq x} m(I_j)$$

$$\mathrm{BEL}(x) = \mathrm{Bel}((\infty, x]) = \sum_{I_j \subseteq (\infty, x]} m(I_j) = \sum_{x \geq r_j} m(I_j). \tag{4}$$

Then from (5.5) and (7.1),

$$\rho(\mathcal{B}(\mathcal{A}))(x) = \mathrm{PL}(x) - \mathrm{BEL}(x) = \sum_{l_j \leq x < r_j} m(I_j) = \sum_{x \in I_j} m(I_j) = \rho(\mathcal{A})(x). \tag{5}$$

Note how crucial the use of half-open intervals is. The weak inequality in (4) results through subtraction in the appropriate half-open interval in (5), and this would have been the case whether the I_j were closed or not. These results can be checked with some simple diagrammatic reasoning between Fig. 4 and Fig. 6.

But conversely, it might be that the trace of a random interval has multiple p-boxes which could generate it.

8. FUTURE WORK

Future development requires the following considerations:

- Given that $\mathcal{B} \to \rho$, then it should be that $\mathcal{A}(\mathcal{B}) \subseteq \mathcal{A}(\mathcal{B}(\rho))$. What about the converse?

- For a given $\mathcal{A}$, compare $\mathcal{A}(\mathcal{B}(\mathcal{A}))$ and $\mathcal{A}(\rho(\mathcal{A}))$.

Comparison of canonical reconstructions:

- For a given $\mathcal{A}$, compare $\mathcal{A}^*(\mathcal{B}(\mathcal{A}))$ and $\mathcal{A}^*(\rho(\mathcal{A}))$.
- Keep going: compare $\rho(\mathcal{A}^*(\mathcal{B}))$ and $\mathcal{B}(\mathcal{A}^*(\rho))$.

Convolutions. Similar questions for convolutions all around. In particular:

- Compare $\mathcal{B}(\mathcal{A}_1 \oplus \mathcal{A}_2)$ with $\mathcal{B}(\mathcal{A}_1) \oplus \mathcal{B}(\mathcal{A}_2)$.
- Compare $\mathcal{A}^*(\mathcal{B}(\mathcal{A}_1 \oplus \mathcal{A}_2))$ with $\mathcal{A}^*(\mathcal{B}(\mathcal{A}_1) \oplus \mathcal{B}(\mathcal{A}_2))$.
- Compare $\rho(\mathcal{A}_1 \oplus \mathcal{A}_2)$ with $\rho(\mathcal{A}_1) \oplus \rho(\mathcal{A}_2)$.
- Compare $\mathcal{A}^*(\rho(\mathcal{A}_1 \oplus \mathcal{A}_2))$ with $\mathcal{A}^*(\rho(\mathcal{A}_1) \oplus \rho(\mathcal{A}_2))$.

Acknowledgements

This work has been supported by a research grant from Sandia National Laboratories as part of the Department of Energy Accelerated Strategic Computing Initiative (ASCI). The first author also wishes to thank the Los Alamos Nuclear Weapons Program for their continued support, especially the Engineering Sciences and Applications Division.

REFERENCES

1. Artstein, Z and Vitale, RA: (1975) "Strong Law of Large Numbers for Random Compact Sets", *Annals of Probability*, v. **3**, pp. 879-882
2. Ben-Haim, Yakov: (2001) *Information-Gap Decision Theory*, Academic Press, London
3. Dempster, AP: (1967) "Upper and Lower Probabilities Induced by a Multivalued Mapping", *Annals of Mathematical Statistics*, v. **38**, pp. 325-339
4. Dempster, AP: (1968) "Upper and Lower Probabilities Generated by a Random Interval", *Annals of Mathematical Statistics*, v. **39**:3, pp. 957-966
5. Dubois, Didier and Prade, Henri: (1978) "Operations on Fuzzy Numbers", *Int. J. Systems Science*, v. **9**, pp. 613-626
6. Dubois, Didier and Prade, Henri: (1987) "Fuzzy Numbers: An Overview", in: *Analysis of Fuzzy Information*, v. **1**, pp. 3-39, CRC Press, Boca Raton
7. Dubois, Didier and Prade, Henri: (1988) *Possibility Theory*, Plenum Press, New York
8. Ferson, Scott; Kreinovich, V; and Ginzburg, L et al.: (2002) "Constructing Probability Boxes and Dempster-Shafer Structures", *SANDReport 2002-4015*, Sandia National Lab, Albuquerque NM, `http://www.sandia.gov/epistemic/Reports/SAND2002-4015.pdf`
9. Gil, MA: (1992) "A Note on the Connection Between Fuzzy Numbers and Random Intervals", *Statistics and Probability Letters*, v. **13**:4, pp. 311-319
10. Goodman, IR: (1982) "Fuzzy Sets as Equivalence Classes of Random Sets", in: *Fuzzy Sets and Possibility Theory*, ed. R Yager, pp. 327-343, Pergamon, Oxford

11. Goodman, IR: (1994) "A New Characterization of Fuzzy Logic Operators Producing Homomorphic-Like Relations with One-Point Coverages of Random Sets", in: *Advances in Fuzzy Theory and Technology*, ed. PP Wang, pp. 133-159, Bookwrights Press, Raleigh NC
12. Goodman, IR; Mahler, Ronald PS; and Nguyen, HT: (1997) *Mathematics of Data Fusion*, Kluwer
13. Guan, J and Bell, DA: (1992) *Evidence Theory and its Applications*, v. **1-2**, North Holland
14. Halmos, PR: (1950) *Measure Theory*, van Nostrand, New York
15. Helton, JC and Oberkampf, WL: (2004) "Special Issue: Alternative Representations of Epistemic Uncertainty", in: *Reliability Engineering and Systems Safety*, in press
16. Joslyn, Cliff: (1996) "Aggregation and Completion of Random Sets with Distributional Fuzzy Measures", *Int. J. of Uncertainty, Fuzziness, and Knowledge-Based Systems*, v. **4**:4, pp. 307-329
17. Joslyn, Cliff: (1997) "Measurement of Possibilistic Histograms from Interval Data", *Int. J. General Systems*, v. **26**:1-2, pp. 9-33
18. Joslyn, Cliff: (2003) "Multi-Interval Elicitation of Random Intervals for Engineering Reliability Analysis", in: *2003 Int. Symp. on Uncertainty Modeling and Analysis (ISUMA 03)*
19. Joslyn, Cliff and Booker, Jane: (2004) "Generalized Information Theory for Engineering Modeling and Simulation", *Engineering Design Reliability Handbook*, CRC Press, in press
20. Joslyn, Cliff and Helton, Jon C: (2002) "Bounds on Plausibility and Belief of Functionally Propagated Random Sets", in: *Proc. Conf. North American Fuzzy Information Processing Society (NAFIPS 2002)*, pp. 412-417
21. Joslyn, Cliff and Kreinovich, Vladik: (2002) "Convergence Properties of an Interval Probabilistic Approach to System Reliabilility Estimation", *Int. J. General Systems*, in press
22. Kendall, DG: (1974) "Foundations of a Theory of Random Sets", in: *Stochastic Geometry*, ed. EF Harding and DG Kendall, pp. 322-376, Wiley, New York
23. Klir, George: (1991) "Generalized Information Theory", *Fuzzy Sets and Systems*, v. **40**, pp. 127-142
24. Klir, George and Smith, Richard M: (2000) "On Measuring Uncertainty and Uncertainty-Based Information: Recent Developments", *Annals of Mathematics and Artificial Intelligence*, v. **32**:1-4, pp. 5-33
25. Klir, George and Wierman, Mark J: (1999) *Uncertainty-Based Information Elements of GIT*, Springer-Verlag, Berlin
26. Klir, George and Yuan, Bo: (1995) *Fuzzy Sets and Fuzzy Logic*, Prentice-Hall, New York
27. Matheron, G: (1975) *Random Sets and Integral Geometry*, Wiley, New York
28. Moore, RM: (1979) *Methods and Applications of Interval Analysis*, in: *SIAM Studies in Applied Mathematics*, SIAM, Philadelphia
29. Nelsen, Roger B: (1999) *Introduction to Copulas*, Springer-Verlag
30. Pawlak, Zdzislaw: (1991) *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer, Boston
31. Shafer, Glen: (1976) *Mathematical Theory of Evidence*, Princeton U Press, Princeton
32. Shannon, CE and Weaver, W: (1964) *Mathematical Theory of Communication*, U Illinois Press, Urbana
33. Tonon, Fulvio and Bernardini, Alberto: (1998) "A Random Set Approach to the Optimization of Uncertain Structures", *Computers and Structures*, v. **68**, pp. 583-600

34. Walley, P: (1999) "An Introduction to the Mathematics of Imprecise Probabilities", in: *Proc. Int. Symp. on Imprecise Probabilities: Theory and Applications*
35. Wang, Zhenyuan and Klir, George J: (1992) *Fuzzy Measure Theory*, Plenum Press, New York
36. Yager, Ronald R: (1986) "Arithmetic and Other Operations on Dempster-Shafer Structures", *Int. J. Man-Machine Studies*, v. **25**, pp. 357-366

Designing Computer Experiments to Determine Robust Control Variables

Jeffrey S. Lehman [a], *Thomas J. Santner* [b], *William I. Notz* [b], *and Donald Bartel*[c]

[a]Home Finance Marketing Analytics, JPMorganChase, Columbus, OH 43240
[b]The Ohio State University, 1958 Neil Ave., Columbus, OH 43210
[c]Cornell University, Upson Hall, Ithaca, NY 14853

Abstract

This manuscript is an extended abstract that outlines the problem solved in the paper of the same name that appears in *Statistica Sinica* and for which additional details are given in the technical report by Lehman et al. (2002).

Keywords: Computer experiments, Robust control variables, Expected improvement, Noise variables, Robust optimization, Sequential design

This research is concerned with the design of computer experiments when there are two types of inputs: control variables and environmental variables. Control variables, also called manufacturing variables, are determined by a product designer while environmental variables, called noise variables in the quality control literature, are uncontrolled in the field but take values that are characterized by a probability distribution. The *objective* is to design the computer experiment so as to find a set of *control variables* which are "robust" in the sense described below.

There are several different notions of robustness that have been proposed in the literature. To explain the current approach and its relationship to other approaches, suppose that $y(\cdot)$ denotes the output of the computer code and $\boldsymbol{x} = (\boldsymbol{x}_c, \boldsymbol{x}_e)$ denotes the input where $\boldsymbol{x}_c$ is the vector of control variables and $\boldsymbol{x}_e$ is the vector of environmental variables. Also let $F(\cdot)$ denote a tentative guess of distribution of the environmental variables, $\boldsymbol{X}_e$. If $F(\cdot)$ is known with *certainty*, we typically focus attention on determining either the distribution of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ ("uncertainty analysis", see, for examples, O'Hagan and Haylock, 1997 or O'Hagan et al., 1999) *or* some summary of this distribution such as its mean $\mu(\boldsymbol{x}_c, F) = E_F\{y(\boldsymbol{x}_c, \boldsymbol{X}_e)\}$ (see, for example, Williams et al., 2000).

If $F(\cdot)$ is *unknown*, either completely or up to a finite vector of parameters, then $\mu(\boldsymbol{x}_c, F)$ may not be useful if its value is "sensitive" to the assumed $F(\cdot)$. The *minimax approach* to robustness assumes that a family $\mathcal{G}$ of distributions can be specified that contains the unknown $F(\cdot)$ (Huber, 1981). This approach defines $\boldsymbol{x}_c^{\mathcal{G}}$ to be $\mathcal{G}$*-robust* if

$$\max_{G \in \mathcal{G}} \quad \mu(\boldsymbol{x}_c^{\mathcal{G}}, G) = \min_{\boldsymbol{x}_c \in \mathcal{X}_c} \max_{G \in \mathcal{G}} \mu(\boldsymbol{x}_c, G).$$

Minimax robustness adopts a pessimistic viewpoint because it attempts to guard against the worst-case scenario among all $\boldsymbol{X}_e$ distributions in $\mathcal{G}$. The *Bayesian approach* to robustness focuses on the mean

$$\mu^{\Pi}(\boldsymbol{x}_c) = \int_{G \in \mathcal{G}} \mu(\boldsymbol{x}_c, G)\, d\,\Pi(G), \tag{1}$$

over the possible $\boldsymbol{X}_e$ distributions in $\mathcal{G}$; here $\Pi(\cdot)$ is a prior distribution on $\mathcal{G}$. A $\boldsymbol{x}_c^{\Pi}$ that minimizes (1) is said to be Π*-robust*.

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

Lehman et al. (2004) adopts a *Taguchi-like approach* to robustness. Assuming that interest lies in $\mu(\boldsymbol{x}_c, F)$, the idea of this type of robustness is that if $y(\boldsymbol{x}_c, \boldsymbol{x}_e)$ is relatively "flat" in $\boldsymbol{x}_e$ for a *given* $\boldsymbol{x}_c$ value, then the mean of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ will be relatively independent of the choice of $F(\cdot)$ (and thus be robust to misspecification of $F(\cdot)$). Formally, we quantify the flatness of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ by $\sigma^2_G(\boldsymbol{x}_c) = \text{Var}_G[y(\boldsymbol{x}_c, \boldsymbol{X}_e)]$, where $G(\cdot)$ is a user-selected distribution on $\boldsymbol{X}_e$. We define $\boldsymbol{x}_c^M$ to be *M-robust* if $\boldsymbol{x}_c^M$ minimizes $\mu(\boldsymbol{x}_c, F)$ subject to a constraint on $\sigma^2_G(\boldsymbol{x}_c)$. Alternatively, and perhaps more in keeping with the quality control concept of having a "target" mean, we define $\boldsymbol{x}_c^V$ to be *V-robust* if it minimizes $\sigma^2_G(\boldsymbol{x}_c)$ subject to a constraint on $\mu_F(\boldsymbol{x}_c)$.

Lehman et al. (2004) present sequential strategies for determing $\boldsymbol{x}_c^M$ and $\boldsymbol{x}_c^V$ based on a notion of "expected improvement" to select successive inputs to the computer code (Schonlau, 1997; Schonlau et al., 1998; Williams et al., 2000). The approach is Bayesian viewpoint. The computer code is treated as a realization of a Gaussian stochastic process; this random function model is the basis for interpolating the response based on a small training sample of computer runs (Sacks et al., 1989; Koehler and Owen, 1996). The predictive interpolator is used in place of the computer code to investigate the input–output relationship. The paper concludes by illustrating the performance of the algorithms proposed with examples that involve several different experimental goals. For the reader's convenience, the entire reference list for the published paper is given below.

ACKNOWLEDGEMENT

We would like to thank Brian Williams for helpful conversation in developing these ideas

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer–Verlag.

Chang, P. B., B. J. Williams, W. I. Notz, T. J. Santner, and D. L. Bartel (1999). Robust optimization of total joint replacements incorporating environmental variables. *Journal of Biomechanical Engineering 121*, 304–310.

Dixon, L. C. W. and G. P. Szego (1978). The global optimisation problem: an introduction. In L. C. W. Dixon and G. P. Szego (Eds.), *Towards Global Optimisation*, Volume 2, pp. 1–15. North Holland, Amsterdam.

Handcock, M. S. and M. L. Stein (1993). A bayesian analysis of kriging. *Technometrics 35*, 403–410.

Haylock, R. G. and A. O'Hagan (1996). On inference for outputs of computationally expensive algorithms with uncertainty on the inputs. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics*, Volume 5, pp. 629–637. Oxford University Press.

Huber, P. J. (1981). *Robust Statistics*. New York: J. Wiley.

Jones, D. R., M. Schonlau, and W. J. Welch (1998). Efficient global optimization of expensive black–box functions. *Journal of Global Optimization 13*, 455–492.

Koehler, J. R. and A. B. Owen (1996). Computer experiments. In S. Ghosh and C. R. Rao (Eds.), *Handbook of Statistics*, Volume 13, pp. 261–308. Elsevier Science B.V.

Lehman, J. (2002). *Sequential Design of Computer Experiments for Robust Parameter Design*. Ph. D. thesis, Department of Statistics, Ohio State University, Columbus, OH USA.

Lehman, J. S., T. J. Santner, and W. I. Notz (2002). Robust parameter design for computer experiments. Technical Report 708, Department of Statistics, The Ohio State University.

Lehman, J. S., T. J. Santner, and W. I. Notz (2004). Robust parameter design for computer experiments. *Statistica Sinica 14*, 571–580.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 21*, 239–245.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia: SIAM.

O'Hagan, A. (1992). Some bayesian numerical analysis. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics*, Volume 4, pp. 345–363. Oxford University Press.

O'Hagan, A. and R. G. Haylock (1997). Bayesian uncertainty analysis and radiological protection. In V. Barnett and K. F. Turkman (Eds.), *Statistics for the Environment*, Volume 3, pp. 109–128. J. Wiley, New York.

O'Hagan, A., M. C. Kennedy, and J. E. Oakley (1999). Uncertainty analysis and other inference tools for complex computer codes. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics*, Volume 6, pp. 503–524. Oxford University Press.

Owen, A. B. (1995). Randomly permuted (t, m, s)-nets and (t, s) sequences. In H. Niederreiter and P. J.-S. Shiue (Eds.), *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pp. 299–317. Springer-Verlag, New York.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989). Design and analysis of computer experiments. *Statistical Sciences 4*, 409–423.

Schonlau, M. (1997). *Computer experiments and global optimization*. Ph. D. thesis, Department of Statistics and Actuarial Science, University of Waterloo.

Schonlau, M., W. Welch, and D. Jones (1998). Global versus local search in constrained optimization of computer models. In N. Flournoy, W. Rosenberger, and W. Wong (Eds.), *New Developments and Applications in Experimental Design*, Volume 34, pp. 11–25. Institute of Mathematical Statistics.

Welch, W. J. (1985). Aced: Algorithms for the construction of experimental designs. *The American Statistician 39*, 146.

Welch, W. J., T.-K. Yu, S. M. Kang, and J. Sacks (1990). Computer experiments for quality control by parameter design. *Journal of Quality Technology 22*, 15–22.

Williams, B. J., J. S. Lehman, T. J. Santner, and W. I. Notz (2003). Constrained optimization of computer experiments having multiple responses. *In Revision*.

Williams, B. J., T. J. Santner, and W. I. Notz (2000). Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica 10*, 1133–1152.

Assa: algorithms for stochastic sensitivity analysis

Michiel J. W. Jansen

Biometris, Wageningen University and Research Centre, P.O. Box 100, 6700 AC Wageningen, The Netherlands
E-mail: michiel.jansen@wur.nl

Abstract: ASSA is a public-domain open-source library of algorithms for stochastic sensitivity analysis in ANSI C. It should serve as a documented collection of basic and more sophisticated algorithms in that field. Its open character ought to advance the applicability, the quality and a restrained kind of completeness of the collection.

Keywords: software, Monte Carlo, latin hypercube, analysis of variance, regression-based, regression-free

1. INTRODUCTION

Various software products exist for stochastic sensitivity analysis (SSA; the adjective 'stochastic' is added to distinguish the subject from deterministic sensitivity analysis). Saltelli, Chan and Scott [1] contains an overview of software available in the year 2000. The software packages mentioned there are closed in the sense that you can hardly change or add components. The section in [1] on generic algorithms is still far from complete. Thus, there does not seem to exist a fairly complete, coherent, and documented collection of algorithms for SSA in a basic programming language like C or Fortran. The ASSA project has the purpose to begin filling this gap. The collection is available in the public domain, in such a form that everyone can use the software freely. It is hoped that users will suggest improvements or additions. The long-term goal is a collection of documented algorithms in the spirit of the famous series of Numerical Recipes [2], but with a slightly different legal status.

Model builders should be enabled to incorporate the algorithms into their own software, for instance in order to accompany model statements with an indication of inaccuracy due to input uncertainty. Another application is inclusion of SSA algorithms into generic frameworks for building, coupling and analysing models.

The language used is ANSI C, written in such a style that translation into another basic programming language should pose no serious problems.

At present, ASSA consists mainly of conventional algorithms for sensitivity analysis. Apart from auxiliary routines, the algorithms can be divided into algorithms for constructing model input samples and algorithms for analysing the corresponding model output samples. All sensitivity analyses are variance-based.

The long-term goals are: a gradually improving and extending collection of basic and advanced algorithms for SSA, leading to a moderate form of standardisation; uniform description of these algorithms via C-programs; a form of publication inviting comments and additions, while enabling flexible use of the algorithms.

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

2. OVERVIEW OF ALGORITHMS

2.1. Input generators

Random generators are used to construct a sample from the distribution of unknown parameters or other model inputs. In the present version, the statistical properties of the inputs can be described in terms of their grade correlation (often loosely called rank correlation). Each of the individual variables is defined through the type and parameters of its distribution.

There are two basic random generators: uniform(0,1) and multinormal(μ, Σ). For the rest, the drawing of random samples is done in two steps.

The first step draws a sample from the k-dimensional unit-hypercube. Each of the k variables thus sampled is more or less randomly and more or less uniformly distributed over the interval (0,1). Some examples: independent; dependent with given grade correlation; latin hypercube [3]; latin hypercube with forced rank correlation [4]; and by way of example a systematic sample constructed from a saturated main-effect design.

The next step transforms these (0,1) variables into variables with the required distribution. The distributions currently available are: uniform, triangular, normal, log-normal, beta and gamma. Auxiliary routines are supplied to derive the standard parameters of distributions from information about means and variances, or the quantiles.

2.2. Analysis

In ASSA's present version, all sensitivity analyses are variance-based, i.e. they perform some kind of analysis of variance on the model output. During the 1990's there seems to have grown consensus that this form of sensitivity analysis is quite adequate for most purposes. The algorithms provide the possibility to estimate the variance contributions of *groups* of inputs, which often facilitates the interpretation of the results, especially when variables belonging to different groups are stochastically independent. There is an algorithm for the most common form of sensitivity analysis: the one based on linear regression. A spline-regression-based analysis is on the list of desiderata. All ingredients are present for a Sobol' type sensitivity analysis [5]. An algorithm for winding stairs analysis [6] is under construction. For the time being, only two simple test functions are included, whose sensitivity properties can be calculated analytically.

2.3. Auxiliary routines

ASSA contains routines to summarise the statistical properties of a sample: variance matrix, correlation matrix, mean, variance, median and rank-correlation. There is an algorithm to check if a symmetric matrix is positive definite. Graphical routines are not included.

3. SUPPORTING SOFTWARE REQUIRED

The current version of ASSA frequently uses algorithms from Numerical Recipes in C (NRC; Press et al., 1992). Thus, one may only use this version of ASSA in applications where one is entitled to use the algorithms from Numerical Recipes. The NRC procedures used for ASSA serve mainly to allocate and free memory space for vectors and matrices, to generate uniform random numbers, and to calculate special functions relating to probability distributions.

4. AVAILABILITY AND LEGAL MATTERS

ASSA's manual and source code can be downloaded free of charge from the website of NPB, the Dutch Nature Policy Assessment Office: www.natuurplanbureau.nl. Once there, click "publicaties" in the left column; next click "werkdocumenten" in the left column; then click "2004" in the central block; and finally download the ASSA manual and source code from the displayed list of documents. A link to that website indicated will be offered to the SAMO site http://sensitivity-analysis.jrc.cec.eu.int.

Permission to use, copy and distribute this software and its documentation for any purpose is granted without fee, provided that the entire package is kept together and that this permission and disclaimer notice appears in all copies. NPB, Biometris, Plant Research International and Wageningen UR make no warranty of any kind, expressed or implied, including without limitation any warranties of merchantability and/or fitness for a particular purpose. NPB, Biometris, Plant Research International and Wageningen UR do not assume any liability for the use of this software. In no event will NPB, Biometris, Plant Research International or Wageningen UR be liable for any additional damages, including any lost profits, lost savings, or other incidental or consequential damages arising from the use of or inability to use, this software and its accompanying documentation, even if NPB, Biometris, Plant Research International or Wageningen UR has been advised of the possibility of such damages.

ACKNOWLEDGEMENTS

The research enabling the construction of ASSA has been performed in the framework of strategic research of Biometris and in various consultation projects. The actual writing up was supported by NPB, the Dutch Nature Policy Assessment Office. The stimulating conversations with Harm Houweling of that office constituted vital support. I am grateful to Jacques Withagen of Biometris for code-checking and comparison of ASSA results with results from other software.

REFERENCES

1. A. Saltelli, K. Chan and E. M. Scott. *Sensitivity analysis*. Wiley, Chichester, 2000.
2. W. H. Press, S. A. Teukolsky, W. T. Vettering and B. P. Flannery. *Numerical recipes in C: the art of scientific computing, second edition*. Cambridge University, Cambridge, 1992.
3. M. D. McKay, R. J. Beckman and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21, 239-245, 1979.
4. R. L. Iman and W. J. Conover. A distribution-free approach to inducing rank correlation among input variables, *Communications in Statistics*. Part B, Simulation and Computation, vol 11, 311-334, 1982.
5. I. M. Sobol'. Sensitivity estimates for nonlinear mathematical models, *Matematicheskoe Modelirovanie* 2, 112-118, 1990, (in Russian). Translated in: *Mathematical Modelling and Computational Experiments*, vol 1, 407- 414, 1993.
6. M. J. W. Jansen, W. A. H. Rossing and R. A. Daamen. Monte Carlo estimation of uncertainty contributions from several independent multivariate sources. In: J. Grasman and G. Van Straten (eds.), *Predictability and Nonlinear Modelling in Natural Sciences and Economics*, p334-343, Kluwer, Dordrecht, 1994.

Case Studies in Gaussian Process Modelling of Computer Codes

Marc C. Kennedy, Clive W. Anderson, Stefano Conti and Anthony O'Hagan

Department of Probability and Statistics, The Hicks Building, University of Sheffield, Sheffield S3 7RH, UK

Abstract: In this paper we present a number of recent applications in which an emulator of a computer code is created using a Gaussian process model. Tools are then applied to the emulator to perform sensitivity analysis and uncertainty analysis. Sensitivity analysis is used both as an aid to model improvement and as a guide to how much the output uncertainty might be reduced by learning about specific inputs. Uncertainty analysis allows us to reflect output uncertainty due to unknown input parameters, when the finished code is used for prediction.

The computer codes themselves are currently being developed within the UK Centre for Terrestrial Carbon Dynamics.

Keywords: Bayesian emulator, Sensitivity analysis, Uncertainty analysis, Carbon budget, Vegetation model

1. INTRODUCTION

Complicated physical processes are increasingly studied by means of sophisticated mathematical models implemented within computer codes. Before relying upon the explanatory and predictive abilities of any computer simulation, however, a variety of validatory checks should be carried out.

The practical complications casting most serious doubts on how adequately and realistically a computer model reproduces reality usually arise from: vague or controversial beliefs about the value of some of the code's parameters; availability of limited and/or inaccurate driving data; restrictions due to the CPU cost required for actually running the program; and incomplete representation of reality by the model. In order to identify and attenuate the main sources of uncertainty hampering a program's performance several statistical methods have already been proposed in the classical literature (see [1] for an exhaustive reference).

The Bayesian Perspective

Over the past decade interesting results have been obtained from addressing problems related to computer model uncertainty in a Bayesian fashion. In particular, a convenient and flexible strategy consists in assigning a semi-parametric Gaussian process prior to the program's response; details of the technique can be found e.g. in [2]. Preliminary *emulation* of a code by such means has already been fruitfully exercised on simulators of

Further author information: (Send correspondence to Marc Kennedy
E-mail: m.kennedy@sheffield.ac.uk)

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

nuclear radiation releases [3] and on models for vehicle crash and spot welding [4]. Besides relevant specific findings, results from these case-studies emphasise how widely applicable and enlightening the principle of Gaussian process-based emulation can be. The case studies described in this paper utilise a Bayesian emulator to deal with the problems of: **prediction:** estimation of (possibly functionals of) model outputs at input configurations other than the available ones; **uncertainty analysis:** exploration of the output distribution induced by assigning some probability distribution to uncertain inputs; **screening:** identification of which of the code inputs are significantly *active*, i.e. most influential on the outputs; **sensitivity analysis:** examination of how model outputs react to changes in appropriate inputs; **code verification:** detection of bugs in the actual implementation of the program. These issues relate to the code output. In this paper we will not consider possible discrepancies between the code and real data.

The simplest sensitivity analysis product derived from the emulator is a set of main effect plots [5]. For each of the emulator inputs, these show how the output responds, on average, to changes in that input. Probability distributions must first be specified so that the averaging correctly accounts for input uncertainty.

The Centre for Terrestrial Carbon Dynamics

The Centre for Terrestrial Carbon Dynamics (*CTCD*) is a consortium of British academic and governmental institutions, established to advance scientific understanding of the role played by terrestrial ecosystems in the carbon cycle, with stress on forest ecosystems. CTCD is funded by the Natural Environment Research Council for 5 years as one of its national centres of excellence in earth observation. The ultimate goals of the project are: to gauge carbon fluxes and their uncertainties at different space/time resolutions; to devise methodological, data and instrument advances for reducing these uncertainties; and to deliver relevant findings in accessible formats to the scientific community and to policy makers. These tasks are pursued with the support of a variety of environmental models designed for simulating carbon patterns over different geographical and climatic scenarios. Unfortunately, such models suffer from coarse reproduction of some underlying physical processes and loose connections to driving data.

Within the Centre, Bayesian methods are being employed for the assessment of the relevant model (and data) developments required for reducing the uncertainty around predictions. We present three case studies of the Bayesian approach addressing these challenges. The first in Section 2 illustrates the use of sensitivity analysis for model testing. In Section 3 the emulator is used for a range of analyses including the creation of a simplified upscaled model. The final case study is part of an assessment of uncertainty in the UK carbon budget calculation.

2. CASE STUDY 1: SHEFFIELD DYNAMIC GLOBAL VEGETATION MODEL

The Sheffield Dynamic Global Vegetation Model, daily version (*SDGVMd*) is described in [6]. It is designed to be able to model generic plant functional types over large areas.

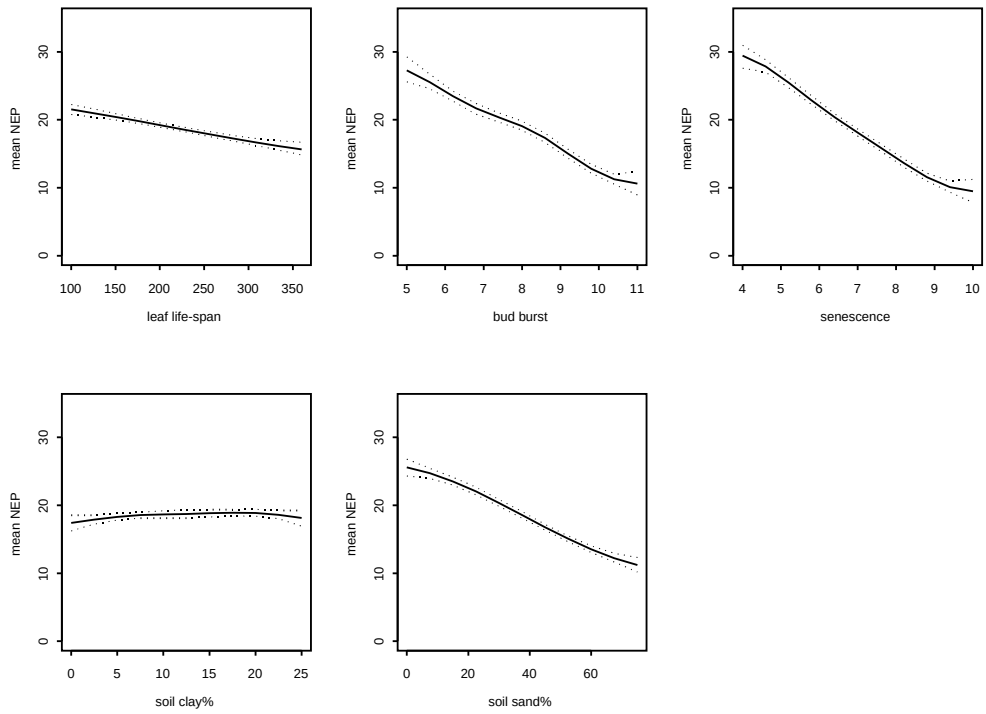


Figure 1. Estimated main effects for SDGVMd inputs. Solid lines represent estimates of the expected output with respect to the unknown input distribution. Dotted lines show 95% pointwise probability bounds for these estimates with respect to the emulator distribution

A variety of extensions and improvements to SDGVMd were undertaken in the first year of CTCD's operation. Simple sensitivity analysis exercises were designed to identify problems with the evolving code.

The five relevant soil and plant inputs that were considered at this stage were: *Leaf life span*, *bud burst temperature*, *senescence temperature*, *soil sand content* (%) and *soil clay content* (%). These were selected after talking with plant scientists following a preliminary sensitivity study. The plant scientists also provided a range of values for these inputs, that were plausible for a deciduous broadleaf plant type. An 80-point maximin latin hypercube was generated in the resulting input space and for each point the average was computed over 100 years for the principal model output (net ecosystem productivity, or *NEP*). A number of coding errors were uncovered during this process, because the code had not been exercised for such varied combinations of input.

Plots of main effects (Figure 1) proved a cheap and effective confirmatory tool for the model developers. They clearly show which of the considered inputs NEP output is significantly sensitive to, and the nature of the various input/output relationships. In calculating the main effects, uniform probability distributions were assumed for these inputs based on the given ranges, while the remainder were fixed at suggested default

values. The plots show that NEP is generally a decreasing function of leaf life span. This goes against the intuition that if leaves live longer they should be able to absorb more carbon, and led the model developers to investigate the phenology routine more closely. They found that a short life span was leading to multiple short growing seasons during the year, and hence higher NEP. A more realistic phenology algorithm has since been developed, and the main effect for leaf life span seen in subsequent sensitivity studies is more realistic (see Section 4). The modellers were satisfied with the relationships revealed by the other plots. Increasing the temperatures of budburst or senescence effectively eats into the growing season at either end of the year, thus reducing total photosynthesis. As expected, these temperatures are critical parameters and effort has been made within CTCD to obtain good phenology information. Output is sensitive to the value of the sand content, but not to clay content over this range. It is clearly important, therefore, to obtain accurate soil sand content data.

3. CASE STUDY 2: SOIL PLANT ATMOSPHERE MODEL

The soil-plant-atmosphere (SPA) model [7] is a detailed model of plant processes operating at a 30 minute time step. It therefore requires 30 minute driver variables in order to run.

3.1. The Aggregated Canopy Model (ACM)

In practice, predictions are required at a coarser temporal scale using a much more restricted set of input data. One solution to this problem is to build a simplified model at the coarse scale by aggregating model output from the fine scale model, and then fitting simpler functional forms to the resulting input/output data set. This approach is described in [8] and can be summarised as follows: (1) Generate 6561 points in the space of 9 daily inputs; (2) Disaggregate each of these daily points into 30-minute time series data; (3) Run SPA with the 30-minute data to produce 6561 daily GPP outputs; (4) Fit a simpler response surface to the daily input and output points.

The resulting aggregated-canopy model (ACM) is a "big-leaf" model of daily gross primary production (GPP) with 9 inputs. The model is much simpler and faster than SPA, requiring daily driving data. are listed in Table 1 with their minimum and maximum values. The target output is aggregate GPP for the given day. Motivated by an earlier investigation [8], a variety of analyses have been performed on ACM and SPA.

3.2. Emulating SPA

The following analysis arises from the recognition that ACM is a kind of emulator of SPA, designed to operate using daily meteorological driving data, when the 30 minute data required by SPA are not available. We expect to meet similar extrapolation problems when applying the more global scale SDGVMd outside the relatively data-rich region of Northern Europe. It was therefore a useful exercise to employ Gaussian process emulation to provide an alternative approximation for the upscaled SPA using far fewer runs.

In the current example we were not able to run the code directly. The following simple algorithm was used to select a subset of 150 points from the 6561 SPA runs already available from the ACM fitting procedure.

Table 1. Input parameters with valid ranges

Input	Symbol	Min.	Max.
Day of year	D	173	230
Leaf Area Index (m^2/m^2)	L	0.1	2.5
Mean foliar N concentration (g N/m^2 leaf are)	N_f	0.32	4.54
Mean daily temperature (°C)	T_m	3	20
Half daily temperature range (°C)	T_{hr}	1	8
Irradiance ($MJ\cdot\ m^{-2}\cdot\ d^{-1}$)	I	4.5	30.6
Leaf water-soil water potential difference (MPa)	Ψ_d	-2.5	-0.5
Ambient CO_2 concentration (μmol/mol)	C_a	173	230
Leaf hydraulic conductance ($mmol\cdot m^{-2}\cdot s^{-1}\cdot MPa^{-1}$)	K_l	0.1	3.0

1. Generate a 150 point maximin Latin hypercube design (D1) in 9 dimensions, with input ranges matching those seen in the SPA run data (Table 1).

2. For each point in D1, select the closest matching point in the big design (excluding those already selected) and add it to the emulator training data.

The emulator can now be used instead of ACM to carry out prediction, uncertainty analysis and sensitivity analysis.

3.2.1. Prediction

The 6411 SPA runs not used to build the emulator are available to test the prediction accuracy of the emulator against that of ACM. The emulator root mean squared error (RMSE) was 0.314, compared with RMSE=0.726 for the ACM. Predicted versus true values of the aggregated SPA output are plotted in figures 2 for both ACM and emulator predictions. Clearly the emulator has smaller errors overall, but not for all regions of the input space. The emulator predicts some small GPP values as being negative. This is physically impossible, and for these values ACM is more accurate because this knowledge is built into the ACM equations. We could of course modify the emulator output so that negative values are set to 0.

As a diagnostic check, we plot the $t_{140}(0, 1)$ Q-Q plot of standardised errors in Figure 3. Most of the points are on the line, indicating that overall the posterior variances are consistent with actual errors. Deviations from the line cast some doubt about the distributional assumptions. In particular, the stationarity assumption may be questionable here.

3.2.2. Sensitivity analysis

Main effects for the emulator inputs are plotted in Figure 4. We assume independent uniform distributions for the inputs according to the ranges in Table 1. The method used in [9] provides an estimate of the uncertainty of the output resulting from the input uncertainty, and a breakdown of the contribution to this uncertainty from each input.

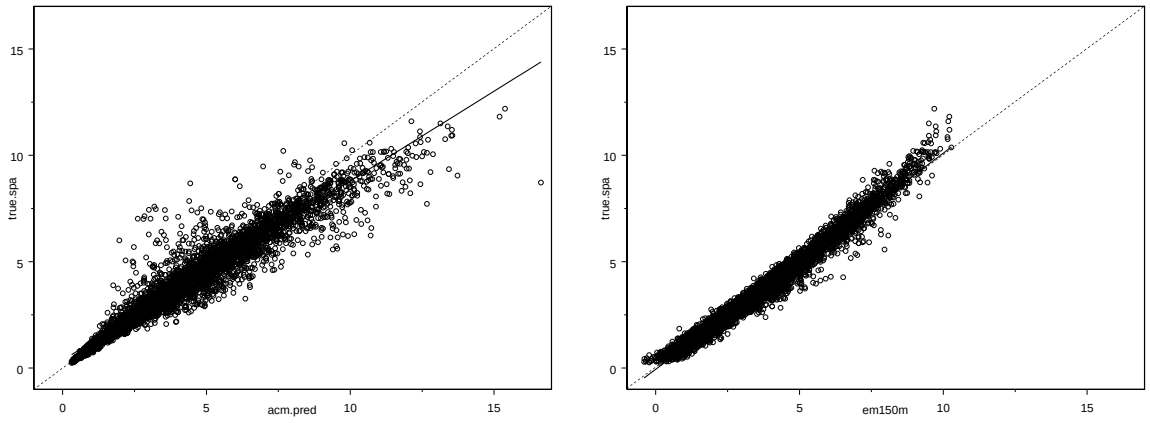


Figure 2. Fitted versus actual values of the aggregated SPA runs: on the left using ACM and on the right using the emulator. 1:1 lines are dashed, regression lines are solid

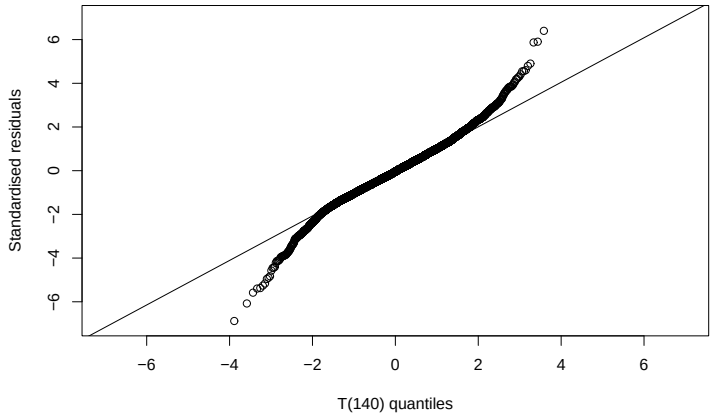


Figure 3. Quantile-Quantile plot for standardised residuals

The total variance is 3.44, and the percentage contributions to this variance from each input are N_f (41.08%), D (18.96%), L (8.63%), I (7.34%), C_a (4.87%), T_m (4.27%), Ψ_d (0.67%), K_l (0.53%), T_{hr} (0.38%). The remaining 13.27% is due to joint effects and higher order interaction effects. These results are consistent with the findings given in [8], yet were obtained in a much simpler way using far fewer runs of SPA.

3.2.3. Uncertainty analysis

Uncertainty analysis is concerned with quantifying the uncertainties in predictions that arise because one or more of the code inputs are unknown. As an example, consider the prediction of GPP at a single site on a given day (site 7000, day 200). Values are available from a data file for each of the required inputs and driving data. The ACM prediction assuming these inputs are exactly known is 3.59. Now suppose that just 1 of these inputs,

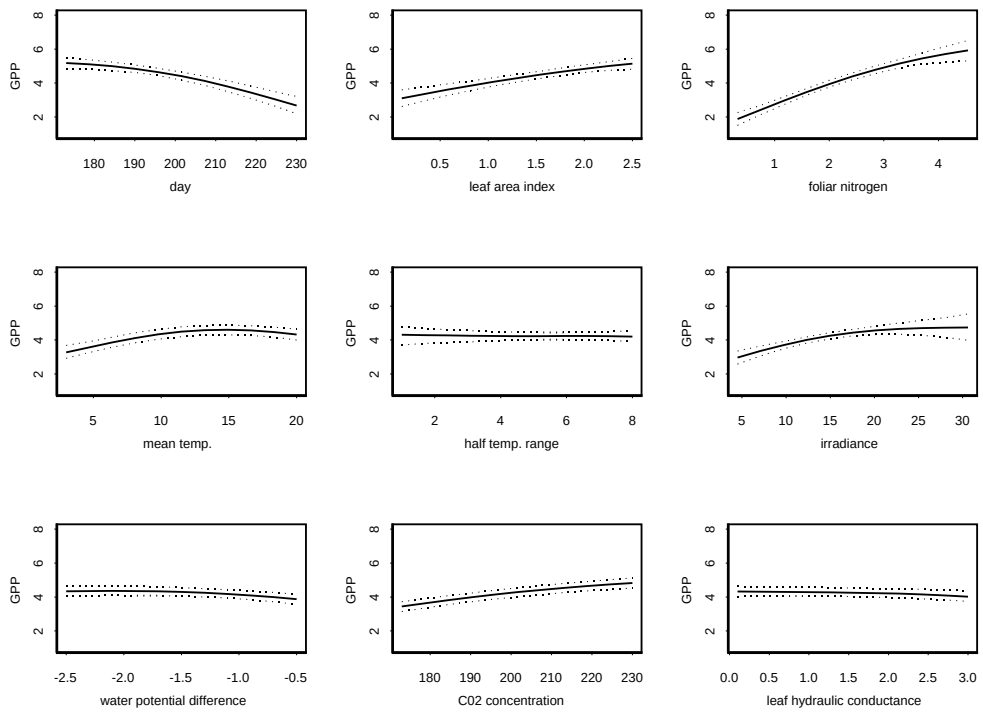


Figure 4. Main effects for each of the input parameters. The solid line is the median. The dashed lines correspond to the 95% point-wise probability band

the irradiance, is uncertain with a $N(15.08, 9)$ distribution. The value 15.08 is the value given in the data files, and a variance of 9 was chosen to match the distribution of errors in irradiance prediction (Figure 2 of [8]). A simple method of propagating this uncertainty is to use a Monte Carlo uncertainty analysis. Running ACM for each of 500 irradiance inputs sampled from this distribution (with all other inputs fixed) produces a sample from the 'true' uncertainty distribution of the GPP output of ACM, which we can obtain in this case only because runs of ACM are essentially instantaneous. The uncertainty distribution has mean 3.56 and variance 0.052.

By comparison, the emulator prediction assuming the inputs are all known is 3.33 (with variance 0.02 due to emulator uncertainty). The emulator prediction assuming a $N(15.08, 9)$ distribution for irradiance is 3.24 (with variance 0.018 due to emulator uncertainty). The variance of the prediction is estimated as 0.15. The conclusion we draw from this is that the ACM is overpredicting the output mean and underpredicting the output variance.

4. CASE STUDY 3: UNCERTAINTY IN THE UK CARBON BUDGET

A major deliverable of CTCD will be an estimate of the UK carbon budget, in April 2004, using SDGVMd. We will quantify uncertainty on the UK carbon budget using Bayesian

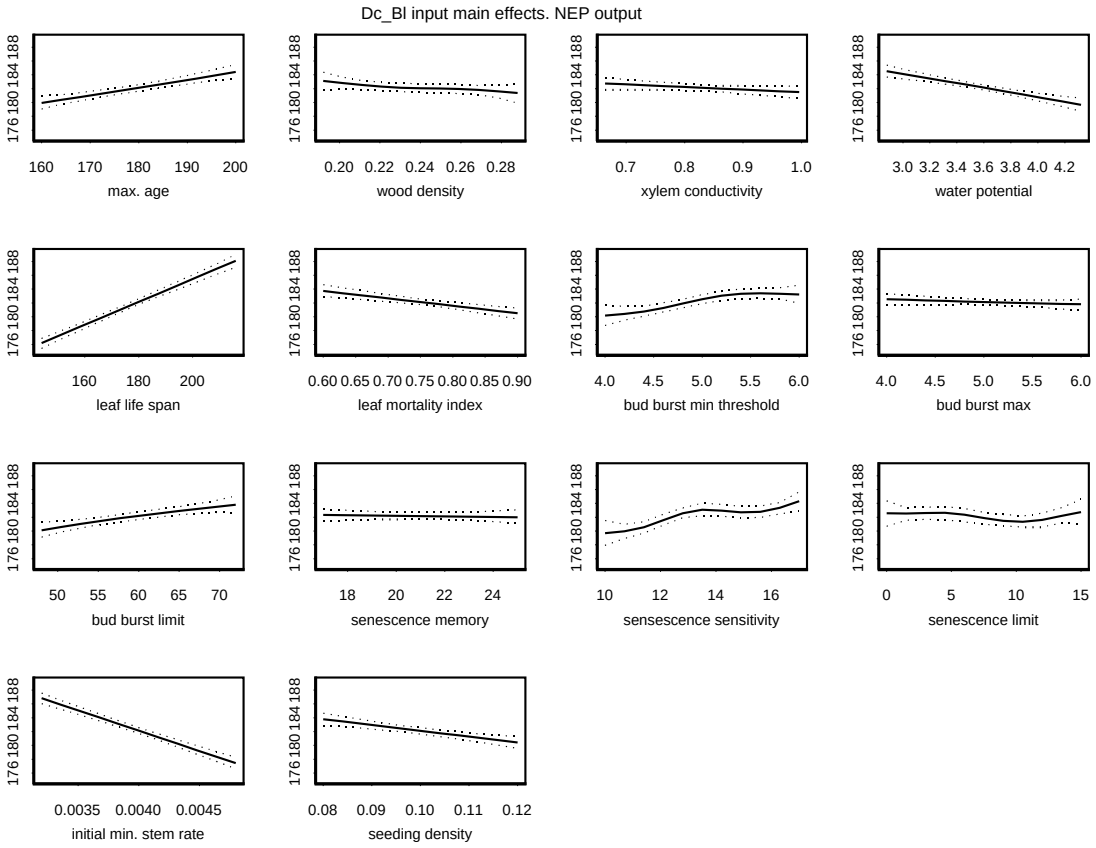


Figure 5. Main effects for each of the input parameters. The solid line is the median. The dashed lines correspond to the 95% point-wise probability bands. The test was carried out using a central UK site and a set of input ranges appropriate for a deciduous broadleaf tree

methods, recognising uncertainty in major model parameters defining vegetation and soil properties. Since SDGVMd is a point model, the first step is to consider uncertainties at individual sites. Nine sites were selected to be representative of the varied climatic conditions in the UK.

The code used here is a more developed version of the one described in Section 2. We performed a more extensive sensitivity study, this time to identify the inputs that would potentially contribute most to the output uncertainty. Figure 4 shows the results from an assessment of 14 plant functional type inputs. Using the same variance decomposition technique as in Section 3.2.2, the most important inputs were found to be leaf life span (days), initial minimum stem rate (millimetres), maximum age (years) and water potential (M Pa). Plant modelling experts were then questioned on their beliefs about these inputs to elicit probability distributions. Different plant functional types were believed to have different probability distributions for some inputs. Each site represents an area covering $10km^2$, so the distributions also account for the fact that multiple species are likely to be represented.

Table 2. Uncertainty analysis results for NEP at the 9 test sites for a deciduous broadleaf plant functional type. The values in parentheses are variances of the mean estimate due to the emulator. Plug-in estimates are obtained by running SDGVMd with input values fixed at their means

Site	Output mean	Output variance	plug-in estimate
S. Ballater (Scotland)	78.10 (1.59)	210.20	89.31
Kielder	65.85 (3.77)	239.73	43.5
New Forest (Hampshire)	207.23 (4.97)	1133.78	269.23
Dartmoor	64.88 (7.63)	472.93	99.93
Lowland (Scotland)	66.35 (7.26)	418.42	73.34
E. Keswick (Lake District)	45.38 (2.56)	183.39	55.19
Barnstaple	137.52 (3.31)	785.95	162.02
Milton Keynes	217.48 (11.54)	494.11	228.43
Stockten on the Forest (Nr York)	218.86 (2.35)	241.39	234.84

Maximum age was agreed as having a $N(180, 100)$ distribution for all types. Leaf life span was agreed as having a $N(200, 625)$ distribution for deciduous types, $N(1500, 10000)$ for evergreen needleleaf and $N(1200, 10000)$ for evergreen broadleaf types. The logarithm of the minimum stem rate was assigned a $N(\ln 0.006, (0.5 \ln 1.5)^2)$ distribution for a broadleaf type and a $N(\ln 0.0015, (0.5 \ln 1.5)^2)$ distribution for a needleleaf type. Water potential was agreed to be distributed as $N(3, 0.25)$ for deciduous broadleaf types, $N(4, 0.25)$ for evergreen broadleaf types, and $N(3.5, 0.25)$ for both needleleaf types. A realistic distribution for the leaf mortality index input has yet to be determined. The sensitivity study was repeated with the refined distributions and ranges to see if anything new would show up. At this point *seeding density* emerged as a significant input.

An uncertainty analysis was carried out at each of the 9 sites, yielding estimates for the mean and variance of NEP output averaged over the decade 1991–2000. Results are given in Table 2. Even after accounting for uncertainty in the emulator, the output means differ noticeably from the plug-in estimates, suggesting non-linearity. In all but Kielder, the plug-in values are overestimating the mean output. We recognise these variances will be underestimates if any of the key inputs, such as seeding density and leaf mortality index are artificially assumed to be fixed or given the wrong distribution. Plant scientists have so far been unable to specify distributions for these inputs, but the process described above has clearly identified these as issues to be resolved by further research. Our results also suggest that the different sites can yield different sets of key inputs, and the process of eliciting prior distributions from the plant scientists will need to be repeated until all uncertainties are accurately represented.

5. CONCLUSIONS

The proposed Bayesian approach to computer experimentation has already supplied useful insights to CTCD modellers and is expected to yield profitable responses when applied to more demanding test beds. Uncertainty and sensitivity analyses will be integral parts of all major CTCD deliverables. The efficiency of the emulator was clearly demonstrated in the case of the aggregated SPA model, where greater accuracy was achieved using only a fraction of the code run data used to derive ACM. Identifying the most significant uncertainty sources will help determine how best to focus future resources in order to reduce overall uncertainty.

ACKNOWLEDGMENTS

The authors are grateful to Mark Lomas, Mat Williams and their colleagues for assistance in running the computer models, and to Ian Woodward for probability distributions on the input parameters to SDGVMd.

The project is funded by the Natural Environment Research Council, contract number F14/G6/105.

REFERENCES

1. A. Saltelli, K. Chan, and E. M. Scott, editors. *Sensitivity analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2000.
2. J. E. Oakley and A. O'Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4):769–784, 2002.
3. M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(3):425–464, 2001.
4. M. J. Bayarri, J. O. Berger, D. Higdon, M. C. Kennedy, A. Kottas, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C. H. Lin, and J. Tu. A framework for validation of computer models. In D. Pace and S. Stevenson, editors, *Foundations for Verifications and Validation in the 21st Century*, 2002.
5. A. Saltelli, S. Tarantola, and F. Campolongo. Sensitivity analysis as an ingredient of modeling. *Statist. Sci.*, 15(4):377–395, 2000.
6. M. R. Lomas, F. I. Woodward, and S. Quegan. The role of dynamic vegetation models. Technical report, University of Sheffield, Sheffield UK, 2002.
7. M. Williams, E. B. Rastetter, D. N. Fernandes, M. L. Goulden, S. C. Wofsy, G. R. Shaver, J. M. Melillo, J. W. Munger, S. M. Fan, and K. J. Nadelhoffer. Modelling the soil–plant–atmosphere continuum in a *Quercus–Acer* stand at Harvard Forest: the regulation of stomatal conductance by light, nitrogen and soil/plant hydraulic properties. *Plant, Cell and Environment*, 19:911–927, 1996.
8. M. Williams, E. B. Rastetter, G. R. Shaver, J. E. Hobbie, E. Carpino, and B. L. Kwiatkowski. Primary production of an arctic watershed: an uncertainty analysis. *Ecological Applications*, 11(6):1800–1816, 2001.
9. J. E. Oakley and A. O'Hagan. Probabilistic sensitivity analysis of complex models: a Bayesian approach. Technical report, Department of Probability and Statistics, University of Sheffield, Sheffield UK, 2002.

Nonparametric Prior Elicitation

Jeremy Oakley and Anthony O'Hagan

Deparment of Probability and Statistics, University of Sheffield, The Hicks Building, Hounsfield Road, Sheffield, S3 7RH, UK;

E-mail: j.oakley@sheffield.ac.uk

1. INTRODUCTION

Global sensitivity analysis is recognized as an essential tool for investigating the effects of input parameter uncertainty in a complex model. To obtain meaningful results from a sensitivity analysis, it is important that the probability distributions for all the uncertain input parameters in the model accurately represent the beliefs of the model user or decision-maker. When little or no data related to these parameters are available, parameter distributions must be specified largely on the basis of expert knowledge. This is rarely a simple task.

A particular difficulty in this scenario is that to perform the global sensitivity analysis the full joint probability distribution is required for all the uncertain input parameters in the model. However, a full probability distribution implies an infinite number of probability judgments by the expert about the parameters, clearly something the expert is unable to provide. In practice it is only going to be possible to elicit a finite and typically small number of probability statements from the expert. These statements will typically take the form quantiles of the distribution, perhaps the mode and sometimes the mean or other moments. Such statements are not enough to identify the expert's probability distribution uniquely, and the usual approach is to fit some member of a convenient parametric family. There are two clear deficiencies in this solution. First, the expert's beliefs are forced to fit the parametric family. The parametric family may imply additional beliefs about the

Further author information:

J.O.: Telephone: +44 (0)114 222 3853

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

parameters that the expert does not agree with. Second, no account is then taken of the many other possible distributions that might have fitted the elicited statements equally well. This clearly has consequences for a global sensitivity analysis; other distributions might produce very different results when the uncertainty they are describe is propagated through the computer model under investigation.

We present an approach which tackles both of these deficiencies. Our model is non-parametric, allowing the expert's distribution to take any continuous form. It also quantifies the uncertainty in the resulting elicited distribution. Formally, the expert's density function is treated as an unknown function, about which we make inference. The result is a posterior distribution for the expert's density function. The posterior mean serves as a 'best fit' elicited distribution, while the variance around this fit expresses the uncertainty in the elicitation.

Specifically, this is achieved by using a Gaussian process to describe our own beliefs about the expert's distribution. Our prior specification contains proper prior beliefs about the smoothness of the expert's distribution, but is ultimately vague in that we do not include any of our own beliefs about likely values of the uncertain input parameter. Data then comes in the form of the expert's summaries, such as their mean and various quantiles. Properties of Gaussian processes can then be exploited to update our beliefs about the expert's distribution analytically, conditional on various hyperparameters in our Gaussian process model. Finally, Markov Chain Monte Carlo methods are used to remove the conditioning on these hyperparameters to give a full, probabilistic description of our uncertainty about the expert's distribution.

Illustrations of our method are given using some simple real elicitation exercises.

2. THE ELICITATION METHOD.

Here we give a brief overview of the elicitation method. Full details can be found in [1]. The idea is to think of eliciting a prior distribution as a standard problem in Bayesian inference. We, the analyst, wish to make inferences about an unknown function $f(\theta)$, the expert's prior density function for θ. We first formulate our own prior beliefs about

$f(\theta)$. We then ask the expert for probability judgments about θ which we think of as data about $f(\theta)$. We then update our beliefs about $f(\theta)$ in light of this data.

2.1. A prior distribution for $f(\theta)$

We assume that the analyst's prior beliefs about $f(\theta)$ can be represented by a Gaussian process. In particular, the analyst's prior distribution for any finite set of points on this function is multivariate normal. Gaussian process priors for functions have been proposed in various different settings, including regression [2] and [3], classification [3] and numerical analysis [4].

The Gaussian process is specified by giving its mean function and variance-covariance function. We will model these hierarchically in terms of a vector α of hyperparameters. First let the analyst's prior expectation of $f(\theta)$ be some member $g(\theta \,|\, u)$ of a suitable parametric family with parameters u. Thus

$$E\{f(\theta) \,|\, \alpha\} = g(\theta \,|\, u). \tag{1}$$

Now it would not be realistic to suppose that the variance of $f(\theta)$ would be the same for all θ. In general, where the analyst expects $f(\theta)$ to be smaller his prior variance should be smaller in absolute terms. We reflect this in our model by supposing that the variance-covariance function has the scaled stationary form

$$Cov\{f(\theta), f(\phi) \,|\, \alpha\} = g(\theta \,|\, u)\, g(\phi \,|\, u)\, \sigma^2 c(\theta, \phi), \tag{2}$$

where $c(\theta, \phi)$ is a correlation function that takes the value 1 at $\theta = \phi$ and is a decreasing function of $|\theta - \phi|$. In general, the function $c(.,.)$ must ensure that the prior variance-covariance matrix of any set of observations of $f(.)$ (or functionals of $f(.)$) is positive semi definite. Here we choose the function

$$c(\theta, \phi) = \exp\{-\frac{1}{2b}(\theta - \phi)^2\}. \tag{3}$$

This will be seen to be a mathematically convenient choice, and implies that $f(.)$ is infinitely differentiable with probability 1.

This formulation was given in [5], who were interested in quadrature for computationally expensive density functions.

Our model represents a belief that the expert's density function $f(\theta)$ will, to some extent, approximate to a member of the parametric family $g(\theta \mid u)$. However, the model is nonparametric and allows the true $f(\theta)$ to have any form at all. The hyperparameter σ^2 specifies how close the true density function will be to its prior mean, and so governs how well it approximates to the parametric family. The hyperparameter b controls the smoothness of the true density. If b is large, then two points $f(\theta)$ and $f(\phi)$ will be highly correlated even if θ and ϕ are far apart.

The hyperparameters of this model are $\alpha = (u, \sigma^2, b)$. Non-informative priors are given for u and σ^2, and a (proper) lognormal prior is assumed for the ratio b/v (reflecting a belief that the expert's density will be smooth over the range implied by v)

2.2. Prior to posterior updating

Data will come in the form of quantiles of the distribution and simple moments. Conditional on α, the posterior distribution of $f(\theta)$ can be derived analytically. Markov Chain Monte Carlo sampling is then used to remove the dependence on α (details in [1]

3. EXAMPLE

We will illustrated the method with a simple synthetic example. Suppose that the expert has the following density function for θ:

$$f(\theta) = \frac{0.4}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(\theta + 2)^2\} + \frac{0.6}{\sqrt{4\pi}} \exp\{-\frac{1}{4}(\theta - 1)^2\}. \tag{4}$$

It is further assumed that the expert can state $P(\theta < x)$ for any x. The expert is asked to give probabilities for the following x: {-3,-2,-1,0,1,2,3}. These probability judgments constitute our data. We do not ask for the mean in this example, though we assume that the expert has given us $P(\theta < \infty) = 1$.

We now use MCMC to sample from the posterior distribution of $\alpha(u, \sigma^2, b)$. For each sampled α, we generate one random density function from the conditional distribution of

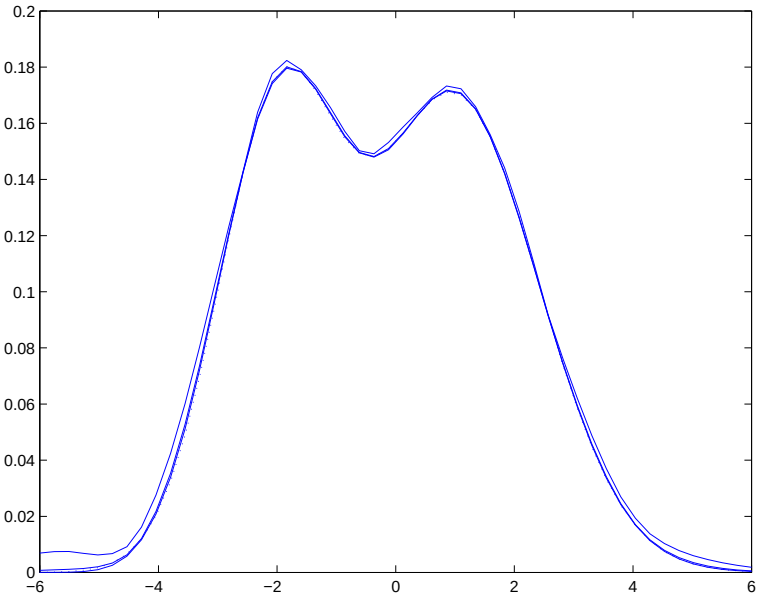


Figure 1. The mean and pointwise 95% intervals for the expert's density function (solid lines), and the true density function (dotted line).

$f(\theta)$ given α and the data. Given that we must have $f(\theta) > 0$, we discard any generated density functions that are negative over the range of interest. We then plot the pointwise mean, 2.5th and 97.5th percentiles from the distribution of the density function.

In figure 1 we can see that, without specifying a bimodal density function $f(\theta)$ in our prior for $f(\theta)$, we have correctly recovered the bimodal shape. We are also able to report our remaining uncertainty about $f(\theta)$ after eliciting the seven percentiles, which in this case is small.

References

[1] J. E. Oakley and A. O'Hagan. Uncertainty in prior elicitations: a nonparametric approach. Technical report, Department of Probability and Statistics, University of Sheffield, 2003.

[2] A. O'Hagan. Curve fitting and optimal design for prediction (with discussion). *J. Roy. Statist. Soc. Ser. B*, 40:1–42, 1978.

[3] R. Neal. Regression and classification using gaussian process priors. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 69–95. Oxford: University Press, 1999.

[4] A. O'Hagan. Some Bayesian numerical analysis. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 345–363. Oxford: University Press, 1992.

[5] M. C. Kennedy and A. O'Hagan. Iterative rescaling for Bayesian quadrature. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 5*, pages 639–645. Oxford: University Press, 1996.

Stochastic Sensitivity Analysis for Computing Greeks

M. Koda, K. Ohmori**, D. Yokomatsu**, and T. Amemiya***

*University of Tsukuba, Tsukuba, Japan, **TGI Network Co. Ltd., Tokyo, Japan
E-mail: * koda@sk.tsukuba.ac.jp

Abstract: In a risk management of derivative securities, sensitivities are important measures of market risk to analyze the impact of a misspecification of some stochastic model on the expected payoff function. We investigate in this paper an application of Malliavin calculus, which enables the computation of sensitivity derivatives, known as Greeks in finance, without resort to a direct differentiation of the complex payoff functions.

Keywords: Stochastic sensitivity analysis, Malliavin calculus, Greeks in finance

1. INTRODUCTION

We consider a stochastic model or, equivalently, a stochastic differential equation in a well-defined framework of Black-Scholes set-up, which is described by

$$S_t = S_0 + \int_0^t rS_\tau d\tau + \int_0^t \sigma S_\tau dW_\tau, \tag{1}$$

where S is the price of underlying asset with S_0 denoting the present (initial) value, r denotes the riskless interest rate, σ the volatility, and $(W_t)_{0\le t\le T}$ is a standard Brownian motion (also known as Wiener process). Note that, in the case of European-type options, we have a closed solution to (1) as follows:

$$S_T = S_0 \exp(\mu T + \sigma W_T), \tag{2}$$

where $\mu = r - \sigma^2/2$ for a fixed expiration or maturity time, T.

We are, in European options, interested in studying how to evaluate the sensitivity with respect to model parameters, e.g., present price S_0, volatility σ, etc., of the expected payoff

$$E[e^{-rT}\Phi(S_T)], \tag{3}$$

for an exponentially discounted value of the payoff function $\Phi(S_T)$, where $E[\square]$ denotes the expectation operator. The sensitivity of more sophisticated payoff functions including path-dependent Asian-type options like

$$E[e^{-rT}\Phi(\frac{1}{T}\int_0^T S_t dt)], \tag{4}$$

may be treated in a similar manner along the lines that will be investigated in the present study. In the Asian option whose payoff functional is defined by (4), we may note that the payoff depends on the average of the asset value in a given period of time.

In finance, this is the so-called model risk problem. Commonly referred to as Greeks, sensitivities in financial market are typically defined as the partial derivatives of the expected payoff function with respect to underlying model parameters. In general, finite difference approximations are heavily used to simulate Greeks by means of Monte Carlo procedures.

Sensitivity Analysis of Model Output
Kenneth M. Hanson and François M. Hemez, eds.
Los Alamos National Laboratory, 2005; http://library.lanl.gov/

However, it is known that the finite difference approximation soon becomes inefficient particularly when payoff functions are complex and discontinuous. This is often the case when we deal with exotic options such as American, lookback, and digital options, etc.

To overcome this difficulty, Broadie and Glasserman [1] proposed a method to put the differential of the payoff function inside the expectation operator required to evaluate the sensitivity. But this idea (i.e., likelihood ratio method) is applicable only when the density of the random variable involved is explicitly known. Recently, Fournie et al. [2] suggested the use of Malliavin calculus, by means of integration by parts, to shift the differential operator from the expected payoff to the underlying diffusion (e.g., Gaussian) kernel, introducing a weighting function.

The real advantage of using Malliavin calculus is that it is applicable when we deal with random variable whose density is not explicitly known as the case of Asian options. Another examples which are similar to the present study and explored by the first author (e.g., Refs. [3,7]) but that are not covered in this paper are models involving a step function and non-smooth objective functions. In these studies, the stochastic sensitivity analysis technique based on the Novikov's identity is used instead of Malliavin calculus.

In this paper, we present a brief introduction of Malliavin calculus, and describe a constructive approach for a stochastic sensitivity analysis for computing Greeks in financial engineering. The present approach enables the simulation of Greeks without resort to direct differentiation of the complex or discontinuous payoff functions.

The remainder of the paper is organized as follows. In Section 2, we briefly review the essence of Malliavin calculus and present integration by parts formula. In Section 3, we describe a constructive approach. Subsection 3.1 presents some explicit formulae for the case of European option. In Subsection 3.2, we investigate the case of Asian option. In Section 4, we present simulation results obtained for the Asian call option. We conclude in Section 5.

2. MALLIAVIN CALCULUS

Following the standard notations that can be found in [6], we present the most concise introduction of Malliavin calculus necessary to our computation.

Let R be the space of random variables of the form $F = f(W_{t_1}, W_{t_2}, \cdots, W_{t_n})$, where f is smooth and W_t denotes the Brownian motion as before. For a smooth random variable $F \in R$, we can define its *derivative* $DF = D_t F$, where the differential operator D is closable. Since D operates on random variables by differentiating functions in the form of partial derivatives, it shares the familiar chain rule property, $D_t(f(F)) = \nabla f(F) \cdot D_t F = f'(F) D_t F$, and other general properties like linearity, etc.

We denote by D^* the Skorohod integral, defined as the adjoint operator of D. If u belongs to Dom(D^*), then $D^*(u)$ is characterized by the following integration by parts formula:

$$E[FD^*(u)] = E[\int_0^T (D_t F) u_t dt]. \tag{5}$$

It is important to note that (5) gives a duality relationship to link operators D and D^*. The adjoint operator D^* behaves like a stochastic integral. In fact, if u_t is an adapted process, then

the Skorohod integral coincides with the classical Ito integral: i.e., $D^*(u) = \int_0^T u_t dW_t$. If u_t is non-adapted or generic, one has

$$D^*(Fu) = FD^*(u) - \int_0^T (D_t F) u_t dt \,. \tag{6}$$

The property (6) follows directly from the duality relation (5) and the product rule of the operator *D*. A heuristic derivation of (6) is demonstrated here. Let us assume that *F* and *G* are any two smooth random variables, and u_t a generic process, then by product rule of *D* one has

$$E[GFD^*(u)] = E[\int_0^T D_t(GF) u_t dt] = E[\int_0^T G(D_t F) u_t dt] + E[\int_0^T (D_t G) F u_t dt]$$
$$= E[G\int_0^T (D_t F) u_t dt] + E[GD^*(Fu)]$$

which implies that

$$E[GD^*(Fu)] = E[G\left(FD^*(u) - \int_0^T (D_t F) u_t dt\right)]$$

for *any* random variables *G*. Therefore, (6) must hold almost everywhere.

In the present study, we frequently use the following formal relationship to remove the derivative from a (smooth) random function *f* as follows:

$$E[\nabla f(X) Y] = E[f'(X) Y] = E[f(X) H_{XY}], \tag{7}$$

where *X*, *Y*, and H_{XY} are random variables. It is noted that (7) can be deduced from the integration by parts formula (5), and we have an explicit expression for H_{XY} as

$$H_{XY} = D^*\left(\frac{Y}{\int_0^T D_t X dt}\right). \tag{8}$$

If higher order derivatives are involved then one has to repeat the procedure (7) iteratively. It may be noted that H_{XY} is not unique and other expressions than (8) can be also possible. For more details, the readers are referred to Koda et al. [4] and Montero and Kohatsu-Higa [5].

3. CONSTRUCTIVE APPROACH

In this section, utilizing the technical framework of Malliavin calculus introduced in Section 2, a constructive approach is presented to compute Greeks of European and Asian options, respectively.

3.1. European Option

In the case of European option whose payoff function is defined by (3), the essence of the present method is that the gradient of the expected (discounted) payoff, $\nabla E[e^{-rT}\Phi(S_T)]$, is evaluated by putting the gradient inside the expectation, i.e., $E[e^{-rT}\nabla\Phi(S_T)]$, which involves computations of $\nabla\Phi(S_T) = \Phi'(S_T)$ and ∇S_T . Further, applying Malliavin calculus techniques, the gradient is rewritten as $E[e^{-rT}\Phi(S_T)H]$ for some random variable *H*. It should be noted, however, that there is no uniqueness in this representation since we can add to *H* any random variables that are orthogonal to S_T . In general, *H* involves Ito or Skorohod integrals.

3.1.1. Delta

Now we compute *Delta*, Δ, the first-order partial differential sensitivity coefficient of the expected outcome of the option, i.e., (3), with respect to the initial asset value S_0:

$$\Delta = \frac{\partial}{\partial S_0} E[e^{-rT}\Phi(S_T)] = e^{-rT} E[\Phi'(S_T)\frac{\partial S_T}{\partial S_0}] = \frac{e^{-rT}}{S_0} E[\Phi'(S_T)S_T]$$

Then, with $X = Y = S_T$ in (7), we perform the integration by parts applying (8) to give

$$\Delta = \frac{e^{-rT}}{S_0} E\left[\Phi(S_T)H_{XY}\right] = \frac{e^{-rT}}{S_0} E\left[\Phi(S_T)D^*\left(\frac{S_T}{\int_0^T D_t S_T dt}\right)\right], \tag{9}$$

which removes the derivative of Φ from the expectation as desired.

Since the integral term in the denominator that appears in (9) can be computed as $\int_0^T D_t S_T dt = \sigma T S_T$, we can evaluate the stochastic integral involved in (9) as

$$H_{XY} = D^*\left(\frac{S_T}{\int_0^T D_t S_T dt}\right) = D^*\left(\frac{1}{\sigma T}\right) = \frac{D^*(1)}{\sigma T} = \frac{W_T}{\sigma T}$$

with the help of (6) applied to u=1 (a constant process which is adapted and Ito integral yields $D^*(1) = W_T$). Then the final expression for Δ reads

$$\Delta = \frac{e^{-rT}}{\sigma T S_0} E[\Phi(S_T)W_T]. \tag{10}$$

We may note that when we deal with European options, the present result (10) coincides with the result that is obtained by the explicit computation of the closed formula for the probability density function of S_T.

3.1.2. Vega

Next Greek *Vega*, *V*, is the index that measures sensitivity of the expected payoff (3) with respect to the volatility σ, which can be computed as

$$V = \frac{\partial}{\partial \sigma} E[e^{-rT}\Phi(S_T)] = e^{-rT} E[\Phi'(S_T)\frac{\partial S_T}{\partial \sigma}] = e^{-rT} E[\Phi'(S_T)S_T\{W_T - \sigma T\}],$$

where we have used the solution (2) to evaluate $\partial S_T / \partial \sigma$. Then, utilizing (7) and (8) again with $X = S_T$ and $Y = S_T(W_T - \sigma T)$, we apply the integration by parts to give

$$V = e^{-rT} E\left[\Phi(S_T)H_{XY}\right] e^{-rT} E\left[\Phi(S_T)D^*\left(\frac{S_T(W_T - \sigma T)}{\int_0^T D_t S_T dt}\right)\right] = e^{-rT} E\left[\Phi(S_T)D^*\left(\frac{W_T}{\sigma T} - 1\right)\right].$$

So, we evaluate the stochastic integral as

$$H_{XY} = D^*\left(\frac{W_T}{\sigma T} - 1\right) = \frac{1}{\sigma T} D^*(W_T) - D^*(1) = \frac{1}{\sigma T} D^*(W_T) - W_T \ .$$

With the help of (6) applied to $u = 1$ (adapted process) and $F = W_T$, we have

$$D^*(W_T) = W_T^2 - \int_0^T D_t W_T \, dt = W_T^2 - \int_0^T 1 dt = W_T^2 - T \ .$$

If we bring together the partial results obtained above, we derive the final expression

$$V = e^{-rT} E\left[\Phi(S_T)\left\{\frac{W_T^2}{\sigma T} - W_T - \frac{1}{\sigma}\right\}\right]. \tag{11}$$

3.1.3. Gamma

The last Greek *Gamma*, Γ, involves a second-order derivative,

$$\Gamma = \frac{\partial^2}{\partial S_0^2} E[e^{-rT}\Phi(S_T)] = \frac{e^{-rT}}{S_0^2} E[\Phi''(S_T) S_T^2] \, .$$

Utilizing (7) and (8) with $X = S_T$ and $Y = S_T^2$, we obtain after a first integration by parts

$$\Gamma = \frac{e^{-rT}}{S_0^2} E\left[\Phi'(S_T) D^*\left(\frac{S_T^2}{\int_0^T D_t S_T \, dt}\right)\right] = \frac{e^{-rT}}{S_0^2} E\left[\Phi'(S_T) D^*\left(\frac{S_T}{\sigma T}\right)\right] \ .$$

With the help of (6) applied to $u = 1/\sigma T$ (constant adapted process) and $F = S_T$, we have

$$D^*\left(\frac{S_T}{\sigma T}\right) = \frac{S_T}{\sigma T} D^*(1) - \frac{1}{\sigma T} \int_0^T D_t S_T \, dt = S_T\left(\frac{W_T}{\sigma T} - 1\right).$$

Then, repeated application of (7) and (8) with $X = S_T$ and $Y = S_T(W_T / \sigma T - 1)$, the second integration by parts yields

$$\Gamma = \frac{e^{-rT}}{S_0^2} E\left[\Phi'(S_T) S_T\left(\frac{W_T}{\sigma T} - 1\right)\right] = \frac{e^{-rT}}{S_0^2} E\left[\Phi(S_T) D^*\left(\frac{S_T}{\int_0^T D_t S_T \, dt}\left\{\frac{W_T}{\sigma T} - 1\right\}\right)\right].$$

With the help of (6) as before, we can evaluate the stochastic integral as

$$D^*\left(\frac{S_T}{\int_0^T D_t S_T \, dt}\left\{\frac{W_T}{\sigma T} - 1\right\}\right) = \frac{1}{\sigma T} D^*\left(\frac{W_T}{\sigma T} - 1\right) = \frac{1}{\sigma T}\left\{\frac{W_T^2}{\sigma T} - W_T - \frac{1}{\sigma}\right\}.$$

If we combine the results obtained above, the final expression becomes

$$\Gamma = \frac{e^{-rT}}{\sigma T S_0^2} E\left[\Phi(S_T)\left\{\frac{W_T^2}{\sigma T} - W_T - \frac{1}{\sigma}\right\}\right]. \tag{12}$$

Comparing (12) with (11), we find the following relationship between V and Γ:

$$\Gamma = \frac{V}{\sigma T S_0^2} \, . \tag{13}$$

Since we have closed solutions for all the Greeks, we can easily check the correctness of the above results.

3.2. Asian Option

In the case of Asian option whose payoff functional is defined by (4), the essence of the present approach is again that the gradient of the expected (discounted) payoff is rewritten as $E[e^{-rT}\nabla\Phi(\frac{1}{T}\int_0^T S_t dt)] = e^{-rT}E[\Phi(\frac{1}{T}\int_0^T S_t dt)H]$, for some random variable H. Different from the European options, however, we do not have a known closed formula in this case.

3.2.1. Delta

Delta in this case is given by

$$\Delta = \frac{\partial}{\partial S_0}E[e^{-rT}\nabla\Phi(\frac{1}{T}\int_0^T S_t dt)] = \frac{e^{-rT}}{S_0}E[\Phi'(\frac{1}{T}\int_0^T S_t dt)\frac{1}{T}\int_0^T S_t dt].$$

There are various ways of performing the integration by parts; e.g., the readers are referred to [2]. In the present approach, utilizing (7) and (8) with $X = Y = \int_0^T S_t dt / T$, we may apply the integration by parts to give

$$\Delta = \frac{e^{-rT}}{S_0}E\left[\Phi(\frac{1}{T}\int_0^T S_t dt)D*\left(\frac{Y}{\int_0^T D_t X dt}\right)\right] = \frac{e^{-rT}}{S_0}E\left[\Phi(\frac{1}{T}\int_0^T S_t dt)D*\left(\frac{\int_0^T S_t dt}{\sigma\int_0^T tS_t dt}\right)\right].$$

With the help of (6) applied to $u = 1/\sigma$ (constant adapted process) and $F = \int_0^T S_t dt / \int_0^T tS_t dt$, we may obtain

$$\Delta = \frac{e^{-rT}}{S_0}E\left[\Phi(\frac{1}{T}\int_0^T S_t dt)\left(\frac{1}{<T>}\left\{\frac{W_T}{\sigma} + \frac{<T^2>}{<T>}\right\} - 1\right)\right], \tag{14}$$

where

$$<T> = \frac{\int_0^T tS_t dt}{\int_0^T S_t dt} \quad \text{and} \quad <T^2> = \frac{\int_0^T t^2 S_t dt}{\int_0^T S_t dt}$$

are the first two moments of the probability density defined by $p(t) = S_t / \int_0^T S_t dt$.

3.2.2. Vega

Vega in this case becomes

$$V = \frac{\partial}{\partial\sigma}E[e^{-rT}\nabla\Phi(\frac{1}{T}\int_0^T S_t dt)] = e^{-rT}E[\Phi'(\frac{1}{T}\int_0^T S_t dt)\frac{1}{T}\int_0^T \frac{\partial S_t}{\partial\sigma}dt]$$
$$= e^{-rT}E[\Phi'(\frac{1}{T}\int_0^T S_t dt)\frac{1}{T}\int_0^T S_t\{W_t - t\sigma\}dt]$$

As before, with the help of (7) and (8) applied to $X = \int_0^T S_t dt / T$ and $Y = \int_0^T S_t\{W_T - t\sigma)dt / T$, we have

$$V = e^{-rT} E\left[\Phi(\frac{1}{T}\int_0^T S_t dt) D * \left(\frac{Y}{\int_0^T D_t X dt} \right) \right] = e^{-rT} E\left[\Phi(\frac{1}{T}\int_0^T S_t dt) D * \left(\frac{\int_0^T S_t W_t dt}{\sigma \int_0^T t S_t dt} - 1 \right) \right],$$

which, with the help of (6), yields the following expression:

$$V = e^{-rT} E\left[\Phi(\frac{1}{T}\int_0^T S_t dt) \left\{ \frac{\int_0^T \int_0^T S_t W_t dt dW_\tau}{\sigma \int_0^T t S_t dt} + \frac{\int_0^T t^2 S_t dt \int_0^T S_t W_t dt}{(\int_0^T t S_t dt)^2} - W_T \right\} \right]. \quad (15)$$

Using the relation (13), it is straightforward to compute *Gamma* as (15) divided by $\sigma T S_0^2$.

4. MONTE CARLO SIMULATION OF ASIAN OPTION

In order to evaluate the results obtained in Section 3, we present in this section the results of Monte Carlo simulation for computing Delta and Vega in the case of Asian Call option whose payoff functional is defined by (4).

4.1. Delta

In Fig. 1, we present the simulation result of Δ given by (14) with parameters r=0.1, $\sigma = 0.25$, T=0.2 (in years), and $S_0 = K = 100$ (in arbitrary cash units) where K denotes the strike price. We have divided the entire interval of integration into 252 pieces, representing the approximate number of trading days in a year.

Fig. 1 shows how the outcome of the simulation progressively attains its own value. We compare the convergence behavior of the present simulation with the results obtained by Broadie and Glasserman [1] where all the parameters take the same values we have used, and which may provide most extensive and detailed results currently available. The result indicates a fairy good convergence to the steady-state value that is attained at 10,000*th* iteration stage in [1]. The standard deviation of the simulation in this case was 0.005.

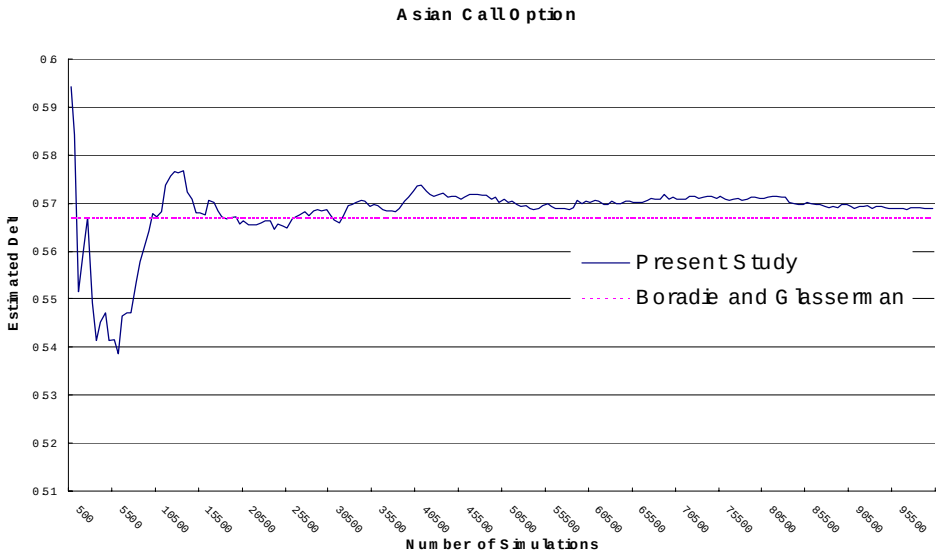


Fig. 1. Estimated *Delta* of Asian Call Option; S_0=K=100, T=0.2, r=0.1, σ=0.25

4.2. Vega

We present in Fig. 2 the result of V given by (15), where all the parameters take the same values we used in the simulation of Δ in Subsection 4.1. Again, we compare the result with the one that is obtained at 10,000*th* iteration stage in [1]. The result indicates that some noticeable bias may remain in the present Monte Carlo simulation, and further study may be necessary to analyze and reduce the bias involved.

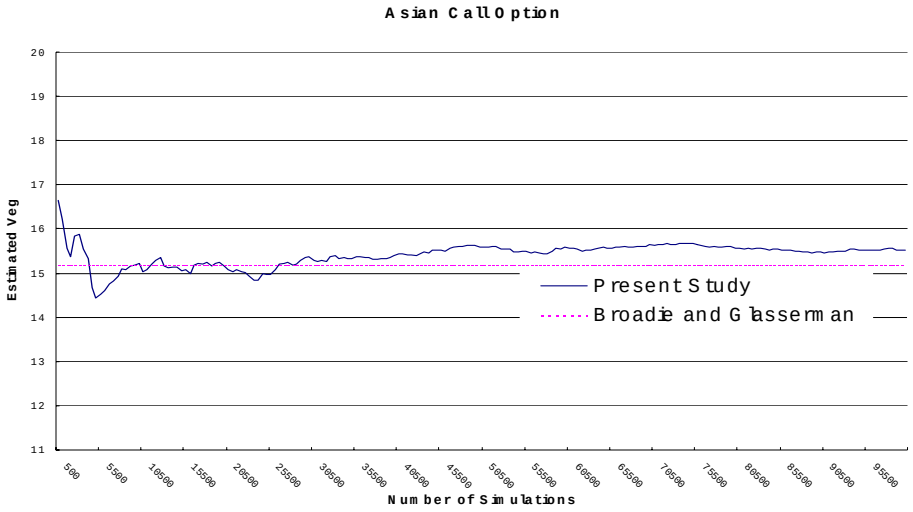


Fig. 2. Estimated *Vega* of Asian Call Option; S_0=K=100, T=0.2, r=0.1, σ=0.25

5. CONCLUSION

We have presented a stochastic sensitivity analysis method, in particular, a constructive approach for computing Greeks in finance using Malliavin calculus. The present approach is useful when the random variables are smooth in the sense of stochastic derivatives. It may be necessary to further investigate and improve Monte Carlo procedures to reduce the bias involved in the simulation of *Vega* in Asian-type options and other sophisticated options.

REFERENCES

1. M. Broadie and P. Glasserman. Estimating security price derivatives using simulation. *Management Science*, 42, 269-285, 1996.
2. E. Fournie, J. M. Lasry, J. Lebuchoux, and P. L. Lions. An application of Malliavin calculus to Monte Carlo methods in Finance II. *Finance and Stochastics*, 5, 201-236, 2001.
3. M. Koda and H. Okano. A new stochastic learning algorithm for neural networks. *Journal of the Operations Research Society of Japan*, 43, 469-485, 2000.
4. M. Koda, A. Kohatsu-Higa, and M. Montero. *An Application of Stochastic Sensitivity Analysis to Financial Engineering.* Discussion Paper Series, No.980, Institute of Policy and Planning Sciences, University of Tsukuba, 2002.
5. M. Montero and A. Kohatsu-Higa. Malliavin calculus applied to finance. *Physica A*, 320, 548-570, 2003.
6. D. Nualert. *The Malliavin Calculus and Related Topics*. Springer, New York, 1995.
7. H. Okano and M. Koda. An optimization algorithm based on stochastic sensitivity analysis for noisy objective landscapes. *Reliability Engineering and System Safety*, 79, 245-252, 2003.

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

Buyung Agusdinata
Delft University of Technology and Faculty of Technology, Policy, and Management
Jaffalaan 5
Delft, 2628 BX
NETHERLANDS
b.agusdinata@tbm.tudelft.nl
Phone: 31-15-2784532
FAX:

Jane Booker
Los Alamos National Laboratory
P.O. Box 1663, MS P946
Los Alamos, NM 87545
USA
jmb@lanl.gov
Phone: 505-667-1479
FAX: 505-665-2137

Cuauhtemoc Aviles-Ramos
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545
USA
aviles@lanl.gov
Phone: 505-665-2657
FAX:

Tom Burr
Los Alamos National Laboratory
P.O. Box 1663, MS F600
Los Alamos, NM 87545
USA
tburr@lanl.gov
Phone: 505-665-7865
FAX:

Lee Balick
Los Alamos National Laboratory
P.O. Box 1663, MS B244
Los Alamos , NM 87545
USA
lbalick@lanl.gov
Phone: 505-665-1012
FAX:

Ulrich Callies
GKSS Research Centre
MAX-PLANCK STR 1
Geesthacht, 21502
GERMANY
ulrich.callies@gkss.de
Phone: +49 4152 87 2837
FAX: +49 4152 87 2818

Jerrell Ballard
U.S. Army ERDC
3909 Halls Ferry Road
Vicksburg, MS 39180
USA
jerry.ballard@erdc.usace.army.mil
Phone: 601-634-2946
FAX:

Katherine Campbell
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos , NM 87545
USA
ksmithcamp@cybermesa.com
Phone: 505-662-3913
FAX:

Nancy Beneda
University of North Dakota
Box 7096 Department of Finance
Grank Forks, ND 58202
USA
nancy.beneda@und.nokak.edu
Phone: 701-777-4690
FAX:

Jonathan Carter
Imperial College London
Department of Earth Science and Engineering
London, SW7 2A
UNITED KINGDOM
j.n.carter@imperial.ac.uk
Phone: +44 20-7594-73-22
FAX:

TotalParticipants 106

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

Flor Castillo
DOW Chemical
,
facastillo@dow.com
Phone:
FAX:

Roger Cooke
Delft University of Technology
Mekelkweg 4
Delft, 2600GA
r.m.cooke@ewi.tu8delft.nl
Phone:
FAX:

James Cavendish
General Motor R&D Center
30500 Mound Rd.
Warren, MI 48090
USA
cavendish@gm.com
Phone: 586-986-1087
FAX: 586-986-0124

Margarida Costa
Dep. Mathematics, Instituto Superior de Gestao
Rua Vitorino Nemesio, n 5
Lisboa, 1750
PORTUGAL
mmcosta@isg.pt
Phone: +351-21-7510053
FAX:

Ismail Celik
West Virginia University
Morgantown, WV 26506-6101
ibcelik@mail.wvu.edu
Phone:
FAX:

Sharon Cummins
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545
USA
scummins@lanl.gov
Phone: 505-667-2350
FAX:

Mike Christie
Heriot Watt University
Institute of Petroleum Engineering
Edinburgh, EH14 4AS
SCOTLAND
mike.christie@pet.hw.ac.uk
Phone: +44 0131 451 8063
FAX: +44 0131 451 8063

Scott Doebling
Los Alamos National Laboratory
P.O. Box 1663, MS T080
Los Alamos, NM 87545
USA
doebling@lanl.gov
Phone: 505-667-8950
FAX: 505-665-6333

Stephano Conti
University of Sheffield
Department of Probability, The Hicks Building
Sheffield, South Yorkshire S3 7RH
UNITED KINGDOM
s.conti@sheffield.ac.uk
Phone: 004 - =(0) 114 2223824
FAX: 0044-(0) 114 2223759

Kevin Dowding
Sandia National Laboratories
P.O. Box 5800, Dept. 9133 MS 0828
Albuquerque, NM
Phone:
FAX:

TotalParticipants 106

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

Scott Ferson
Applied Biomathematics
100 North Country Road
Setauket, NY 11733
USA
scott@ramas.com
Phone: 631-751-4350
FAX: 631-751-3435

Jim Hall
University of Bristol
Department of Civil Engineering, Queens Building
University Walk, Bristol BS8 1TR
UNITED KINGDOM
Jim.Hall@bristol.ac.uk
Phone: 44 117 928 9763
FAX:

Michael Flechsig
Potsdam Institute for Climate Impact Research
Telegrafenberg
Potsdam, D-14473
GERMANY
flechsig@pik-potsdam.de
Phone: ++49-331-288-2604
FAX: ++49-331-288-2600

Michael Hamada
Los Alamos National Laboratory
P.O. Box 1663, MS F600
Los Alamos, NM 87545
USA
hamada@lanl.gov
Phone: 505-667-1843
FAX: 505-667-7440

David Gay
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185
USA
dmgay@sandia.gov
Phone: 505-284-1456
FAX:

Clifford Hansen
Sandia National Laboratories MS 1395
4100 National Parks Highway
Carlsbad, NM 88220
USA
cwhanse@sandia.gov
Phone: 505-234-0103
FAX:

Robert Gover
Naval Research Laboratory
4555 Overlook Ave. S.W. Washington
Washington , D.C. 20375
USA
gover@ninja.nrl.navy.mil
Phone: 202-404-7678
FAX:

Ken Hanson
Los Alamos National Laboratory
P.O. Box 1663, MS D413
Los Alamos , NM 87545
USA
kmh@lanl.gov
Phone: 505-667-1402
FAX:

Todd Graves
Los Alamos National Laboratory
P.O. Box 1663, MS F600
Los Alamos, NM 87545
USA
tgraves@lanl.gov
Phone: 505-667-9186
FAX:

Hanae Hara
Computer Sciences Corporation
4045 Hancock Street
San Diego, CA 92110
USA
hanae@spawar.navy.mil
Phone: 619-553-3803
FAX:

TotalParticipants 106

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

John Helton
Sandia National Laboratories
Dept. 6849, MS 0779
Albuquerque, NM 87185
USA
jchelto@sandia.gov
Phone:
FAX:

Dave Higdon
Los Alamos National Laboratory
P.O. Box 1663, MS F600
Los Alamos, NM 87545
USA
ahigdon@lanl.gov
Phone:
FAX:

Francois Hemez
Los Alamos National Laboratory
P.O. Box 1663, MS T006
Los Alamos, NM 87545
USA
hemez@lanl.gov
Phone: 505-663-5204
FAX: 505-663-5225

Toshimitsu Homma
Japan Atomic Energy Research Institute
2-4, Shirakata Shirane
Tokaimura, Ibaraki 319-1195
JAPAN
homma@popsvr.tokai.jaeri.go.jp
Phone: +81 29 282 6862
FAX: +81 29 282 6147

Michael Hemsch
NASA Langley Research Center
17 W. Taylor St. MS 286
Hampton, VA 23681-2199
USA
michael.j.hemsch@nasa.gov
Phone: 757-864-2882
FAX: 757-864-8192

Henry Hsieh
Lawrence Livermore National Laboratory
700 East Avenue
Livermore, CA 94551
USA
hsieh6@llnl.gov
Phone: 925-422-4541
FAX:

Cynthia Hendershot
US Government
9800 Savage Rd. Suite 6678
Ft Meade, MD 20755-6678
USA
clhende@ncsc.mil
Phone: 301-688-2847
FAX: 301-688-2354

Bertrand Iooss
CEA Cadarache
DER/SESI/LCFR, Bat 212
Saint Paul lez Durance cedex, 13108
FRANCE
bertrand.iooss@cea.fr
Phone: (33) 4 42 25 7273
FAX: (33) 4 42 25 24 08

Rudolph Henninger
Los Alamos National Laboratory
P.O. Box 1663, MS D413
Los Alamos, NM 87545
USA
rjh@lanl.gov
Phone: 505-665-1562
FAX:

Bethanna Jackson
Imperial College London
Department of Civil and Environmental Engineering
London, SW7 2AZ
UNITED KINGDOM
b.m.jackson@imperial.ac.uk
Phone: +44-207-594-6115
FAX:

TotalParticipants 106

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

Hank Jackson
Air Force
AFFTC/OL-AB, MZ1100, AF Plant #4, Box 371
Ft. Worth , TX 76101-0371
USA
hank.jackson.af@dcma.mil
Phone: 817-763-4436
FAX: 817-77-4911

Donald Jarvis
Naval Research Laboratory
4555 Overlook Ave. S.W. Washington
Washington , D.C. 20375
USA
jarvis@ninja.nrl.navy.mil
Phone: 202-404-7682
FAX:

Julien Jacques
INRIA Rhone-Alpes
655 avenue de l'Europe
Montbonnot Saint Martin, 38330
FRANCE
julien.jacques@inrialpes.fr
Phone: 00 33 476 615 342 / 00 33 476 615 342
FAX: 00 33 476 615 477

Nina Jeliazkova
Institute of Parallel Processing, Bulgarian Academy of Sciences
25a "acad. G.Bonchev" stir
Sofia, 1113
BULGARIA
nina@acad.bg
Phone: +359 2 979 6616
FAX:

Michiel Jansen
Wageningen University and Research Centre
P.O. Box 100, 6700C Wsgeningen
Bergpad 1, 6703CL
NETHERLANDS
michiel.jansen@wur.nl
Phone: +31 317 417624
FAX:

Cliff Joslyn
Los Alamos National Laboratory
P.O. Box 1663, MS B265
Los Alamos, NM 87545
USA
joslyn@lanl.gov
Phone:
FAX:

Peter Janssen
RIVM/MNP
P.O Box 1
Bilthoven, 3720 BA
NETHERLANDS
peter.janssen@rivm.nl
Phone: +31 30 2742760
FAX: +31 30 2744427

Masato Koda
University Tsukuba
1-1-1 Tennoudai
Tsukuba,
JAPAN
koda@sk.tsukuba
Phone:
FAX: +81-29-853-5222

Kenneth Jarmen
PNNL
444 Palm Drive
Richland, VA 99352
USA
kj@pnl.gov
Phone:
FAX:

Snezano Konecni
Los Alamos National Laboratory
P.O. Box 1663, MS P946
Los Alamos , NM 87545
USA
Phone: 505-665-5546
FAX:

TotalParticipants 106

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

Bernard Krzykacz-Hausmann
GRS
GRS, Forschungsinstitute 85748 Garching
,
GERMANY
krb@grs.de
Phone: ++ 49 89 32004 394
FAX: ++49 89 32004 301

Luca Maciucescu
Los Alamos National Laboratory
P.O. Box 1663, MS P946
Los Alamos, NM 87545
USA
luca@lanl.gov
Phone:
FAX:

Tom Lauwagie
Katholieke Universiteit Leuven
Celestijnenlaan 300 B 3001 Heuerlee
, 3001
BELGIUM
tom.lauwagif@mech.kuleuven.ac.be
Phone: 00 32 16 32 86 06
FAX:

Sankaran Mahadevan
Vanderbilt University
Box 1831 - B
Nashville , TN 37235
USA
sankaran.mahadevan@vanderbilt.edu
Phone: 615-322-3040
FAX: 615-322-3365

Stella Ledbetter
Los Alamos National Laboratory
P.O. Box 1663, MS P366
Los Alamos, NM 87545
USA
sjled@lanl.gov
Phone: 505-667-4889
FAX: 505-667-7530

David Makowski
INRA
UMR Agromonie INRA/INA-PG
B.P. 01, 78850
FRANCE
makowski@grignon.inra.fr
Phone: 33 1 30 81 59 92
FAX: 33 1 30 81 54 25

Bruce Letellier
Los Alamos National Laboratory
P.O. Box 1663, MS K557
Los Alamos , NM 87545
USA
bcl@lanl.gov
Phone: 505-665-5188
FAX:

Michael McKay
Los Alamos National Laboratory
P.O. Box 1663, MS F600
Los Alamos, NM 87545
USA
mdm@lanl.gov
Phone: 505-667-6227
FAX:

Albert Liebetrau
PNNL
11822 SW Lodi Court
Powell Butte, OR 97753
USA
liebetraualsue@cs.com
Phone: 541-504-4751
FAX: 541-504-4751

Robert McLeod
Institute of Industrial Mathematical Sciences, University of Manitoba
420 Machray Hall
Winnipeg, Manitoba, R3T 2N2
CANADA
rob_mcleod@umanitoba.ca
Phone: 204-474-6724
FAX: 204-474-7602

TotalParticipants 106

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

Scott Mitchell
Sandia National Laboratories
P.O. Box 5800, MS 0847
Albuquerque, NM 87185
USA
samitch@sandia.gov
Phone: 505-845-7594
FAX: 505-844-9247

Hisashi Ninokata
Tokyo Institute of Technology
RLNR 2-12-1 O-okayama Megurotu
Tokyo, 152-8550
JAPAN
hninokat@nr.titech.ac.jp
Phone: +81-3-5734-3056
FAX: +81-3-5734-3056

Herve Monod
INRA
Unite MIA, Domaine de Vilvert, Jouy en Josas
, F78 352
FRANCE
herve.monod@jouy.inra.fr
Phone: +33 1 34 65 28 45
FAX: +33 1 34 65 22 17

Anthony O'Hagan
University of Sheffield
Department of Probability and Statistics, Hicks Bldg.
Sheffield, 53 7RH
UNITED KINGDOM
a.ohagan@sheffield.ac.uk
Phone: +44 114 222 3773
FAX: +44 114 222 3759

Douglas Mooney
Battelle
505 King Ave.
Columbus, OH 43201
USA
mooneyd@battelle.org
Phone: 614-424-7527
FAX: 614-458-7527

Alannah O'Sullivan
Heriot Watt University
Institute of Petroleum Engineering
edinburgh, EH14 4AS
SCOTLAND
alannah.osullivan@pet.hw.ac.uk
Phone: +44 0131 451 8298
FAX: +44 0131 451 8298

Lisa Moore
Los Alamos National Laboratory
P.O. Box 1663, MS 1663
Espanola, NM 87545
USA
lmoore@lanl.gov
Phone: 505-665-3958
FAX: 505-667-4470

Jeremy Oakley
University of Sheffield
Department of Probability and Statistics, The Hicks Building
hounsfield Road, Sheffield S3 7RH
UNITED KINGDOM
j.oakley@sheffield.ac.uk
Phone: +44 114 2223853
FAX: +44 114 2223759

Max Morris
Iowa State University
Department of Statistics, Snedecor Hall, ISU
Ames, IA 50011
USA
mmorris@iastate.edu
Phone: 515-294-2775
FAX: 515-294-4040

Bill Oberkampf
Sandia National Laboratories
MS 0828
Albuquerque, NM 87185
USA
wloberk@sandia.gov
Phone: 505-844-3799
FAX:

TotalParticipants 106

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

Keiji Ohmori
6-35-9 Kounandai Kounanku Yokohama
, 234-0054
JAPAN
bigforest@mub.biglobe.ne.jp
Phone: 81-3-6688-6251
FAX:

Ramesh Rebba
Vanderbilt University, Department of Civil and Environmental Engineering
2201 West End Avenue
Nashville, TN 37235
USA
rebba.ramesh@vanderbilt.edu
Phone: 615-343-4658
FAX:

Jennifer Park
SPAWAR S4S CEN San Diego
53560 Hull Street
San Diego, CA 92152
USA
jennifer@navy.mil
Phone: 619-553-2848
FAX:

Dave Robinson
Sandia National Laboratories
P.O. Box 5800, MS 0748
Albuquerque, NM 87185
USA
drobin@sandia.gov
Phone:
FAX:

Carlos Perez
Facultad de Veterinaria, Departmento de Matematicas
Avda, Universidad s/n. Caceres
, 10071
SPAIN
carper@unex.es
Phone: +34 927 257146
FAX: +34 927 257110

Tommy Rodgers
Lockheed Martin
P.O. Box 748
Fort Worth , TX 76101
USA
tommy.j.rodgers@lmco.com
Phone: 817-763-1472
FAX: 817-763-4790

Luca Podofillini
Politecnico Di Milano
Via Ponzio, 34/3 Milian
Milan, 20133
ITALY
luca.podofillini@polim.it
Phone: +39 02 23996340
FAX:

Renata Romanowicz
Lancaster University
Department of Envirnmental Sciences
Lancaster, LA1 4YQ
UNITED KINGDOM
r.romanowicz@lancaster.ac.uk
Phone: +44 1524 593894
FAX:

Guillaume Puel
LMT-Cachan
61 avenue du President Wilson
cachan Cedex, 94235
FRANCE
guillaume.puel@lmt.ens-cachan.fr
Phone: +33 1 47 40 28 75
FAX: +33 1 47 40 27 85

Vicente Romero
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185
USA
vjromer@sandia.gov
Phone: 505-844-5890
FAX:

TotalParticipants 106

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

Paolo Ruffo
ENI - E&P Division
GEBA Dept. Via Emilia 1 - S. Donato Milanese (MI)
, 20097
ITALY
paolo.ruffo@agip.it
Phone: 39.02.520.61777
FAX:

John Schuenemeyer
Southwest Statistical Consulting
960 sligo St.
Cortez, CO 81321
USA
jackswsc@charter.net
Phone: 970-564-9346
FAX:

Brian Rutherford
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185
USA
bmruth@sandia.gov
Phone: 505-844-3120
FAX:

Patrick Sheridan
University of Giessen
Senckenbrgstrasse 3
, 35390
GERMANY
patrick.sheridan@agrar.uni-giessen.de
Phone: +49-641-0037246
FAX:

Cedric Sallaberry
Sanida National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0776
USA
cnsalla@sandia.gov
Phone: 505-284-4813
FAX:

Hugo Sol
Vrije Universiteit Brussel (VUB)
Dept. MEMC - Faculty of Engineering
Pleinlaan 2, B1050
BELGIUM
hugos@vub.ac.be
Phone: ++32 2 629 29 61
FAX: +32 2 629 29 28

Andrea Saltelli
Joint Research Centre - European Commission
Via E. Fermi, 1 - TP361
Ispra (VA), 21020
ITALY
andrea.saltelli@jrc.it
Phone: 0039 0332 789686
FAX: 0039 0332 785733

Normand St-Pierre
Ohio State University
221-A Animal Science Building
Columbus, OH 43210
USA
st-pierre.8@osu.edu
Phone: 614-292-6507
FAX: 614-292-1515

Thomas Santner
Ohio State University
404 Cockins Hall
Columbus , OH 43210
USA
santner.1@osu.edu
Phone: 614-292-3593
FAX: 614-292-2096

Marius Stan
Los Alamos National Laboratory
P.O. Box 1663, MS G755
Los Alamos, NM 87545
USA
mastan@lanl.gov
Phone: 505-667-8726
FAX: 505-667-8021

TotalParticipants 106

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

Laura Swiler
Sandia National Laboratories
P.O. Box 5800, MS 0847
Albuquerque, NM 87185
USA
lpswile@sandia.gov
Phone: 505-844-8093
FAX:

Charles Tong
Lawrence Livermore National Laboratory
7000 East Avenue
Livermore, CA 94551
USA
chtong@llnl.gov
Phone: 925-422-3411
FAX:

Stefano Tarantola
Joint Research Centre - European Commission
Via E. Fermi, 1 - TP361
Ispra (VA), 21020
ITALY
stefano.tarantola@jrc.it
Phone: 0039 0332 789928
FAX:

Timothy Trucano
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185
USA
tgtruca@sandia.gov
Phone: 505-844-8812
FAX: 505-844-0918

Ben Thacker
Southwest Research Insitiute
6220 Culebra Road
San Antonio, TX 78238
USA
ben.thacker@swri.org
Phone: 210-522-3896
FAX: 210-522-6965

Velmir Vesselinov
Los Alamos National Laboratory
P.O. Box 1663, MS T003
Los Alamos, NM 87545
USA
vvv@lanl.gov
Phone: 505-665-1468
FAX:

Vincent Thomas
Los Alamos National Laboratory
P.O. Box 1663, MS T086
Los Alamos, NM 87545
USA
vat@lanl.gov
Phone: 505-665-3672
FAX:

Carlos Villa
DOW Chemical
2301 N. Brazosport Blvd. B-1217
Freeport, TX 77541
USA
cmvilla@dow.com
Phone: 979-238-5554
FAX:

Alison Tomlin
University of Leeds
Energy & Resources Research Institute
Leeds, 1S2 9JT
UNITED KINGDOM
fueast@leeds.ac.uk
Phone: +44 113 3432500
FAX: +44 113 2467310

Eric Wainwright
Decisioneering, Inc.
1515 Arapahoe St. Suite 1330
, 80202
USA
ewainwright@crystalball.com
Phone: 303-626-0103
FAX:

TotalParticipants 106

Sensitivity Analysis of Model Output
March 8-11, 2004
Participant List

Joanne Wendelberger
Los Alamos National Laboratory
P.O. Box 1663, MS F600
Los Alamos , NM 87545
USA
joanne@lanl.gov
Phone: 505-665-4840
FAX: 505=667-4470

Judit Zador
Eotvos University
Pazmany P. stny. 1/A
Budapest, H-1117
HUNGARY
zador@chem.elte.hu
Phone: +36-1-209-0555 x 1509
FAX: +36-1-209-0602

Brian Williams
Los Alamos National Laboratory
P.O. Box 1663, MS F600
Los Alamos, NM 87545
USA
bnanw@lanl.gov
Phone: 505-667-2331
FAX: 505-667-4470

Steven Wojtkiewicz
Sandia National Laboratories
P.O. Box 5800, MS 0847
Albuquerque, NM 87185
USA
sfwojtk@sandia.gov
Phone: 505-284-5482
FAX: 505-844-9297

Sheldon Wolk
Naval Research Laboratory
4555 Overlook Ave. S.W. Washington
Washington , D.C. 20375
USA
sheldon.wolk@nrl.navy.mil
Phone: 202-404-8784
FAX:

Kathie Womack
Los Alamos National Laboratory
P.O. Box 1663, MS T006
Los Alamos, NM 87545
USA
womack@lanl.gov
Phone: 505-663-5206
FAX: 505-663-5225

TotalParticipants 106