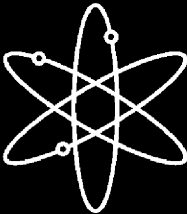# Combined Estimation of Hydrogeologic Conceptual Model and Parameter Uncertainty

Pacific Northwest National Laboratory

University of Arizona

**U.S. Nuclear Regulatory Commission
Office of Nuclear Regulatory Research
Washington, DC 20555-0001**

Combined Estimation of Hydrogeologic Conceptual Model and Parameter Uncertainty

Prepared by
P.D. Meyer, M. Ye, S.P. Neuman (UA),
K.J. Cantrell

Pacific Northwest National Laboratory
Richland, WA 99352

Subcontractor:
University of Arizona
Tucson, AZ 85721

T.J. Nicholson, NRC Project Manager

Abstract

The objective of the research described in this report is the development and application of a methodology for comprehensively assessing the hydrogeologic uncertainties involved in dose assessment, including uncertainties associated with conceptual models, parameters, and scenarios. This report describes and applies a statistical method, Maximum Likelihood Bayesian Model Averaging (MLBMA), to quantitatively estimate the combined uncertainty in model predictions arising from conceptual model and parameter uncertainties. The method relies on model averaging to combine the predictions of a set of alternative models. Implementation is driven by the available data. When there is minimal site-specific data the method can be carried out with prior parameter estimates based on generic data and subjective prior model probabilities. For sites with observations of system behavior (and optionally data characterizing model parameters), the method uses model calibration to update the prior parameter estimates and model probabilities based on the correspondence between model predictions and site observations. The set of model alternatives can contain both simplified and complex models, with the requirement that all models be based on the same set of data.

MLBMA was applied to the geostatistical modeling of air permeability at a fractured rock site. Seven alternative variogram models of log air permeability were considered to represent data from single-hole pneumatic injection tests in six boreholes at the site. Unbiased maximum likelihood estimates of variogram and drift parameters were obtained for each model. Standard information criteria provided an ambiguous ranking of the models, which would not justify selecting one of them and discarding all others as is commonly done in practice. Instead, some of the models were eliminated based on their negligibly small updated probabilities and the rest were used to project the measured log permeabilities by kriging onto a rock volume containing the six boreholes. These four projections, and associated kriging variances, were averaged using the posterior model probabilities as weights. Finally, cross-validation was conducted by eliminating from consideration all data from one borehole at a time, repeating the above process, and comparing the predictive capability of the model-averaged result with that of each individual model. Using two quantitative measures of comparison, the model-averaged result was superior to any individual geostatistical model of log permeability considered.

Contents

Figures

Tables

Executive Summary

In its performance assessments of decommissioning sites and other nuclear facilities, the U.S. Nuclear Regulatory Commission (NRC) staff uses a risk-informed, performance-based approach in which evaluation of risk is an integral part of, but not the sole basis for, decision making. The risk is, in part, manifested as uncertainty in estimates of dose. The importance of assessing uncertainty in dose is made clear by considering the following:

- long regulatory time frames (e.g., 1000 years),
- complex exposure pathways involving multiple media,
- relatively small incremental doses, and
- potentially limited site-specific characterization data.

The objective of the research described in this report is the development and application of a methodology for comprehensively assessing the hydrogeologic uncertainties involved in dose assessment modeling. For methodological purposes, prediction uncertainty is classified as being associated with one of three basic components of dose assessment models:

- the conceptual-mathematical basis of the model,
- model parameters, or
- the scenario to which the model is applied.

This report describes and applies a method to estimate the combined uncertainty in model predictions arising from conceptual model and parameter uncertainties. A future report will include the analysis of scenario uncertainty.

The primary steps involved in addressing uncertainty in model parameters are

- characterization of parameter uncertainty,
- propagation of parameter uncertainty into model output uncertainty, and
- parameter sensitivity analysis.

Parameter estimation, including the characterization of parameter uncertainty, is driven by the available data and information. In the most data-limited case, prior parameter estimates are based on available information that does not include site-specific parameter measurements. These estimates represent the largest degree of uncertainty. Meyer and Gee (1999) discuss data sources for characterizing hydrogeologic parameter uncertainty in the context of dose assessment modeling for license termination decisions. They

suggest the application of a hierarchy of data from national-scale databases (referred to as generic information) to site-specific measurements of parameter values. Site-specific parameter measurements, when available, can be used to update the prior estimates (Meyer et al., 1997), thereby decreasing parameter uncertainty. A similar methodology for the characterization of probability distributions for (adsorption) distribution coefficients is being developed as part of the research reported here (see Appendix A).

When observations of system state variables (e.g., hydraulic head, radionuclide concentration) are available at a site, formal calibration methods, using an inverse model, can be used to improve parameter estimates and characterize the uncertainty of these estimates. Calibrated parameter estimates represent the application of the maximum amount of data/information and yield parameters with the minimum uncertainty (Wang et al., 2003). Because they rely on an inverse model, calibrated parameter estimates are model-dependent. In fact, most calibration methods assume the model is correct. Errors thus represent the uncertainty in parameters given that the model is correct. This will underestimate parameter uncertainty.

Relying on a single conceptual representation of a system has two potential pitfalls: the rejection by omission of valid conceptual model alternatives, and reliance on an invalid representation by failing to adequately test it. The potential consequences are underestimation of uncertainty by under-sampling model space and biased results by relying on an invalid model. To obtain realistic risk estimates, effort should thus be made to evaluate multiple, alternative conceptualizations of the system being analyzed.

Any approach based on evaluation of a discrete set of alternative models will only be as good as the set of alternatives. That is, if the set of alternatives does not represent the full range of possibilities, conceptual model uncertainty will be underestimated. In Neuman and Wierenga's (2003) extensive discussion of conceptual model uncertainty they provide some advice on the generation of alternatives, summarized as follows.

- From the assembled database of site-specific data and other relevant information, consider alternative representations of space-time scales, number and type of hydrogeologic

units, flow and transport property characterization, system boundaries, initial conditions, fast flow paths, controlling transport phenomena, etc.

- Each conceptual model alternative should be supported by key data.
- Minimize inconsistencies, anomalies, and ambiguities.
- Apply the principle of Occam's window according to which one considers only a relatively small set of the most parsimonious models among those which, *a priori*, appear to be hydrologically most plausible in light of all knowledge and data relevant to the purpose of the model and, *a posteriori*, explain the data in an acceptable manner.
- Maximize the number of experts involved in the generation of alternative conceptualizations.
- Articulate uncertainties associated with each alternative conceptualization.

Having defined the set of alternatives, the options for addressing conceptual model uncertainty include the following.

- Evaluate each alternative and select the best model, either through an informal comparison or through evaluation of a formal model selection criterion.
- Evaluate each alternative and combine the results using some weighting scheme.

When multiple model conceptualizations are consistent with the available data, it may not be justifiable to rely on a single model structure. The method described here relies on model averaging to combine the predictions of alternative models. The weights applied to each model's predictions are estimated model probabilities.

The method uses a Maximum Likelihood implementation of Bayesian Model Averaging (MLBMA) described by Neuman (2003). If Δ is the predicted quantity (e.g., dose), its posterior distribution given a set of data $\boldsymbol{D}$ is

$$p\left(\Delta|\boldsymbol{D}\right) = \sum_{k=1}^{K} p\left(\Delta|M_k,\boldsymbol{D}\right) p\left(M_k|\boldsymbol{D}\right) \qquad \text{(E-1)}$$

where $\mathbf{M} = \left(M_1,...,M_K\right)$ is the set of all models considered. The posterior probability for model M_k is a function of the prior model probability and the model likelihood, as given by Bayes' rule,

$$p\left(M_k|\boldsymbol{D}\right) = \frac{p\left(\boldsymbol{D}|M_k\right) p\left(M_k\right)}{\sum_{l=1}^{K} p\left(\boldsymbol{D}|M_l\right) p\left(M_l\right)} \qquad \text{(E-2)}$$

The solution of these equations is accomplished by maximum likelihood estimation of each model's parameters.

Prior model probabilities in Equation E-2 [$p(M_k)$ and $p(M_l)$]are subjective values reflecting a belief about the <u>relative</u> plausibility of each model based on its apparent consistency with available knowledge and data. Posterior model probabilities are modifications of these subjective values based on an objective evaluation of each model's consistency with available data. Hence, the posterior probabilities are valid only in a comparative, not in an absolute, sense.

The maximum likelihood method can be applied to complex and simplified models as long as each model in the set of alternatives is based on the same data (Ye et al., 2003). It can be applied to deterministic models and also to stochastic models based on moment equations (Hernandez et al., 2003). Application of maximum likelihood also yields parameter sensitivity information.

Including prior information in the maximum likelihood calibration is an option, which allows one to condition the parameter estimates not only on site monitoring (observational) data but also on site characterization data, potentially rendering the model a better predictor.

Maximum likelihood allows the statistical parameters characterizing the parameter and state variable errors to be estimated. When these statistical parameters are known (i.e., not estimated), maximum likelihood reduces to generalized least squares estimation. In this case, available codes such as PEST and UCODE can be applied.

Maximum likelihood estimation yields an approximate covariance matrix for the parameter estimation errors. Assuming these errors to be Gaussian or log Gaussian, the probability distribution of model output [$p\left(\Delta|M_k,\boldsymbol{D}\right)$ in Equation E-1] can be determined by Monte Carlo simulation of Δ through random perturbation of the parameters. If the model is a geostatistical model or a stochastic moment model, it yields the expected value and variance of its output directly without Monte Carlo simulation.

In the most data-limited application, one in which there are no system observations with which to calibrate a model and the only available parameter information is

that available from generic databases, Equation E-1 reduces to

$$p(\Delta) = \sum_{k=1}^{K} p(\Delta|M_k) p(M_k).$$

That is, model predictions can still be made using prior parameter estimates and model averaging can still be carried out, but only with prior model probabilities. Since the predictions and model probabilities are not conditioned on state variable observations, however, the results are expected to be more uncertain and potentially more biased.

To implement MLBMA the following steps are followed.

(1) Postulate alternative conceptual-mathematical models for a site using guidance provided in Neuman and Wierenga (2003).

(2) Assign a prior probability to each model.

(3) Optionally assign prior probabilities to the parameters of each model, using, for example, guidance provided in Meyer and Gee (1999).

(4) Obtain posterior maximum likelihood parameter estimates, and estimation covariance, for each model by inversion (model calibration). In many cases, available codes such as PEST and UCODE can be applied to this step.

(5) Calculate a posterior probability for each model using the model calibration results and the prior model probabilities.

(6) Predict quantities of interest using each model.

(7) Assess prediction uncertainty (distribution, variance) for each model using Monte Carlo or stochastic moment methods.

(8) Weight predictions and uncertainties by the corresponding posterior model probabilities.

(9) Sum the results over all models.

To evaluate MLBMA, it was applied to seven alternative variogram models of log air permeability data from single-hole pneumatic injection tests in six boreholes at the Apache Leap Research Site (ALRS) in central Arizona. Unbiased ML estimates of variogram and drift parameters were obtained using adjoint state maximum likelihood cross validation in conjunction with universal kriging and generalized least squares. Standard information criteria provided an ambiguous ranking of the models, which did not justify selecting one of them and discarding all others as is commonly done in practice. Instead, three of the models were eliminated based on their negligibly small posterior probabilities and the remaining four models were used to project the measured log permeabilities by kriging onto a rock volume containing the six boreholes. These four projections, and associated kriging variances, were averaged using the posterior probability of each model as weight.

Finally, the results were cross-validated by eliminating from consideration all data from one borehole at a time, repeating the above process, and comparing the predictive capability of MLBMA with that of each individual model. The predictive capabilities of the alternative models and the MLBMA result were compared through their log scores. The lower the predictive log score of a model, the smaller the amount of information lost upon eliminating a borehole's data from the original dataset (i.e., the higher the probability that the model based on the reduced dataset would reproduce the eliminated borehole's data).

Another measure of model performance is its predictive coverage. This is the percent of measurements from the eliminated borehole's data that fall within a given prediction interval generated by conducting Monte Carlo simulations of log air permeability conditioned on the data from the remaining boreholes.

The table below lists the average log score for the three model alternatives with the highest posterior probability, as well as the average of corresponding MLBMA scores. The average predictive log score of MLBMA is seen to be smaller than that of any individual model, indicating that MLBMA is a better predictor than any of the single model alternatives. The table also shows the predictive coverage of MLBMA, which is larger than that of any individual model, attesting once again to its superior performance.

Table E-1. Comparison of MLBMA with individual model alternatives

Model	Pow0	Exp0	Exp1	MLBMA
Predictive log score	34.1	35.2	34.0	31.4
Predictive coverage (%)	86.5	80.8	83.7	87.5

Foreword

This technical contractor report was prepared by Pacific Northwest National Laboratory[1] (PNNL) under their DOE Interagency Work Order (JCN Y6465) with the U.S. Nuclear Regulatory Commission. This research report describes an approach for integrating two methodologies developed to assess uncertainties: one for evaluating hydrologic conceptual model uncertainty as documented in NUREG/CR-6805, and the second for estimating hydrologic parameter uncertainty as documented in NUREG/CR-6767. This report provides both the logic developed and examples demonstrating the approach using field data. The detailed input and analyses for the real-world examples are presented in the report's appendix and may be useful in decommissioning reviews of complex sites. This report is consistent with the NRC strategic performance goal of making NRC activities and decisions more effective, efficient, and realistic by identifying and estimating uncertainties.

The report demonstrates, using examples relevant to decommissioning analyses, that sources of uncertainty can be identified, quantified, and integrated using a comparative model analysis approach. The report illustrates the effectiveness of the integrated methodology to estimate uncertainty in model predictions arising from both conceptual and parameter uncertainties. This information will assist NRC licensing staff, Agreement State regulators, and licensees in their decision making by identifying and quantifying overall uncertainties in performance assessment models.

This report, as with the previous reports on individual sources of uncertainty, is not a substitute for NRC regulations, and compliance is not required. The approaches and/or methods described in this NUREG/CR are provided for information only. Publication of this report does not necessarily constitute NRC approval or agreement with the information contained herein. Use of product or trade names is for identification purposes only and does not constitute endorsement by the NRC or Pacific Northwest National Laboratory.

Cheryl A. Trottier, Chief
Radiation Protection, Environmental Risk and Waste Management Branch
Division of System Analysis and Regulatory Effectiveness
Office of Nuclear Regulatory Research

---

[1] Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract DE-AC06-76RLO 1830.

Acknowledgments

1   Introduction

In its performance assessments of decommissioning sites and other nuclear facilities, the U.S. Nuclear Regulatory Commission (NRC) staff uses a risk-informed, performance-based approach in which evaluation of risk is an integral part of, but not the sole basis for, decision making. NRC regulatory criteria are often written in terms of dose. For example, the primary regulatory criterion for license termination is a maximum dose for the period up to 1000 years from the time of decommissioning (see Table 1-1). One might argue that risk (such as the risk of premature death) could be derived from knowledge of exposure to a particular dose. When, however, estimating that dose involves predictions of contaminant transport and exposure via complex contaminant exposure pathways over a 1000-year period, then there is an obvious additional component of uncertainty contributing to risk. That component is the uncertainty in the estimate of dose. The importance of assessing uncertainty in dose is made clear by considering

- the long regulatory time frame,
- complex exposure pathways involving multiple media,
- the relatively small incremental dose specified in the regulations, and
- potentially limited site-specific characterization data.

In the license termination case, the result of a quantitative assessment of this uncertainty will be an estimate of the probability distribution of dose to the average member of the critical group for the 1000-year period following decommissioning.

There are numerous sources of uncertainty that are potentially significant contributors to an estimate of the probability of dose. This is a consequence of the multiple potential exposure pathways. The analysis presented in this report only addresses the pathways involving transport of radionuclides in water. For license termination, that includes a residential farmer scenario in which exposure comes from the use of contaminated groundwater for home, garden, and farm. Thus, the uncertainties considered are those related to transport from a source (typically near the ground surface) through unsaturated soils and groundwater to an exposure point via a pumped well or surface water body. The methods described here are general, however, and could be applied to other exposure scenarios.

Although the analysis described here is limited to hydrogeologic uncertainty, it is comprehensive in the sense that all types of hydrogeologic uncertainty are considered. Uncertainty is defined, for the purposes of this study, as a lack of certainty due to

- incomplete knowledge of the system being analyzed;
- measurement or sampling error in characterizing the system's features, events, and processes;
- variability in the system's properties;
- disparity among the sampling, simulation, and actual scales of the system's features, events, and processes; and
- randomness in the system's stresses, particularly transient external stresses, often in a short-time context.

Table 1-1. Summary of radiological criteria for license termination (10 CFR Part 20 Subpart E) (from Meyer and Gee, 1999). (TEDE – Total Effective Dose Equivalent; ALARA – As Low As Reasonably Achievable)

	Unrestricted Release	Restricted Release	
Dose Criterion	25 mrem TEDE per year peak annual dose to the average member of the critical group	25 mrem TEDE per year peak annual dose to the average member of the critical group while controls are in place	100 or 500 mrem TEDE per year peak annual dose to the average member of the critical group upon failure of the controls
Time Frame	1000 years	1000 years	
Other Requirements	ALARA	ALARA, financial assurance, public participation	

Note that this definition includes uncertainty that can be reduced with sufficient data (sometimes referred to as subjective or epistemic uncertainty; see Helton 1996) and uncertainty that is an irreducible characteristic of the system (sometimes referred to as stochastic or aleatory uncertainty). An example of the former is uncertainty about the continuity (thickness) of a low permeability hydrostratigraphic unit. Examples of the latter are the annual recharge rate over the next 1000 years, or the stage of a river hydraulically connected to a groundwater system. It is often argued that these two broad types of uncertainty should be kept separate in the application of uncertainty analysis methods (Helton, 1994; Ayyub and McCuen, 2003); this may improve the ability to draw correct conclusions about the important factors leading to system success/failure and the value of additional data. Winkler (1996) suggests that uncertainties that appear irreducible may, in fact, often be a function of the available knowledge (i.e., subjective). For example, river stage may be inherently variable, but that variability could, in principle, be entirely accounted for if a sufficiently detailed hydrologic model and the associated data were available. Winkler (1996) argues that distinctions between types of uncertainty are largely related to sources of information and that it is more useful to think in terms of what is needed to accomplish the modeling task: adequate decomposition of the problem, combining various sources of information, assessing the value of additional data, and effectively utilizing sensitivity analysis. This is the viewpoint adopted in this report.

Models are generally used to make consistent, quantitative assessments of future dose required by criteria such as that given in Table 1-1. Although we do not strictly distinguish between subjective and stochastic types of uncertainty, from a methodological perspective we classify uncertainty as being associated with one of three basic components of dose assessment models:

- the conceptual-mathematical basis of the model,
- model parameters, or
- the scenario to which the model is applied.

The model conceptual basis can be thought of as a hypothesis about the behavior of the system being modeled and the relationships between the components of the system. This conceptualization is typically represented mathematically to render quantitative predictions; thus it is appropriate to talk about a conceptual-mathematical model (sometimes referred to as model structure). The model parameters are the quantities required to obtain a solution from the model

(and thus are model-specific). A scenario is defined here as a future state or condition assumed for a system that is the result of an event, process, or feature that was not assumed in the initial base case definition of the system and diverges significantly from the initial base case. A scenario may be imposed by humans (e.g., irrigation schemes and ground-water extraction) but may also be natural (e.g., glaciation and flooding). Scenarios are often considered in a long-time context. Only hydrologically related aspects of scenario uncertainty are included in this analysis.

The objective of the research described in this report is the development and application of a methodology for comprehensively assessing the uncertainties involved in dose assessment, including uncertainties associated with conceptual models, parameters, and scenarios. In addressing this problem we have generally adopted a Bayesian viewpoint. The merits of a Bayesian (subjectivist) approach to probability relative to a classical (frequentist) approach have been discussed in many publications (e.g., Martz and Waller, 1988; Abramson, 1988). Our approach is Bayesian primarily for practical reasons. Quantification of hydrogeologic uncertainty for dose assessments must often deal with very limited observations of site characteristics. Generic and indirect data can be and generally are used to infer site properties. For example, geologic characteristics may be inferred from analysis of outcrops, hydraulic characteristics may be estimated from soil-textural information, and radionuclide adsorption characteristics may be assigned from a database of values measured at other sites under a variety of conditions. In addition, the assessment of conceptual model and scenario probabilities seems inherently subjective. The Bayesian approach provides a means to incorporate different types of data and subjective judgments into the assessment of uncertainty.

This report describes and applies a method to estimate the combined uncertainty in model predictions arising from conceptual model and parameter uncertainties. The inclusion of scenario uncertainty will be described in a future report. Chapter 2 provides some background on the quantification of parameter and conceptual model uncertainty. A related discussion of an approach being developed as part of this project for evaluating uncertainty in the distribution coefficient parameter is included in Appendix A. Chapter 3 describes the maximum likelihood Bayesian model averaging method, a general method for combining quantitative estimates of conceptual model and parameter uncertainty. Chapter 4 is an application of this method to the geostatistical modeling of air permeability at a fractured rock site. This application was chosen

2

because the site is a relatively well-controlled experimental research site with good characterization data. In addition the results of past studies at the site were available to us. Applications that are more reflective of actual NRC-regulated sites will be the focus of future efforts.

The developments described here are being coordinated with other Federal agencies cooperating under the Interagency Steering Committee on Multimedia Environmental Models Memorandum of Understanding (ISCMEM MOU) (see http://ISCMEM.org). Results reported here have been discussed with members of the Working Group on Uncertainty and Parameter Estimation organized under the steering committee and have been presented at the International Workshop on Uncertainty, Sensitivity, and Parameter Estimation for Multimedia Environmental Modeling, held August 19-21, 2003, at NRC Headquarters and organized by the Working Group.

2 Quantification of Parameter and Conceptual Model Uncertainty

2.1 Parameter Uncertainty

2.1.1 Sources of Parameter Uncertainty

The sources of uncertainty outlined in the previous chapter that contribute to hydrogeologic parameter uncertainty can be clearly illustrated with the aid of Figure 2-1, a photo of a trench face from an excavation in the 200 Area of the Hanford Site. A large variation in soil particle size can be seen, ranging from fine silts to very coarse gravels. The profile shows a layered structure with evidence of cross-bedding; the scale of the structures is on the order of a few centimeters. This variation results in hydraulic and transport properties that may vary over several orders of magnitude on this

same small scale. Measurements are likely to be made on a somewhat larger scale, perhaps 10 cm or more. Exhaustive sampling to determine the exact nature of the subsurface at this scale will be impossible, thus requiring interpolation between measurements and other indirect methods to estimate properties at unmeasured locations. In addition, the simulation scale for most practical applications (and thus the scale of the parameters) is likely to be significantly larger than the measurement scale, from a few tens of centimeters to many meters.

The impact of measurement errors on parameter uncertainty is often felt to be small relative to other sources of uncertainty and easily quantified. Holt et al. (2002) provide some evidence that relatively simple

Figure 2-1. Photograph of a trench face from an excavation in the 200 Area of the Hanford Site, Washington (photograph by John Selker, Oregon State University).

measurement errors can introduce significant parameter uncertainties. They simulated tension infiltrometer measurements with added pressure transducer error (observation error) and contact error (inversion error). They used the simulated measurements to estimate the variance and correlation length of the parameters of the Gardner hydraulic conductivity model over a range of true values representing poorly-sorted to well-sorted silt to coarse sand. The ratio of estimated to true parameter values (for the variances and correlation lengths) ranged from less than 0.5 to more than 2.5. These are significant errors for parameters representing the average characteristics of a site. Holt et al. (2002) also observed that the modeled errors produced spurious parameter correlations, an effect that has likely been poorly appreciated in most applications.

An additional source of parameter uncertainty that has likely not been fully appreciated can be illustrated using results presented in Zimmerman et al. (1998). They compared results from seven models calibrated on the same set of data by different participant groups using different inverse methods. The ratio of estimated to true parameter values for the variance and correlation length of the transmissivity are shown in Figure 2-2 for each of the inverse methods used. The true transmissivity field was synthetically generated.

An exponential model was fit to the average empirical variogram for a set of realizations obtained from each inverse method. The results shown are for Test Problem 1, the simplest transmissivity model used (an isotropic, exponential variogram). Nonetheless, the parameter errors resulting simply from the use of different inverse methods (and participants) were significant.

2.1.2 Analysis of Parameter Uncertainty

The analysis of parameter uncertainty has received much attention in the literature. Helton (1993) and McKay (1995) provide discussions of parameter uncertainty that are particularly relevant to dose assessment modeling. The primary steps involved in addressing uncertainty in model parameters are

- characterization of parameter uncertainty,
- propagation of parameter uncertainty into model output uncertainty, and
- parameter sensitivity analysis.

Parameter estimation, including the characterization of parameter uncertainty, is driven by the available data and information. Figure 2-3 is a simple representation of the parameter estimation process, where it is
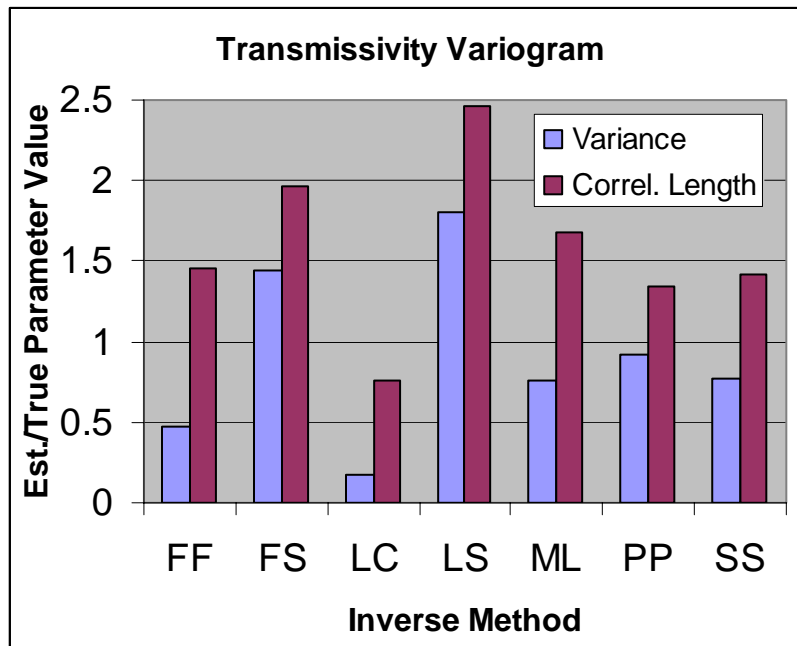
Figure 2-2. Ratio of estimated to true parameter values for variance and correlation length of transmissivity for seven different inverse methods. Results from Test Problem 1 of Zimmerman et al. (1998). (FF=Fast Fourier Transform, FS=Fractal Simulation, LC=Linearized Cokriging, LS=Linearized Semianalytical, ML=Maximum Likelihood, PP=Pilot Point, SS=Sequential Self-Calibration)
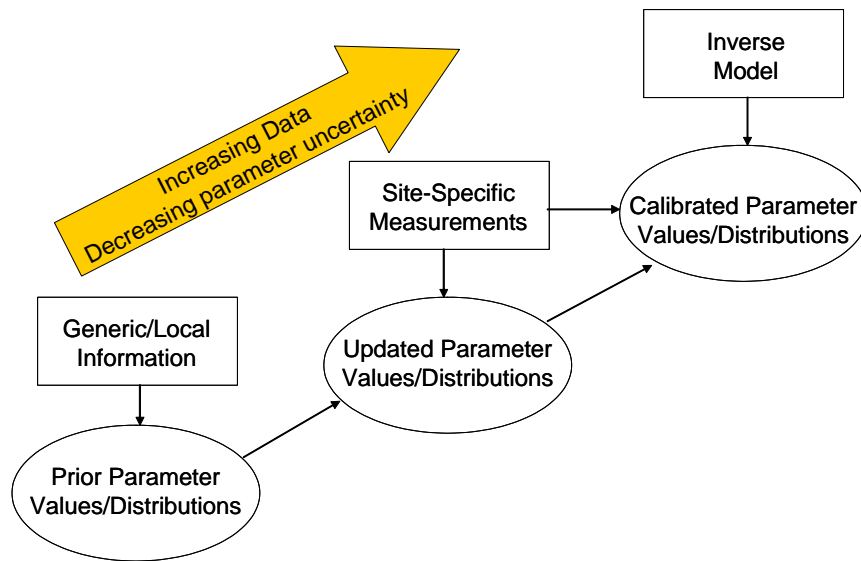
Figure 2-3. Use of data/information in parameter estimation

assumed that the process provides not only parameter estimates, but also some measure of the parameter uncertainty. This could take the form of bounding values, variances, or specific distributional forms. At the lower left are prior parameter estimates based on available information that does not include site-specific parameter measurements. This information may include a compilation of parameter values from numerous sites, or data from analogous sites. The prior parameter estimates represent the largest degree of uncertainty and the least amount of site-specific data. In the center of Figure 2-3 are updated (posterior) parameter estimates that are based on the prior estimates but include the effect of site-specific parameter measurements. They represent a decrease in parameter uncertainty from the prior estimates.

Meyer and Gee (1999) discuss data sources for characterizing hydrogeologic parameter uncertainty in the context of dose assessment modeling for license termination decisions. They suggest the application of a hierarchy of data from national-scale databases (referred to as generic information) to site-specific measurements of parameter values. Their methodology is represented schematically in Figure 2-4. Information from the national-scale databases is used by Meyer and Gee (1999) to specify prior parameter distributions that can be updated subsequently in a Bayesian approach using site-specific parameter data (Meyer et al., 1997), which is expected to be sparse or non-existent at many of the decommissioning sites. In data-limited applications parameter probability distributions can also be based on the subjective opinions of one or more

experts. Formal procedures are available to provide consistency in the elicitation of expert opinions regarding probabilities (Morgan and Henrion, 1990). A methodology relying on generic databases, however, has the advantage of being less expensive and more easily applied to a wide variety of sites. The methodology is currently being extended to include the characterization of probability distributions for (adsorption) distribution coefficients of selected radionuclides (see Appendix A).

When observations of state variables (e.g., hydraulic head, radionuclide concentration) are available at a site, formal calibration methods can be used to improve parameter estimates and characterize the uncertainty of these estimates (Hill, 1998). As shown in the upper right of Figure 2-3, this involves the application of an inverse model. These calibrated parameter values may include the effect of the site-specific parameter measurements. In this case the updated parameter estimates shown in Figure 2-3 are referred to as the prior parameter estimates for the calibration. Calibrated parameter estimates represent the application of the maximum amount of data/information and yield parameters with the minimum uncertainty. An application to unsaturated flow presented in Wang et al. (2003) illustrates the relationships between the data used in parameter estimation and the resulting prediction uncertainty.

Note that prior and updated parameter estimates may be independent of a model. As discussed in Meyer and Gee (1999), however, there must be a correspondence

7

Types of Information   ## Application of Information

Generic Information from Regional/National Sources e.g., UNSODA, NUREG/CR-6565	Parameter Uncertainty Distributions, Bounding Values, Best Estimates in the absence of site-specific data
Local Information from Regional/National Sources, e.g., Natl.Soil Char. Dbase, STATSGO, SSURGO, NCDC, NWIS, GWIS	Modify Uncertainty Distributions and Bounding Values, Best Estimates in the absence of site-specific data
Local Information from Local Sources e.g., Extension Service, State Agencies, University/Industry Experts	Modify Uncertainty Distributions and Bounding Values, Best Estimates in the absence of site-specific data
Site-specific direct measurements	Best Estimates, Modify Uncertainty Distributions

Increasingly Site-Specific (left vertical arrow)
Reduced Uncertainty (right vertical arrow)

Figure 2-4. Types and uses of data sources and information for characterizing hydrogeologic parameter uncertainty in dose assessments for license termination decisions (from Meyer and Gee, 1999). Acronyms refer to various databases.

between the estimates and the parameters assigned those estimates, e.g., a model that has a single value of a parameter representing a site must be assigned a value that represents a mean. Similarly, the uncertainty in that parameter value must represent uncertainty in the mean. Because they rely on an inverse model, calibrated parameter estimates are model-dependent. In fact, most calibration methods assume the model is correct. Errors thus represent the uncertainty in parameters given that the model is correct. This will underestimate parameter uncertainty.

Zimmerman et al. (1998) evaluated a variety of calibration methods using a set of hypothetical (generated) data based on the Waste Isolation Pilot Plant site. Transmissivity fields for two-dimensional groundwater flow models were calibrated on four test problems. One of their conclusions was that the calibrated models consistently underestimated the "true" variability in transport. The maximum likelihood (Carrera and Neuman, 1986a) and sequential self-calibration (Gomez-Hernandez et al., 1997) methods were consistently ranked higher than the other methods. The sequential self-calibration method offers the advantage of producing spatially variable transmissivity fields that honor the spatial statistics of the transmissivity field. A calibrated, stochastic groundwater simulation can be carried out using a set of these fields in a Monte Carlo simulation. The maximum likelihood method is more general, however, and can be applied to the calibration of a wide variety

of parameters, including statistical parameters. Maximum likelihood is used in the method and application described in Chapters 3 and 4.

Computer codes that can be adapted to the calibration of any simulation model have recently become available (Poeter and Hill, 1998; Doherty, 2002). One of these codes, PEST (Doherty, 2002) was used in the application presented in Chapter 4. A method for calibrating geostatistically-simulated parameter fields (similar to the sequential self-calibration method) has recently been demonstrated using PEST (Doherty, 2003).

A variety of methods for propagating parameter uncertainty are available, including Monte Carlo simulation, the first-order, second-moment method (Kunstmann et al., 2002), the stochastic response surface method (Isukapalli et al. 1998), and stochastic moment methods (Dagan and Neuman, 1997; Zhang, 2001). Monte Carlo simulation is the most generally applicable method and was used in the application presented in Chapter 4. The stochastic moment methods are appealing because of their potential computational advantage over Monte Carlo simulation. Recent progress in handling conditions that introduce nonstationarities (Zhang, 2001) have made these methods more generally applicable.

Uncertainties must be defined on a site-specific basis and the importance of individual sources may vary site

by site or even with different objectives at the same site. Determination of the parameters that are most important to the prediction uncertainty is the final element of an assessment of parameter uncertainty. This is generally carried out through the implementation of sensitivity analysis (Saltelli et al., 2000a; Helton, 1993). Meyer and Taira (2001) applied differential, graphical, and sampling-based methods of sensitivity analysis to decommissioning problems. Sensitivity measures may also be obtained during the calibration procedure (Hill, 1998; Tiedeman et al., 2003). Global sensitivity methods (Borgonovo et al, 2003; Saltelli et al., 2000b; McKay, 1995) partition the total prediction variance according to the contribution of each parameter and also determine the contribution to prediction variance due to interactions between parameters. A sensitivity analysis was not conducted for the application described in Chapter 4.

2.2 Conceptual Model Uncertainty

The sources of uncertainty described in the previous sections result in multiple valid representations of parameter values. That is, for a given model structure, there will be multiple sets of parameter values that provide valid representations of observed system behavior. In a similar manner, the same sources of uncertainty may result in valid alternative model structures or conceptualizations. When multiple model conceptualizations are consistent with the available data, it may not be justifiable to rely on a single model structure. Relying on a single conceptual representation of a system has two potential pitfalls: the rejection by omission of valid alternatives, and reliance on an invalid representation by failing to adequately test it. The potential consequences are underestimation of uncertainty by under-sampling model space and biased results by relying on an invalid model.

When discussing model uncertainty, it is instructive to view model structure as the combination of a conceptual model and a mathematical model: a conceptual-mathematical model (Neuman and Wierenga, 2003). The conceptual model can be thought of as a hypothesis about the system behavior and the relationship between system components. It is primarily qualitative and comprehensive. The mathematical model can be thought of as a process to test the conceptual model hypothesis. It is a quantitative, possibly simplified implementation of the conceptual model.

Figure 2-5 illustrates the relationship between alternative conceptual-mathematical models. Each conceptual model is based on the available site data and other relevant information and represents a distinct conceptualization of system characterization or behavior. For example, alternative conceptual models might be represented by the presence and absence of leakage from an underlying aquifer; or the presence and absence of matrix-fracture interaction in a fractured rock. In addition, a single conceptual model may be implemented in more than one way: for example, a fractured rock may be represented as an equivalent porous medium or as a discrete network of fractures. The process of conceptual-mathematical model development may be iterative as additional site data becomes available and conceptual models are updated.

In this report, "conceptual model uncertainty" should be interpreted as "conceptual-mathematical model uncertainty," representing uncertainty in either the conceptual model or its mathematical implementation.

2.2.1 Analysis of Conceptual Model Uncertainty

Methods for the quantification of conceptual model uncertainty are much less well established than those addressing parameter uncertainty. Mosleh et al. (1994) provide a good introduction to the issues involved.
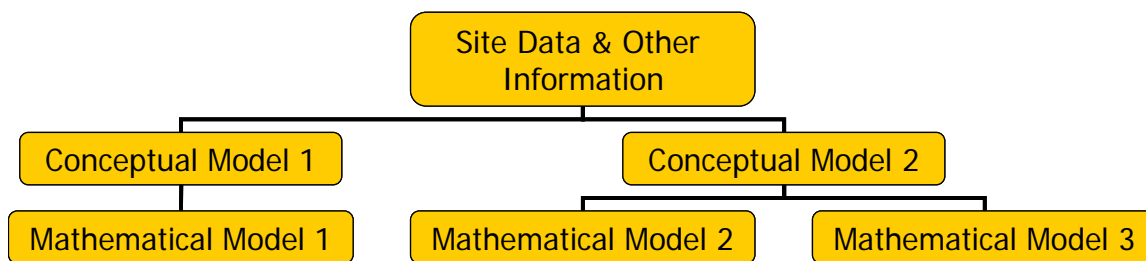

Figure 2-5. A schematic representation of the relationship between alternative conceptual-mathematical models

Neuman and Wierenga (2003) discuss a wide variety of issues related to hydrogeologic conceptual model uncertainty, including many instances of its practical importance.

While it is generally possible to specify a reasonable probability distribution representing the complete set of possibilities for the value of a parameter, it is not generally possible to specify the complete set of possible conceptual model alternatives. As a result, conceptual model uncertainty has generally been represented as a discrete distribution, with a small number of model alternatives taken as the complete set of possibilities. In the generic example of Figure 2-5, the complete set of possibilities consists of three conceptual-mathematical model alternatives. Having defined the set of alternatives, the options for addressing conceptual model uncertainty include the following.

- Evaluate each alternative and select the best model. This may be carried out through an informal comparison (James and Oldenburg, 1997; Cole et al., 2001) or through evaluation of a formal model selection criterion (Burnham and Anderson, 2002). As discussed previously, selection of a single model may not always be justifiable.
- Evaluate each alternative and combine the results using some weighting scheme, such as the likelihood-based weighting of Beven and Freer (2001), the multimodel ensemble approach of Krishnamurti et al. (2000), the model likelihood weighting of Burnham and Anderson (2002), and the model probability weighting of Draper (1995).

Neuman (2003) reviews a number of approaches that have been used to address conceptual model uncertainty. The method he proposes, a version of the model averaging method described in Draper (1995), was used here and is discussed in detail in the following chapter.

Any approach based on evaluation of a discrete set of alternative models will only be as good as the set of alternatives. That is, if the set of alternatives does not represent the full range of possibilities, conceptual model uncertainty will be underestimated. In Neuman and Wierenga's (2003) extensive discussion of conceptual model uncertainty, they provide some advice on the generation of alternatives, summarized as follows.

- From the assembled database of site-specific data and other relevant information, consider alternative representations of space-time scales, number and type of hydrogeologic units, flow and transport property characterization, system boundaries, initial conditions, fast flow paths, controlling transport phenomena, etc.
- Each conceptual model alternative should be supported by key data.
- Minimize inconsistencies, anomalies, and ambiguities.
- Apply the principle of Occam's window (Jefferys and Berger, 1992; Madigan and Raftery, 1994) according to which one considers only a relatively small set of the most parsimonious models among those which, *a priori*, appear to be hydrologically most plausible in light of all knowledge and data relevant to the purpose of the model and, *a posteriori*, explain the data in an acceptable manner.
- Maximize the number of experts involved in the generation of alternative conceptualizations.
- Articulate uncertainties associated with each alternative conceptualization.

Because the set of alternative conceptual models is unlikely to represent the full range of possibilities, evaluations of model uncertainty should be viewed as relative comparisons. That is, they may be used to conclude that one model is better than another for the intended purpose, but they cannot necessarily be used to conclude that any model is a good model. In addition, as stated above, the contribution of model uncertainty to overall prediction uncertainty will be underestimated.

Gaganis and Smith (2001) presented a unique analysis based on Bayes Factors for calculating an absolute measure of conceptual model uncertainty for a single model (that is, without comparison to alternative models). We evaluated this method using two synthetic examples of groundwater flow in which model structural errors were introduced through a boundary flux and a source term error. Parameter uncertainty was represented by a random field of transmissivity. Although head and parameter measurements were error-free and all driving forces other than the specified model errors were known, the method of Gaganis and Smith (2001) provided inconsistent estimates of the (known) model uncertainty. Based on these results, we feel the method is, at best, not generally applicable.

3 Combining Parameter and Conceptual Model Uncertainty

This chapter discusses a method to provide an optimal way of combining the predictions of several alternative models and assessing their joint predictive uncertainty, with consideration of parameter and conceptual model uncertainty. This method relies on the specification of a set of alternative models (with the consequent limitations discussed in the previous chapter and below) and weights the alternative model results by a measure of the model probabilities. The method was originally proposed by Neuman (2002).

3.1 Bayesian Model Averaging

A formal method of evaluating prediction uncertainty with full consideration of model uncertainty is Bayesian Model Averaging (BMA) (Draper, 1995; Hoeting et al., 1999). Using the notation of Hoeting et al. (1999), if Δ is the predicted quantity, its posterior distribution given a set of data $\boldsymbol{D}$ is

$$p\left(\Delta|\boldsymbol{D}\right) = \sum_{k=1}^{K} p\left(\Delta|M_k, \boldsymbol{D}\right) p\left(M_k|\boldsymbol{D}\right) \qquad (1)$$

where $\mathbf{M} = \left(M_1, ..., M_K\right)$ is the set of all models considered, at least one of which must be correct. As discussed in the previous chapter, Neuman and Wierenga (2003) provide guidance on selecting a set of models that is small enough to be computationally feasible yet large enough to represent the breadth of significant possibilities.

In (1), $p\left(\Delta|\boldsymbol{D}\right)$ is the average of the posterior distributions $p\left(\Delta|M_k, \boldsymbol{D}\right)$ under each model, weighted by their posterior model probabilities $p\left(M_k|\boldsymbol{D}\right)$. The posterior probability for model M_k is given by Bayes' rule,

$$p\left(M_k|\boldsymbol{D}\right) = \frac{p\left(\boldsymbol{D}|M_k\right) p\left(M_k\right)}{\sum_{l=1}^{K} p\left(\boldsymbol{D}|M_l\right) p\left(M_l\right)} \qquad (2)$$

where

$$p\left(\boldsymbol{D}|M_k\right) = \int p\left(\boldsymbol{D}|\boldsymbol{\theta}_k, M_k\right) p\left(\boldsymbol{\theta}_k|M_k\right) d\boldsymbol{\theta}_k \qquad (3)$$

is the integrated likelihood of model M_k,

$\boldsymbol{\theta}_k$ is the vector of parameters associated with model M_k, $p\left(\boldsymbol{\theta}_k|M_k\right)$ is the prior density of $\boldsymbol{\theta}_k$ under model M_k, $p\left(\boldsymbol{D}|\boldsymbol{\theta}_k, M_k\right)$ is the joint likelihood of model M_k and its parameters $\boldsymbol{\theta}_k$, and $p\left(M_k\right)$ is the prior probability that M_k is the correct model. All probabilities are implicitly conditional on $\mathbf{M}$.

The posterior mean and variance of Δ are (Draper, 1995)

$$E\left[\Delta|\mathbf{D}\right] = \sum_{k=1}^{K} E\left[\Delta|\mathbf{D}, M_k\right] p\left(M_k|\mathbf{D}\right) \qquad (4)$$

$$Var\left[\Delta|\mathbf{D}\right] = \sum_{k=1}^{K} Var\left[\Delta|\mathbf{D}, M_k\right] p\left(M_k|\mathbf{D}\right) +$$

$$\sum_{k=1}^{K} \left(E\left[\Delta|\mathbf{D}, M_k\right] - E\left[\Delta|\mathbf{D}\right]\right)^2 p\left(M_k|\mathbf{D}\right) \qquad (5)$$

In (5), the first term on the right-hand side represents within-model variance; the second term represents between-model variance. Note that the predictive probabilities (1) and leading moments (4) – (5) are weighted by the posterior probabilities of the individual models.

3.1.1 Interpretation of Model Probability

Philosophical difficulties with the BMA approach have been discussed by Winkler (1993) and center on the interpretation of $p\left(M_k\right)$ as the probability that M_k is the correct model and the method's requirement that one of the M_k is in fact the correct model. Winkler (1993) argues that, although this interpretation is intuitively appealing, the existence of a "correct" model is questionable since all models are approximations of reality.

One approach to these philosophical difficulties is to interpret model probability in relative terms (e.g., Zio and Apostolakis, 1996), where the model with the greatest probability is the "best" model (and all model probabilities sum to one). Winkler (1993) suggests that this means $p\left(\Delta|M_k, \boldsymbol{D}\right)$ must be interpreted as being conditional to the "best" model, and asks whether there is utility in that interpretation if the "best" model is not very good. As discussed in the previous chapter, basing the analysis on a set of model alternatives that do not encompass all possibilities implies a relative

comparison between models. We thus interpret prior model probabilities to be subjective values reflecting the analyst's belief about the <u>relative</u> plausibility of each model based on its apparent (qualitative, *a priori*) consistency with available knowledge and data.

Whereas prior model probabilities must in our view remain subjective, the posterior model probabilities are modifications of these subjective values based on an objective evaluation of each model's consistency with available data. Hence, the posterior probabilities are valid only in a comparative, not in an absolute, sense. They are conditional on the choice of models (in addition to being conditional on the data) and may be sensitive to the choice of prior model probabilities (as demonstrated later by example). This sensitivity is expected to diminish with increased level of conditioning on data.

3.1.2 Specifying Prior Model Probability

Given a set of alternative models, **M**, one formally assumes that their prior probabilities sum up to one,

$$\sum_{k=1}^{K} p(M_k) = 1 . \qquad (6)$$

This implies that all possible models of relevance are included in **M** (the set is collectively exhaustive), and that all models in **M** differ from each other sufficiently to be considered mutually exclusive (the joint probability of any two models is zero), at the outset. Mutually exclusive models are not redundant; they produce different results for the same set of inputs. In practice, it may be impossible to demonstrate that the set of models is collectively exhaustive. In this case, model uncertainty may be underestimated, a condition implied by the fact that all probabilities computed using BMA are conditional on **M**, as stated previously.

With regard to prior model probability, when there is insufficient prior reason to prefer one model over another, a "reasonable 'neutral' choice" (Hoeting et al., 1999) is to assume that all models are *a priori* equally likely. Draper (1995) and George (1999) express concern that if two models are near equivalent as regards predictions (i.e., redundant), treating them as separate equally likely models amounts to giving double weight to a single model of which there are two slightly different versions, thereby "diluting" the predictive power of BMA. One way to minimize this effect is to eliminate at the outset models that are deemed potentially inferior. Another is to retain only models that are structurally distinct and non-collinear. Otherwise, one should consider reducing (diluting) the

prior probabilities assigned to models that are deemed closely related. We explore this idea through an example in the following chapter.

3.2 Maximum Likelihood Bayesian Model Averaging (MLBMA)

Computational difficulties in the BMA approach include the calculation of $p(\Delta|M_k, \boldsymbol{D})$ in (1) and $p(\boldsymbol{D}|M_k)$ in (3), which may require exhaustive Monte Carlo simulations of the prior parameter space $\boldsymbol{\theta}_k$ for each model. This may be computationally and hydrologically very demanding. Approximating $p(\Delta|M_k, \boldsymbol{D})$ by $p(\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D})$, where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood (ML) estimate of $\boldsymbol{\theta}_k$ based on the likelihood $p(\boldsymbol{D}|\boldsymbol{\theta}_k, M_k)$, was suggested by Taplin (1993) and was shown to be useful in the BMA context by Draper (1995), Raftery et al. (1996) and Volinsky et al. (1997).

Neuman (2002, 2003) proposed evaluating the posterior model probability, $p(M_k|\boldsymbol{D})$, based on a result due to Kashyap (1982) and referred to the resulting method as Maximum Likelihood BMA (MLBMA). Kashyap derived an expression for $p(M_k|\boldsymbol{D})$ by expanding the terms in the integrand of (3) in a Taylor series about $\hat{\boldsymbol{\theta}}_k$. A related approach based on Laplace approximations has been used in the BMA context by Draper (1995) and Kass and Raftery (1995). Kashyap's expression can be written (Ye et al., 2003) as

$$p(M_k|\mathbf{D}) = \frac{\exp\left(-\frac{1}{2}\Delta KIC_k\right) p(M_k)}{\sum_{l=1}^{K} \exp\left(-\frac{1}{2}\Delta KIC_l\right) p(M_l)} \qquad (7)$$

where

$$\Delta KIC_k = KIC_k - KIC_{\min} , \qquad (8)$$

$$KIC_k = NLL_k + N_k \ln\left(\frac{N}{2\pi}\right) + \ln\left|\boldsymbol{F}_k\left(\mathbf{D}\big|\hat{\boldsymbol{\theta}}_k, M_k\right)\right| \quad (9)$$

KIC_k is the so-called Kashyap information criterion for model M_k, KIC_{min} is its minimum value over all candidate models, and $NLL_k = -2\ln p\left(\mathbf{D}\big|\hat{\boldsymbol{\theta}}_k, M_k\right) - 2\ln p\left(\hat{\boldsymbol{\theta}}_k\big|M_k\right)$ the negative log likelihood of M_k evaluated at $\hat{\boldsymbol{\theta}}_k$. Here

N_k is the dimension of $\boldsymbol{\theta}_k$ (number of parameters associated with model M_k), N is the dimension of $\mathbf{D}$ (number of discrete data points), and $\mathbf{F}_k$ is the normalized (by N) observed (as opposed to ensemble mean) Fisher information matrix having components

$$F_{k,ij} = -\frac{1}{N}\frac{\partial^2 \ln p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)}{\partial\theta_i\partial\theta_j}\Bigg|_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_k} \qquad (10)$$

In the absence of prior information about the parameters, one simply drops the term $-2\ln p(\hat{\boldsymbol{\theta}}_k|M_k)$ from NLL_k. This reflects common practice in model calibration.

Increasing the number of parameters N_k allows $-\ln p(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k)$ to decrease and $N_k \ln N$ to increase. When N_k is large, the rate of decrease does not compensate for the rate of increase and KIC_k grows while $p(M_k|\mathbf{D})$ diminishes. This means that a more parsimonious model with fewer parameters is ranked higher and assigned a higher posterior probability. On the other hand, $-\ln p(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k)$ diminishes with N at a rate higher than linear so that as the latter grows, there may be an advantage to a more complex model with larger N_k.

The last term in (9) reflects the information content of the available data. It thus enables consideration of models of growing complexity as the data base improves in quantity and quality. As illustrated by Carrera and Neuman (1986b), KIC_k recognizes that when the data base is limited and/or of poor quality, one has little justification for selecting an elaborate model with numerous parameters. Instead, one should prefer a simpler model with fewer parameters, which nevertheless reflects adequately the underlying hydrologic structure and regime of the system. Stated otherwise, KIC_k may cause one to prefer a simpler model that leads to a poorer fit with the data over a more complex model that fits the data better.

As shown in Ye et al. (2003), alternative models can have different types and numbers of parameters, but the latter must be estimated and the models compared considering a single data set $\mathbf{D}$. For a comparison of two- and three-dimensional models, data distributed in three-dimensional space may need to be projected onto a two-dimensional plane as done by Ando et al. (2003) or averaged in the third dimension as suggested by Neuman and Wierenga (2003, Appendix B).

3.2.1   A Few Words About *KIC*

Previously, KIC_k has been used (e.g., Carrera and Neuman, 1986a,b; Samper and Neuman, 1989a,b) as an optimum decision rule for the ranking of competing models. The highest-ranking model is that corresponding to KIC_{min}. Note that KIC has no intrinsic meaning; it is only the differences between KIC values that have meaning. Thus the use of ΔKIC in (7) reflects the interpretation of $p(M_k|\mathbf{D})$ as a relative probability suitable for comparing the models within the set $\mathbf{M}$.

The Fisher information matrix term in (9) tends to a constant as N becomes large, so that KIC_k becomes asymptotically equivalent to the Bayes information criterion

$$BIC_k = NLL_k + N_k \ln N \qquad (11)$$

derived on the basis of other considerations by Akaike (1977), Rissanen (1978) and Schwarz (1978). Raftery (1993) proposed adopting the asymptotic *BIC* approximation, without the prior information term $-2\ln p(\hat{\boldsymbol{\theta}}_k|M_k)$, for BMA (see also Raftery et al. 1996; Volinsky et al. 1997; Hoeting et al. 1999). From (11) it follows that (7) tends asymptotically to

$$p(M_k|\mathbf{D}) = \frac{\exp\left(-\frac{1}{2}\Delta BIC_k\right)p(M_k)}{\sum_{l=1}^{K}\exp\left(-\frac{1}{2}\Delta BIC_l\right)p(M_l)} \qquad (12)$$

where

$$\Delta BIC_k = BIC_k - BIC_{min} \qquad (13)$$

and BIC_{min} is the smallest value of BIC_k over all candidate models (see also Burnham and Anderson, 2002, pp. 297).

Since hydrologic models are often data limited, the asymptotic expression (12) is less general than the nonasymptotic expression (7) that is at the heart of MLBMA. Indeed, Carrera and Neuman (1986a,b) and Samper and Neuman (1989a,b) found KIC_k to provide more reliable rankings of alternative groundwater flow and geostatistical models than do BIC_k or two other commonly used information criteria:

$AIC_k = NLL_k + 2N_k$ (Akaike, 1974) and

$HIC_k = NLL_k + 2N_k \ln(\ln N)$ (Hannan, 1980).

For a recent overview of various information criteria the reader is referred to Burnham and Anderson (2002, p. 284).

3.2.2 Applicability of MLBMA

Using the maximum likelihood method has several advantages. It can be applied to both complex and simplified models. It can be applied to deterministic models as described by Carrera and Neuman (1986a,b) and Carrera et al. (1997) and also to stochastic models based on moment equations as demonstrated by Hernandez et al. (2002, 2003). Application of maximum likelihood also yields parameter sensitivity information.

Including prior information in the maximum likelihood calibration is an option that allows one to condition the parameter estimates not only on site monitoring (observational) data but also on site characterization data, from which prior parameter estimates are usually derived. When both sets of data are considered to be statistically meaningful, the posterior parameter estimates are compatible with a wider array of measurements than they would be otherwise and are therefore better constrained (potentially rendering the model a better predictor).

Maximum likelihood yields a negative log likelihood criterion NLL_k that includes two weighted square residual terms: a generalized sum of squared differences between simulated and observed state variables arising from $-2\ln p\left(\mathbf{D}\middle|\hat{\mathbf{\theta}}_k, M_k\right)$, and a generalized sum of squared differences between posterior and prior parameter estimates arising from $-2\ln p\left(\hat{\mathbf{\theta}}_k\middle|M_k\right)$. The first is weighted by a matrix proportional to the inverse covariance matrix of state observation errors. The second is weighted by a matrix proportional to the inverse covariance matrix of prior parameter estimation errors. Maximum likelihood allows the statistical parameters of the errors to be estimated. When these statistical parameters are known (i.e., not estimated), maximum likelihood reduces to generalized least squares estimation. In this case, available codes such as PEST (Doherty, 2002) and UCODE (Poeter and Hill, 1998) can be applied.

Maximum likelihood estimation yields an approximate covariance matrix for the estimation errors of $\hat{\mathbf{\theta}}_k$. Upon considering the parameter estimation errors of a calibrated deterministic model M_k to be Gaussian or log Gaussian, one easily determines $p\left(\Delta\middle|M_k, \hat{\mathbf{\theta}}_k, \mathbf{D}\right)$

by Monte Carlo simulation of Δ through random perturbation of the parameters. The simulation also yields corresponding approximations $E\left[\Delta\middle|M_k, \hat{\mathbf{\theta}}_k, \mathbf{D}\right]$ of $E\left[\Delta\middle|M_k, \mathbf{D}\right]$, and $Var\left[\Delta\middle|M_k, \hat{\mathbf{\theta}}_k, \mathbf{D}\right]$ of $Var\left[\Delta\middle|M_k, \mathbf{D}\right]$, in (4) and (5). If M_k is a geostatistical model (as in the example below) or a stochastic moment model (of the kind considered by Hernandez et al. (2002, 2003), it yields $E\left[\Delta\middle|M_k, \hat{\mathbf{\theta}}_k, \mathbf{D}\right]$ and $Var\left[\Delta\middle|M_k, \hat{\mathbf{\theta}}_k, \mathbf{D}\right]$ directly without Monte Carlo simulation.

One final point regarding the applicability of MLBMA. In the most data-limited application, one in which there are no system observations with which to calibrate a model and the only available parameter information is that available from generic databases, Equation 1 reduces to

$$p(\Delta) = \sum_{k=1}^{K} p\left(\Delta\middle|M_k\right) p\left(M_k\right).$$

That is, model predictions can still be made using prior (or updated) parameter estimates (see Figure 2-3) and model averaging can still be carried out, but only with prior model probabilities. Since the predictions and model probabilities are not conditioned on state variable observations, however, the results are expected to be more uncertain and potentially more biased.

3.3 Summary of MLBMA

To implement MLBMA the following steps are followed.

(1) Postulate alternative conceptual-mathematical models for a site using guidance provided in Neuman and Wierenga (2003).

(2) Assign a prior probability to each model.

(3) Optionally assign prior probabilities to the parameters of each model, using, for example, guidance provided in Meyer and Gee (1999).

(4) Obtain posterior maximum likelihood parameter estimates, and estimation covariance, for each model by inversion (model calibration). In many cases, available codes such as PEST (Doherty, 2002) and UCODE (Poeter and Hill, 1998) can be applied to this step.

(5) Calculate a posterior probability for each model using the model calibration results and the prior model probabilities as expressed in Equations 7 to 9.

(6) Predict quantities of interest using each model.

(7) Assess prediction uncertainty (distribution, variance) for each model using Monte Carlo or stochastic moment methods.

(8) Weight predictions and uncertainties by the corresponding posterior model probabilities.

(9) Sum the results over all models.

A flowchart illustrating the MLBMA approach to combined estimation of conceptual model and parameter uncertainty is shown in Figure 3-1. Numbers in parentheses above the boxes refer to the numbered steps above.

The following chapter provides an example application of MLBMA and an evaluation of its performance and suitability.
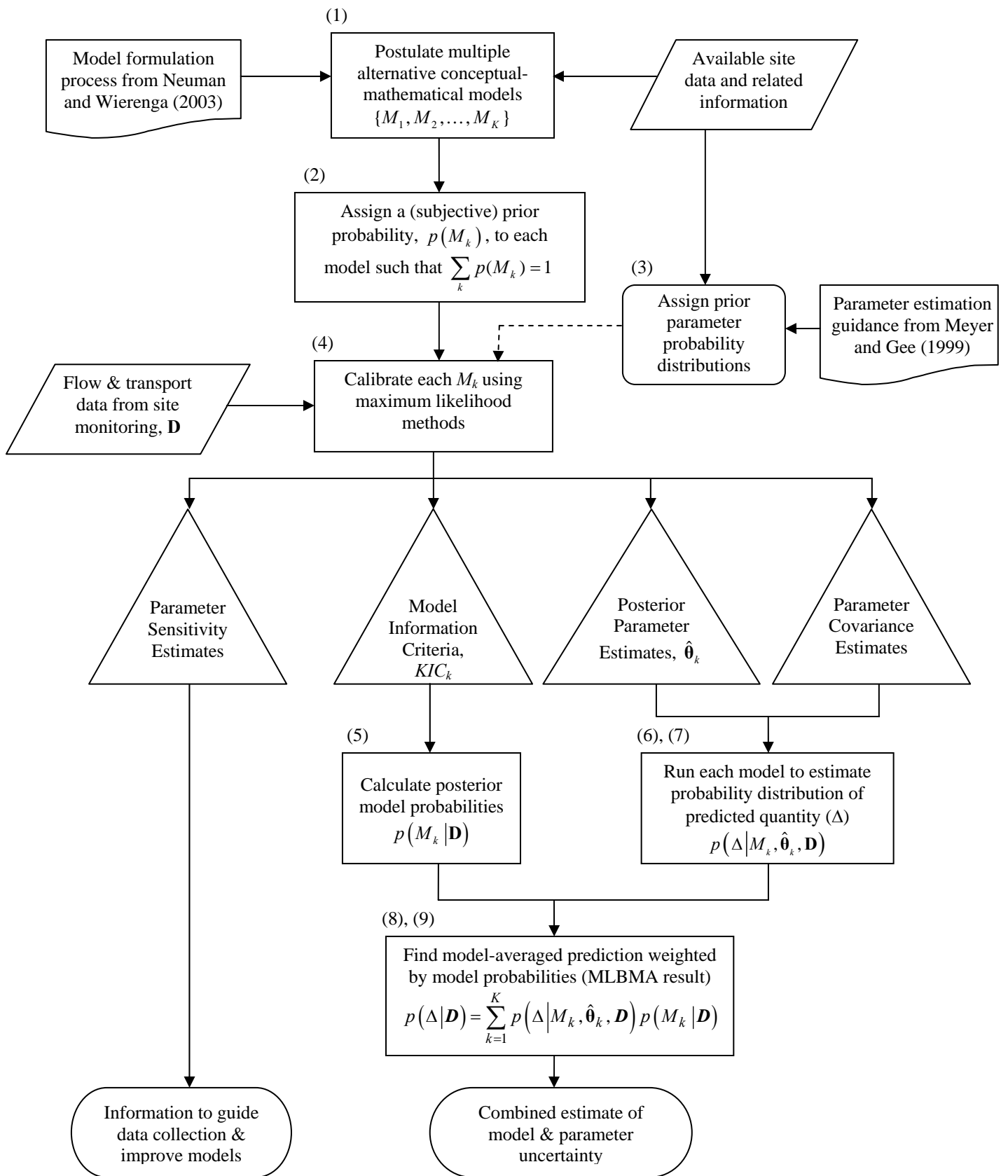
Figure 3-1. Maximum Likelihood Bayesian Model Averaging (MLBMA) approach to combined estimation of model and parameter uncertainty

4   Example Application

4.1   Implementation of MLBMA

To demonstrate the application of MLBMA and to evaluate the results, we apply it to alternative geostatistical models of log air permeability variations in unsaturated fractured tuff at the Apache Leap Research Site (ALRS) in central Arizona. This site was chosen for an initial application of MLBMA because it is a relatively well-controlled experimental research site with good characterization data. The results of past studies at the site were available to us as well, which facilitated the application of MLBMA. In addition, the models considered (geostatistical models of permeability) are relatively simple, thus reducing the computational effort required to complete the application. We recognize that an example considering groundwater flow and transport would better reflect NRC-regulated sites. However, we see no fundamental barrier in applying MLBMA to the more complex models required in such applications. Any difficulties in applying MLBMA to groundwater flow and transport applications will be explored in a case-study using actual field data that is the focus of future efforts.

4.2   ALRS Data and Previous Efforts

Spatially distributed log air permeability data were obtained by Guzman et al. (1994, 1996) based on a steady state interpretation of 184 pneumatic injection tests in 1-m-length intervals along 6 vertical and inclined (at 45°) boreholes at the site (Figure 4-1). Five of the boreholes (V2, W2A, X2, Y2, Z2) are 30-m long and one (Y3) has a length of 45 m; five (W2A, X2, Y2, Y3, Z2) are inclined at 45° and one (V2) is vertical.

Figure 4-2 shows an omni-directional sample variogram of corresponding $\log_{10} k$ data. Chen et al. (2000) fitted three variogram models to these and some 3-m-scale data using an adjoint state maximum likelihood cross-validation (ASMLCV) method developed for this purpose by Samper and Neuman (1989a,b), coupled with a generalized least squares (GLS) drift removal approach due to Neuman and Jacobson (1984). The three models included (1) power (characteristic of a random fractal), (2) exponential with a linear drift, and (3) exponential with a quadratic drift. The data did not support accounting for directional effects by considering the variograms to be anisotropic.

The authors found that whereas the exponential variogram model with a quadratic drift provided a best fit to the data (as measured and implied by the smallest negative log-likelihood model fit criterion, *NLL*), four model discrimination criteria (*AIC*, *BIC*, *HIC*, *KIC*) consistently ranked the power model as best, and the former model as least acceptable. The reason was that
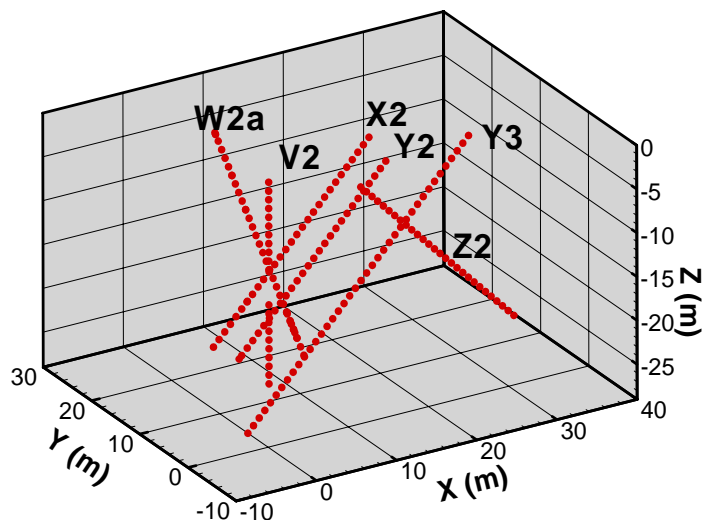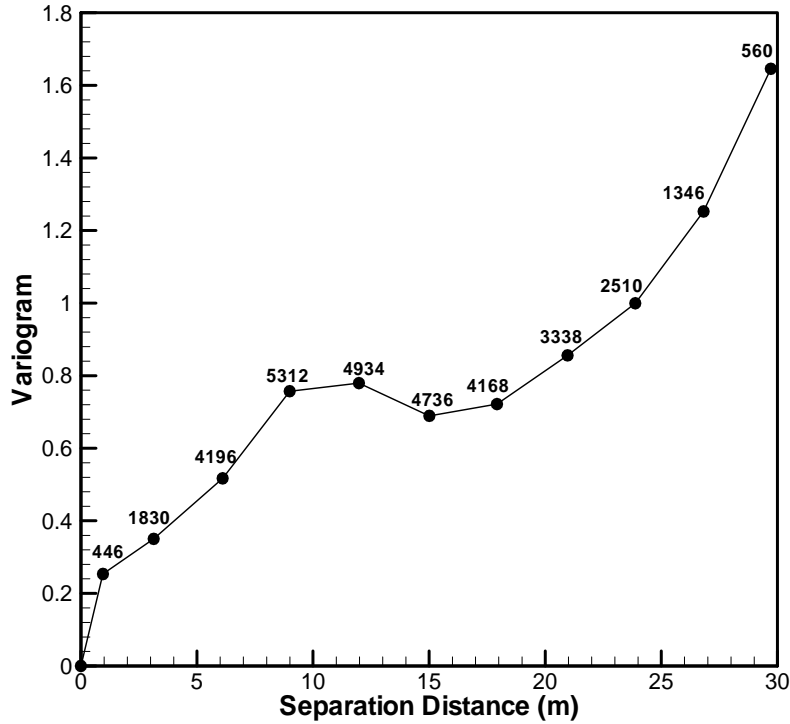

Figure 4-1. Spatial locations of 184 1-m-scale log$_{10}$k data at ALRS

Figure 4-2. Omni-directional sample variogram of 1-m-scale $\log_{10}k$ data at the ALRS and numbers of data pairs

whereas all three models provided an almost equally good fit to the data, the power model was most parsimonious with only two parameters, and the exponential variogram model with second-order drift was least parsimonious with twelve parameters. They therefore adopted the power model and discarded all other variogram models from further consideration.

4.2.1 Alternative Models and Maximum Likelihood Parameter Estimation

For purposes of MLBMA we expand the range of variogram models postulated for 1-m-scale $\log_{10}k$ at the ALRS to seven: (1) Power (*Pow0*), (2) exponential without a drift (*Exp0*), (3) exponential with a linear drift (*Exp1*), (4) exponential with a quadratic drift (*Exp2*), (5) spherical without a drift (*Sph0*), (6) spherical with a linear drift (*Sph1*), and (7) spherical with a quadratic drift (*Sph2*).

To estimate the parameter vector $\boldsymbol{\beta}$ of drift-free variogram models (*Pow0*, *Exp0*, *Sph0*) we use ASMLCV as described in Ye et al. (2003), implemented in a computer code slightly modified after Samper (1998, personal communication). To do the same for models with drift (*Exp1*, *Exp2*, *Sph1*, *Sph2*), we decompose the *N*-dimensional data vector **D** of

$\log_{10}k$ measurements into a deterministic drift vector $\boldsymbol{\mu}$ and a random residual vector **R**,

$$\mathbf{D} = \boldsymbol{\mu} + \mathbf{R} \qquad (14)$$

$$\boldsymbol{\mu}(\mathbf{x}) = \sum_{k=0}^{p} g_k(\mathbf{x})a_k = \mathbf{Ga} \qquad (15)$$

where $\mathbf{a} = \left(a_0, a_1, ..., a_p\right)^T$ is a vector of $p+1$ drift coefficients and **G** is a $N \times (p+1)$ matrix of linearly independent monomial functions $g_k(\mathbf{x})$ evaluated at the data points $\mathbf{x}_n$, $n = 1, 2, ..., N$.

Assuming that **D** is multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}_R$ (Vesselinov 2000 has shown that the data pass the Kolmogorof-Smirnov test of univariate Gaussianity at a significance level of 0.05), the joint negative log likelihood function of drift and variogram parameters takes the form

$$NLL(\mathbf{a}, \boldsymbol{\beta} \mid \mathbf{D}) = -2 \ln p(\mathbf{D} \mid \mathbf{a}, \boldsymbol{\beta})$$
$$= N \ln 2\pi + \ln\left|\mathbf{C}_R(\boldsymbol{\beta})\right| + (\mathbf{D} - \mathbf{Ga})^T \mathbf{C}_R^{-1}(\boldsymbol{\beta})(\mathbf{D} - \mathbf{Ga}) \ (16)$$

Minimizing (16) jointly with respect to **a** and $\boldsymbol{\beta}$ yields biased estimates of the variogram parameters, a problem that can be remedied through the use of a

restricted ML (RML) approach (Hoeksema and Kitanidis, 1985; Kitanidis and Lane, 1985; Cressie, 1991, p. 92). We solve the problem differently by formally decoupling the ML estimations of **a** and **β**. First, we obtain unbiased ML estimates $\hat{\boldsymbol{\beta}}$ of the variogram parameters using ASMLCV in conjunction with universal kriging (ASMLCV-UK, Samper 1998, personal comm.), which does not require knowledge of the drift coefficients (Ye et al., 2003). Next, we compute corresponding unbiased ML estimates $\hat{\mathbf{a}}$ of the drift coefficients through minimization of

$$NLL(\mathbf{a}, \hat{\boldsymbol{\beta}} \mid \mathbf{D})$$
$$= N \ln 2\pi + \ln\left|\mathbf{C}_R(\hat{\boldsymbol{\beta}})\right| + (\mathbf{D} - \mathbf{Ga})^T \mathbf{C}_R^{-1}(\hat{\boldsymbol{\beta}})(\mathbf{D} - \mathbf{Ga}) \quad (17)$$

with respect to **a** by generalized least squares, a task we accomplish using PEST-ASP (Doherty, 2002). Our optimum *NLL* is then given by

$$NLL(\hat{\mathbf{a}}, \hat{\boldsymbol{\beta}} \mid \mathbf{D})$$
$$= N \ln 2\pi + \ln\left|\mathbf{C}_R(\hat{\boldsymbol{\beta}})\right| + (\mathbf{D} - \mathbf{G}\hat{\mathbf{a}})^T \mathbf{C}_R^{-1}(\hat{\boldsymbol{\beta}})(\mathbf{D} - \mathbf{G}\hat{\mathbf{a}}) \quad .(18)$$

Figure 4-3 depicts profiles of $NLL(\mathbf{a}, \boldsymbol{\beta} \mid \mathbf{D})$ in (16) versus each parameter of model *Exp1* when the remaining parameters are fixed. It clearly demonstrates that $\hat{\boldsymbol{\beta}}$ (the marked values of sill and integral scale [m] does not correspond to the minimum of $NLL(\mathbf{a}, \boldsymbol{\beta} \mid \mathbf{D})$, which would therefore yield biased estimates of variogram parameters.

The estimation covariance matrix of $\hat{\boldsymbol{\theta}} = \left(\hat{\mathbf{a}}, \hat{\boldsymbol{\beta}}\right)^T$ is generally represented by its asymptotic lower or Cramer-Rao bound, given by the inverse Fisher information matrix (e.g., Carrera et al., 1997). Components of the observed Fisher information matrix (10) are proportional to those of the Hessian matrix **H** which, in turn, can be approximated as (Kitanidis and Lane, 1985)

$$H_{k,ij} = -\left.\frac{\partial^2 \ln p\left(\mathbf{D} \mid \boldsymbol{\theta}_k, M_k\right)}{\partial \theta_i \partial \theta_j}\right|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k}$$
$$\approx \frac{1}{2} Tr\left(\mathbf{C}_R^{-1} \frac{\partial \mathbf{C}_R}{\partial \boldsymbol{\theta}_i} \mathbf{C}_R^{-1} \frac{\partial \mathbf{C}_R}{\partial \boldsymbol{\theta}_j}\right) + \left.\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}_i} \mathbf{C}_R^{-1} \frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}_j}\right|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k} \quad (19)$$

This approximation obviates the need to calculate second-order derivatives of the log likelihood function, which would be computationally more demanding than computing first-order derivatives of $\mathbf{C}_R$ and **R**. In our

case, the latter are easy to obtain analytically as done for exponential and spherical variogram models with drift (Ye, et al., 2003). An alternative, which in our case yields very similar results, is to compute the observed Fisher information matrix numerically using methods such as the Ridder algorithm (Press et al., 1992, pp. 180).

4.2.2 Posterior Model Probabilities

Table 4-1 confirms that increasing the number of parameters associated with a given class of variogram model (exponential or spherical) brings about an improvement in model fit, as indicated by a reduction in the negative log likelihood criterion *NLL*. Whereas the exponential variogram model with a quadratic drift (*Exp2*) fits the data best (ranks first in terms of fit due to its smallest *NLL* value), it is ranked second by *AIC* and sixth by *BIC* and *KIC*. Whereas the power model (*Pow0*) shows a relatively poor fit with the data (rating fifth), it is ranked highly (first through third) by all three information criteria. The reason is that the difference in fit between the two models is not enough to compensate for the much more parsimonious nature of *Pow0* (with 2 parameters) than that of *Exp2* (with 12 parameters).

The rankings of the seven models by *AIC*, *BIC* and *KIC* are not entirely consistent. None of these information criteria provide justification for retaining one model while discarding all other models as is commonly done in practice. Nor do they provide clear justification for retaining some models while discarding the rest. We therefore consider all seven models to be valid initial candidates for MLBMA.

Upon assigning an equal prior probability of 1/7 to each model, we find on the basis of *KIC* via (7) that the first three models (*Pow0*, *Exp0*, *Exp1*) have much higher posterior probabilities than do the rest. Three of the models (*Exp2*, *Sph0*, *Sph2*) have zero probabilities (to three significant figures) and can therefore be eliminated (considering the low posterior probability of *Sph1*, there is almost equal justification for eliminating it too, but we retain it at this stage for the sake of illustration). Doing so and assigning an equal prior probability of 1/4 to each of the retained models is seen to have no impact on their posterior probabilities. In both cases the posterior probabilities are markedly different from their prior values, reflecting the strong impact of conditioning on data.

4.2.2.1 Sensitivity to Prior Model Probabilities

To investigate the influence of prior probability selection on the outcome, consider assigning an equal

probability of 1/3 to each of the three classes of models (power, exponential and spherical) and also assigning equal probability to models within each class. This results in a prior probability of 1/3 for *Pow0* and of 1/9 for each of the other six models. Though this brings about a marked increase in the posterior probability of *Pow0* and a decrease in those of *Exp0* and *Exp1*, once again the posterior probabilities of *Exp2*, *Sph0* and *Sph2* are zero while that of *Sph1* is very close to zero. Eliminating the three models with zero posterior probability and redistributing the prior probabilities among the remaining models as shown in the next-to-

last row of Table 4-1 brings about a decrease in the posterior probability of *Pow0* and an increase in the posterior probabilities of *Exp0* and *Exp1*. We conclude that posterior model probabilities exhibit some degree of sensitivity to the choice of prior probabilities but expect this sensitivity to diminish with improved conditioning.

4.2.3 Kriging Results

We continue our analysis by retaining four (*Pow0*, *Exp0*, *Exp1*, *Sph1*) of the seven models (with the
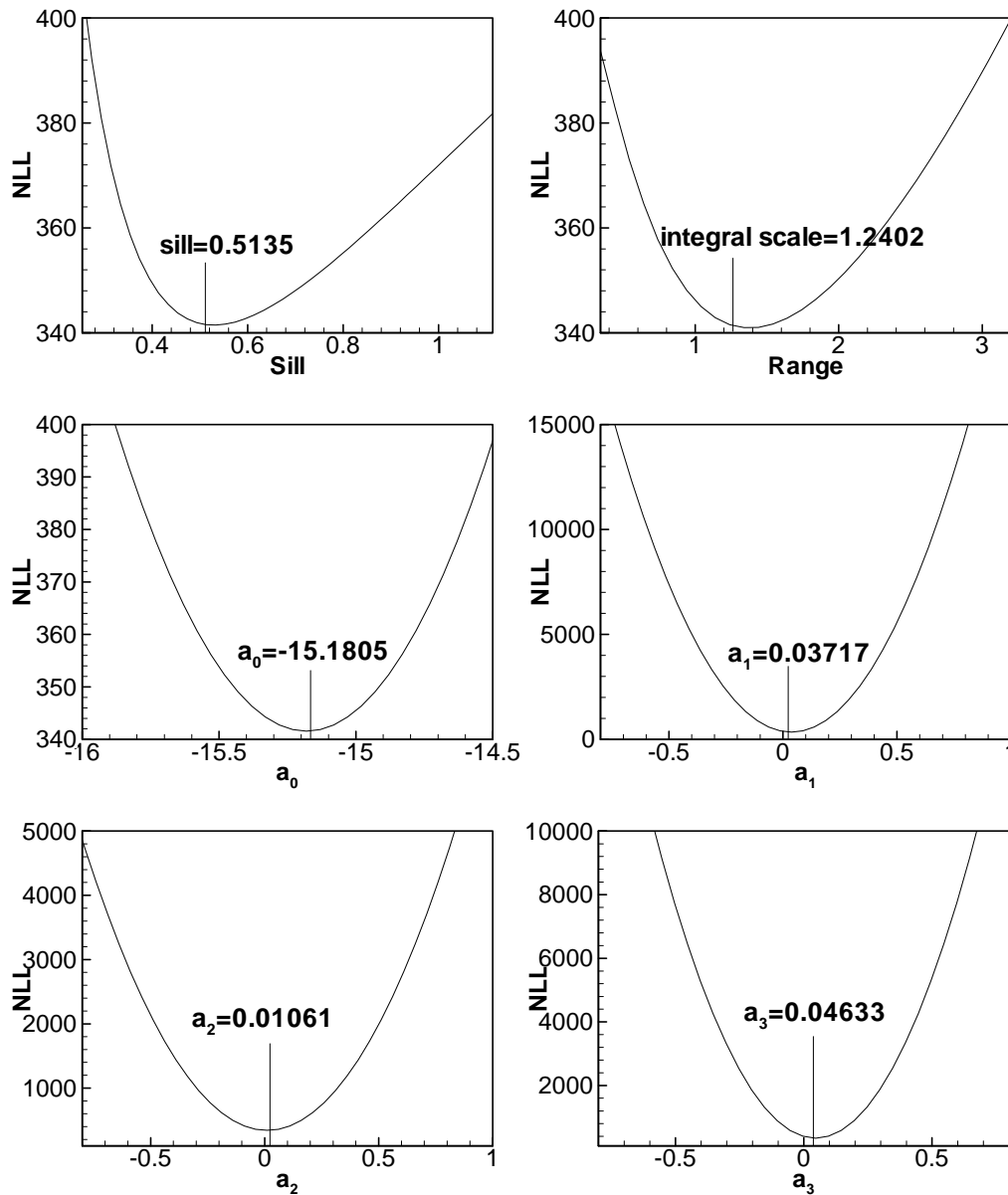

Figure 4-3. Negative log likelihood functions (*NLL*) as function of each variogram parameter and drift coefficient for exponential model with linear drift (*Exp1*). Vertical lines indicate unbiased ML estimates.

20

Table 4-1. Quality criteria, rankings and prior/posterior probabilities associated with alternative geostatistical models of $\log_{10}k$ at the ALRS

Model	Pow0	Exp0	Exp1	Exp2	Sph0	Sph1	Sph2
Number of parameters	2	2	6	12	2	6	12
Sill/Coefficient	0.286	0.718	0.514	0.501	0.749	0.664	0.662
Correlation/Power	0.460	1.840	1.240	1.198	3.184	2.849	2.835
NLL	352.2	361.0	341.6	330.4	379.1	349.6	338.8
Rank	5	6	3	1	7	4	2
AIC	356.2	365.0	353.6	354.4	383.1	361.6	362.8
Rank	3	6	1	2	7	5	4
BIC	362.6	371.4	372.9	392.9	389.5	380.9	401.4
Rank	1	2	3	6	5	4	7
KIC	369.6	370.1	369.5	416.7	390.5	378.1	424.6
Rank	2	3	1	6	5	4	7
$p(M_k)$	1/7	1/7	1/7	1/7	1/7	1/7	1/7
$p(M_k|\mathbf{D})$(%)	35.3	26.6	37.6	0	0	0.5	0
$p(M_k)$	1/4	1/4	1/4	-	-	1/4	-
$p(M_k|\mathbf{D})$(%)	35.3	26.6	37.6	-	-	0.5	-
$p(M_k)$	1/3	1/9	1/9	1/9	1/9	1/9	1/9
$p(M_k|\mathbf{D})$(%)	62.1	15.6	22.0	0	0	0.3	0
$p(M_k)$	1/3	1/6	1/6	-	-	1/3	-
$p(M_k|\mathbf{D})$(%)	52.0	19.6	27.7	-	-	0.7	-

corresponding ML parameter estimates) and assigning to each of them an equal prior probability of 1/4. Using each of these models, we project the available $\log_{10}k$ data by ordinary (in the case of drift-free models) or universal (otherwise) kriging onto a grid of $50 \times 40 \times 30$ 1-m^3 cubes contained within the coordinate ranges $-10 \le x \le 40$ m, $-10 \le y \le 30$ m and $-30 \le z \le 0$ m in Figure 4-1.

If one thinks of Δ as a random value of $\log_{10}k$ in a given grid block then our kriging estimates represent $E[\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}]$ and their variances stand for $Var[\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}]$, the ML approximations of $E[\Delta|M_k, \mathbf{D}]$ and $Var[\Delta|M_k, \mathbf{D}]$ in (4) and (5), respectively.

Figure 4-4 to Figure 4-7 show the kriged estimates and variances of $\log_{10}k$ on a vertical plane $y = 6.5$ m for the four models. Conditioning on borehole data is evident to a lesser degree in the images of $\log_{10}k$ estimates than in those of their variances. Averaging the kriging results across all models using an ML approximation of (4) and (5) yields corresponding MLBMA estimates and variances of the kind depicted for $y = 6.5$ m in Figure 4-8. Figure 4-9 shows a decomposition of the MLBMA estimation variance in Figure 4-8b into its within- and between-model components. The largest values of these two components throughout the three-dimensional grid are 1.1 and 0.38, respectively. Whereas the within-model MLBMA variance in Figure 4-9a reflects conditioning on borehole measurements, it is difficult to discern such conditioning in the image of between-model variance (Figure 4-9b) due to the faint

reflection of such conditioning in the underlying images of $\log_{10}k$ estimates.

Figure 4-10 shows univariate cumulative distributions of kriging estimates corresponding to each of the four models and MLBMA. The distributions are seen to be sensitive to the choice of model with MLBMA providing a weighted compromise. The same is reflected in the variances of these kriged estimates, listed in Table 4-2.

Table 4-2. Variance of kriged estimates across the grid obtained with alternative models and MLBMA

Model	Variance
Pow0	0.334
Exp0	0.134
Exp1	0.467
Sph1	0.404
MLBMA	0.405

(a) Kriged estimate
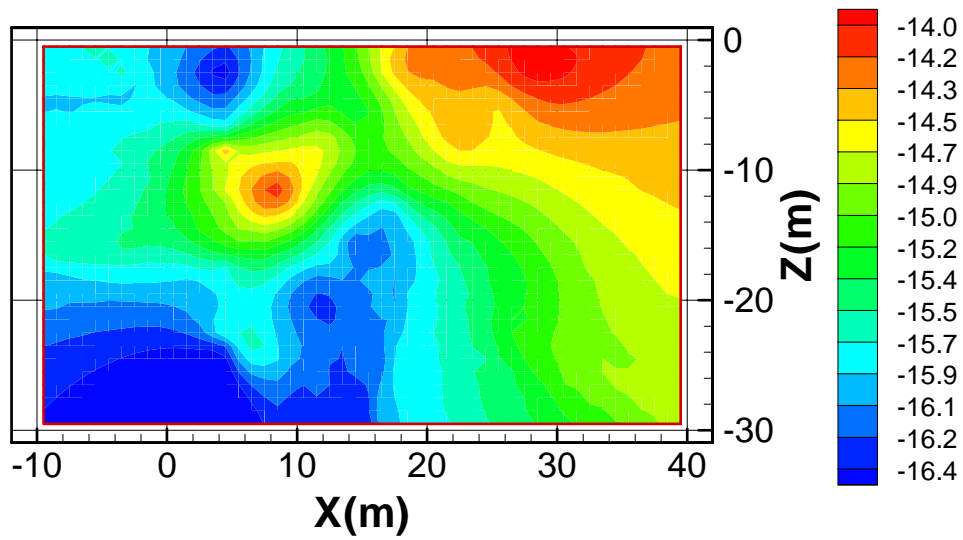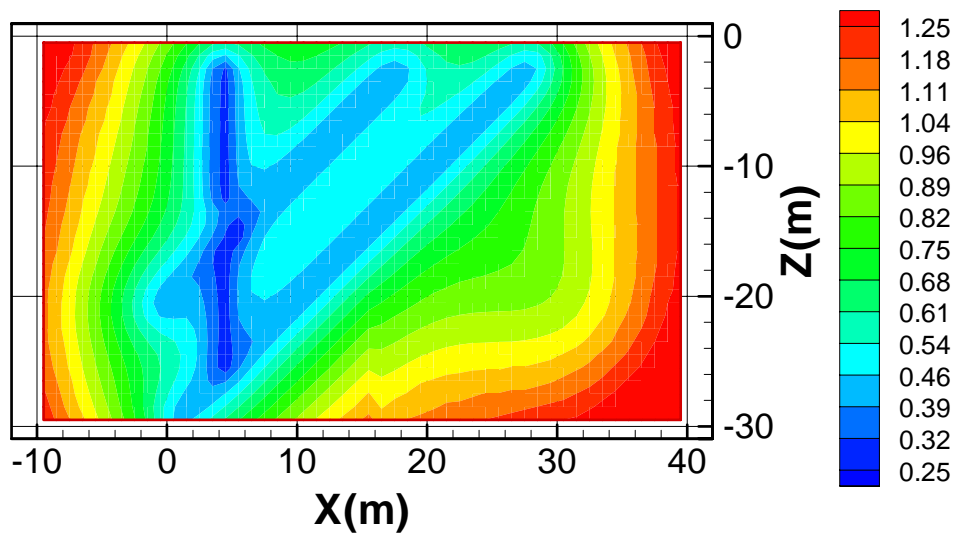
(b) Kriged variance

Figure 4-4. Kriged (a) estimate and (b) variance of log$_{10}$k at y = 6.5 m obtained using the power model (*Pow0*)

(a) Kriged estimate
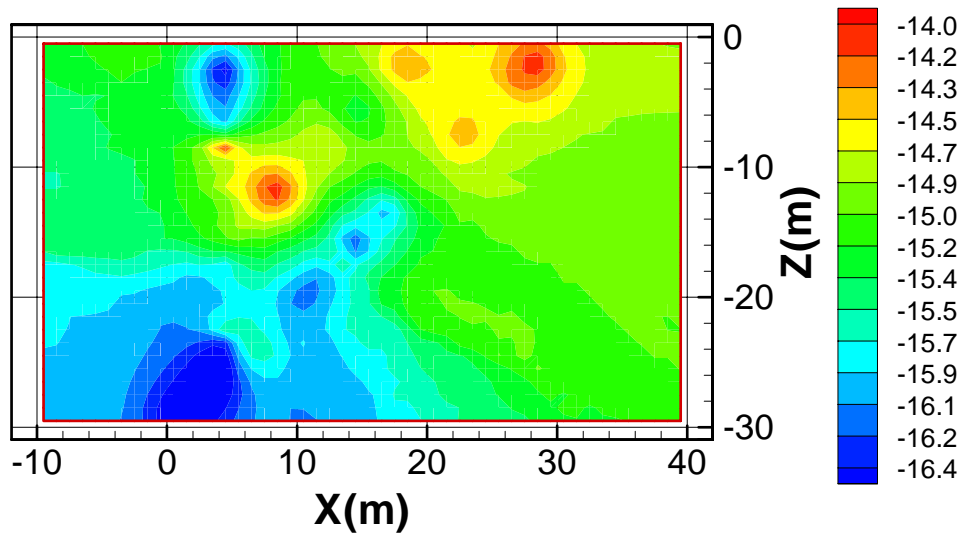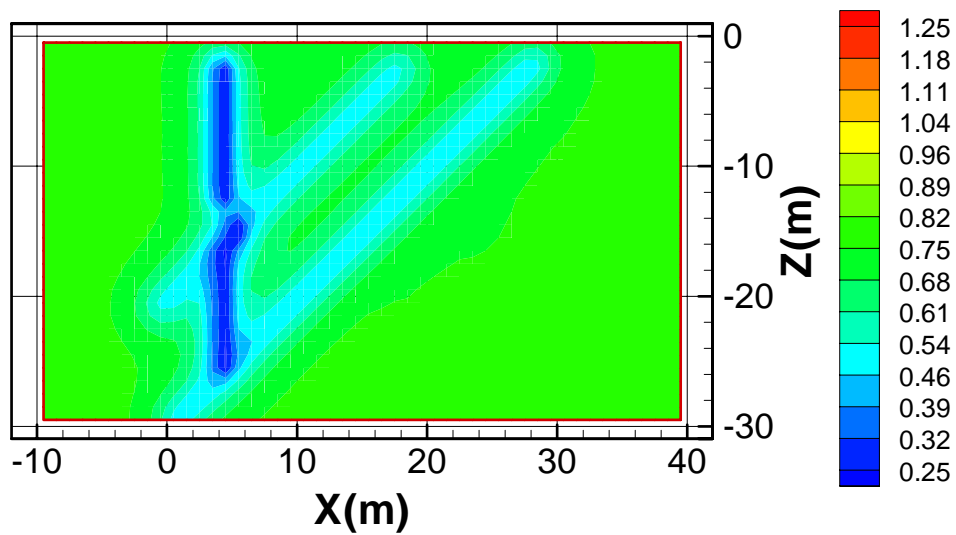
(b) Kriged variance

Figure 4-5. Kriged (a) estimate and (b) variance of $\log_{10}k$ at $y = 6.5$ m obtained using the exponential model without drift (*Exp0*)
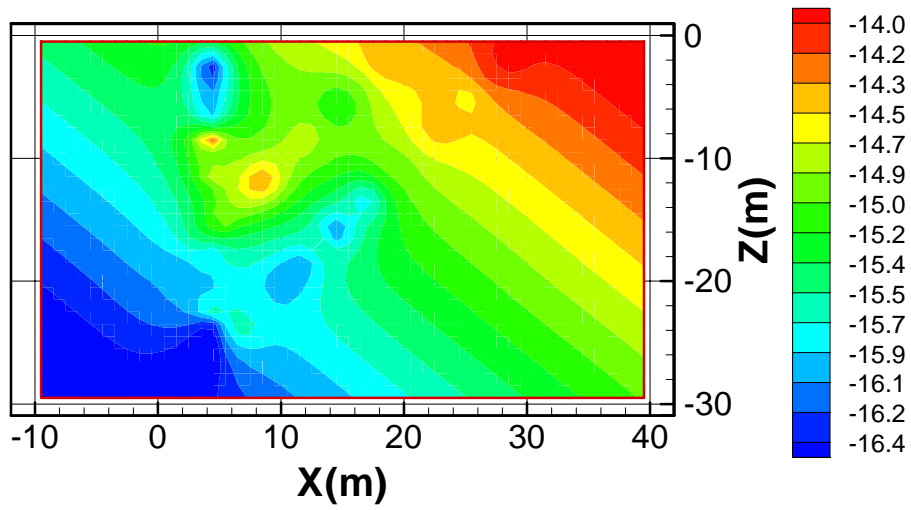
(a) Kriged estimate
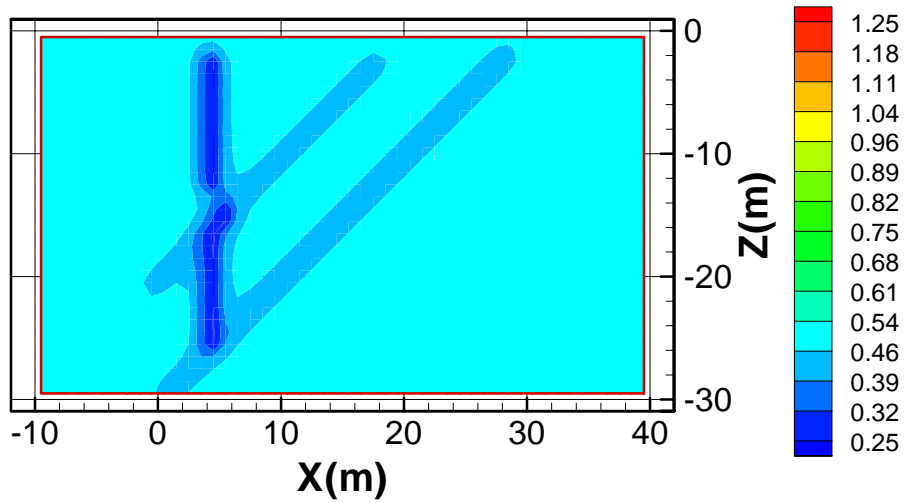
(b) Kriged variance

Figure 4-6. Kriged (a) estimate and (b) variance of $\log_{10}k$ at $y = 6.5$ m obtained using the exponential model with first-order drift (*Exp1*)
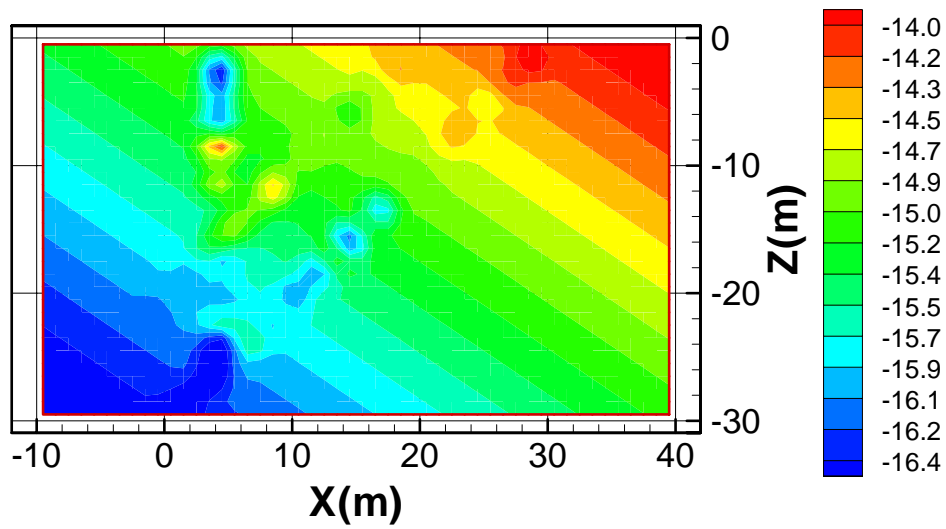
(a) Kriged estimate
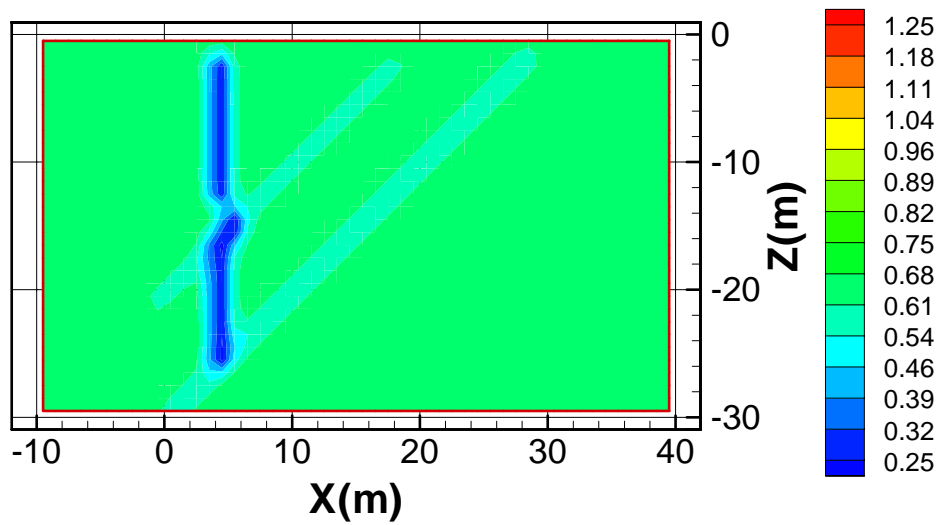
(b) Kriged variance

Figure 4-7. Kriged (a) estimate and (b) variance of log$_{10}$k at y = 6.5 m obtained using the spherical model with first-order drift (*Sph1*)

(a) Kriged estimate
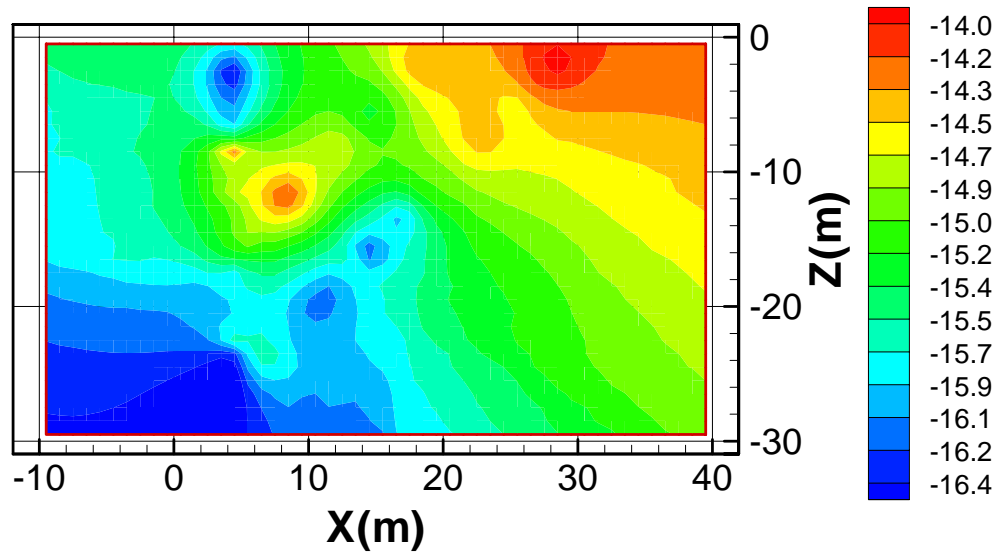
(b) Kriged variance

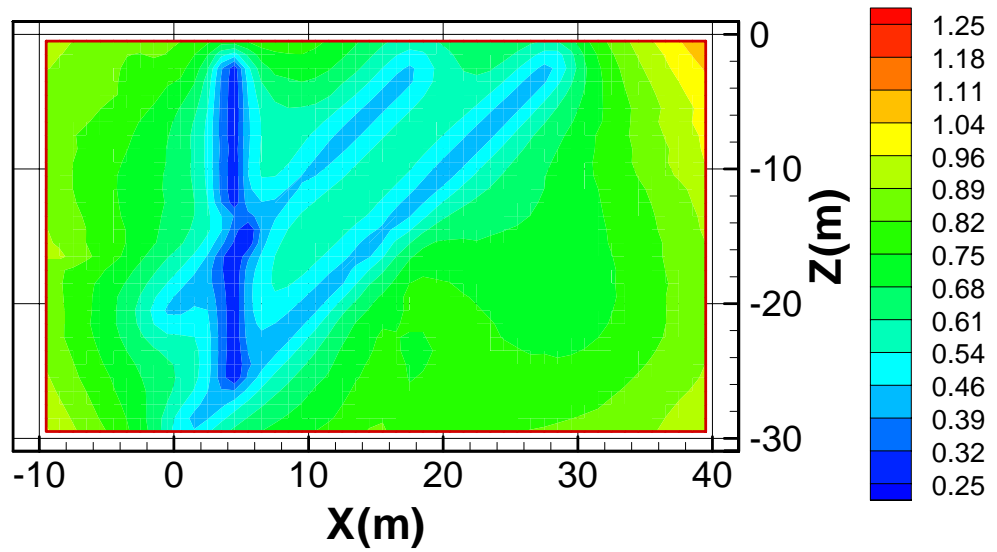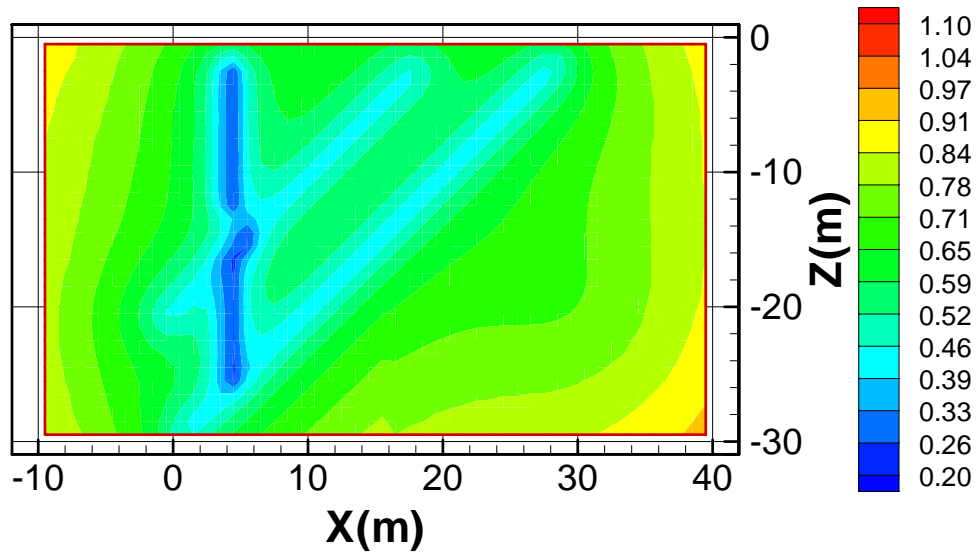Figure 4-8. Kriged (a) estimate and (b) variance of $\log_{10}k$ at $y = 6.5$ m obtained using MLBMA

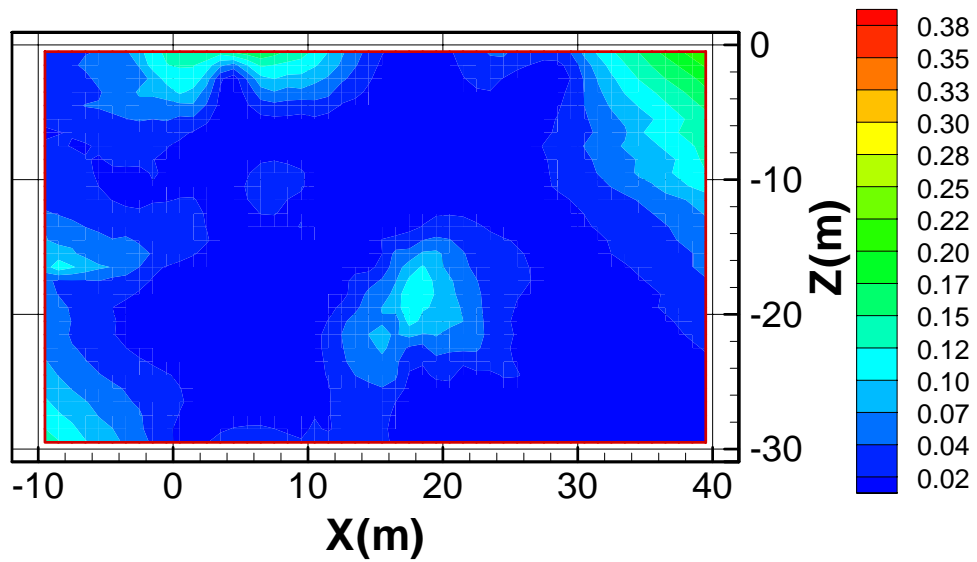# (a) Within-model variance

(b) Between-model variance

Figure 4-9. (a) Within- and (b) between-model variance of MLBMA $\log_{10}k$ estimates at $y = 6.5$ m
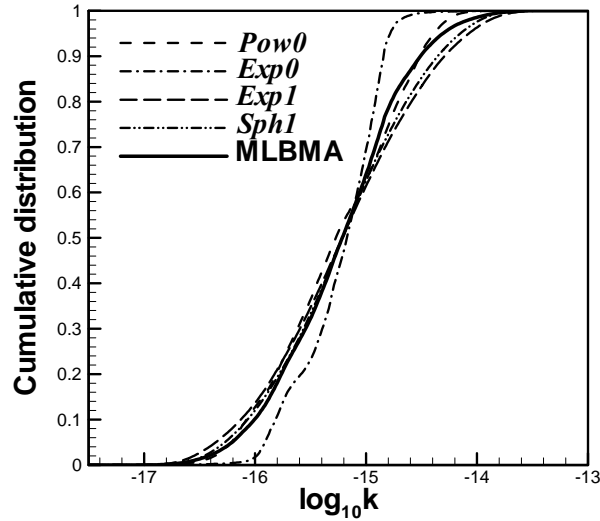
Figure 4-10. Cumulative distribution of kriged $\log_{10}k$ estimates obtained using various models and MLBMA

4.3 Assessment of Predictive Performance

To assess the predictive performance of MLBMA, we cross-validate the above results by (1) splitting the data **D** into two parts, $\mathbf{D}^A$ and $\mathbf{D}^B$; (2) obtaining ML estimates of model parameters and posterior probabilities conditional on $\mathbf{D}^A$; (3) using these to render MLBMA predictions $\hat{\mathbf{D}}^B$ of $\mathbf{D}^B$; and (4) assessing the quality of the predictions. We do so by eliminating from consideration all $\log_{10}k$ data from one borehole at a time and predicting them with models conditioned on the remaining data. The number and corresponding percentage of data in $\mathbf{D}^A$ for each cross validation case are listed in Table 4-3. As *Sph1* has a very small posterior probability in comparison to *Pow0*, *Exp0,* and *Exp1* (Table 4-1), we limit the cross-validation to the latter three geostatistical models and recalculate their posterior probabilities by assigning to each of them a prior probability of 1/3.

Figure 4-11 shows that eliminating data from one borehole at a time may, but need not, have a significant impact on the omni-directional sample variogram of $\log_{10}k$. The impact that such elimination has on parameter estimates and model quality criteria associated with *Pow0* is indicated in Figure 4-12. Figure 4-13 demonstrates that posterior model probability is sensitive to the choice of conditioning data. This sensitivity is greater when posterior probability is computed using *KIC* in (7) than *BIC* in (12). This illustrates that the non-asymptotic criterion *KIC* is more informative than the asymptotic criterion

Table 4-3. Number of $\log_{10}k$ data in D^A of each cross validation case and their percentage of the entire data set.

Well	Number	Percentage (%)
V2	163	89.1
X2	154	83.7
Y2	156	84.8
Y3	144	78.3
Z2	156	84.8
W2A	147	79.9

BIC, supporting the choice of the former as the basis for MLBMA (Neuman, 2002, 2003).

4.3.1 Predictive Log Score

One way to compare the predictive capabilities of alternative models is through their log scores, $-\ln p(\mathbf{D}^B \mid M_k, \mathbf{D}^A)$ (Good, 1952; Volinsky et al., 1997). The lower the predictive log score of model M_k based on data $\mathbf{D}^A$, the smaller the amount of information lost upon eliminating $\mathbf{D}^B$ from the original dataset **D** (i.e., the higher the probability that M_k based on $\mathbf{D}^A$ would reproduce the lost data, $\mathbf{D}^B$). The predictive log score associated with BMA is

29

$$-\ln p(\mathbf{D}^B \mid \mathbf{D}^A)$$

$$= -\ln \sum_{k=1}^{K} p\left(\mathbf{D}^B \middle| M_k, \mathbf{D}^A\right) p\left(M_k \middle| \mathbf{D}^A\right) \qquad (20)$$

Approximating $p(\mathbf{D}^B \mid M_k, \mathbf{D}^A)$ by

$p(\mathbf{D}^B \mid M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}^A)$, and computing $p\left(M_k \middle| \mathbf{D}^A\right)$ via

(7) after replacing $\mathbf{D}$ by $\mathbf{D}^A$, yields a corresponding log score for MLBMA.

Let $\hat{\mathbf{D}}^B$ be kriged estimates of $\log_{10}k$ data $\mathbf{D}^B$ along a borehole obtained using variogram model M_k with

ML parameters $\hat{\boldsymbol{\theta}}_k$ based on $\log_{10}k$ data $\mathbf{D}^A$ in other boreholes. Then the ML log score for drift-free models *Pow0* and *Exp0* is (Ye et al., 2003)

$$-\ln p(\mathbf{D}^B \mid M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}^A)$$

$$= \frac{N_d}{2}\ln(2\boldsymbol{\pi}) + \frac{1}{2}\sum_{i=1}^{N_d}\sigma_i^2 + \frac{1}{2}\sum_{i=1}^{N_d}\frac{(\hat{D}_i^B - D_i^B)^2}{\sigma_i^2} \quad (21)$$

where N_d is the dimension of $\mathbf{D}^B$, D_i^B are its components, and σ_i^2 is given in Ye et al. (2003, Equation B5). In analogy to (17), the ML log score for *Exp1* is

$$-\ln p(\mathbf{D}^B \mid M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}^A) = \frac{N_d}{2}\ln(2\boldsymbol{\pi}) + \frac{1}{2}\ln(|\mathbf{C}_R(\hat{\boldsymbol{\beta}}_k)|)$$

$$+ \frac{1}{2}(\mathbf{D}^B - \mathbf{G}_k\hat{\mathbf{a}}_k)^T \mathbf{C}_R^{-1}(\hat{\boldsymbol{\beta}})(\mathbf{D}^B - \mathbf{G}_k\hat{\mathbf{a}}_k) \qquad (22)$$

Predictive log scores were obtained for each model upon eliminating data from one of six boreholes at a time. Table 4-4 lists the average of these six scores for each model, as well as the average of corresponding MLBMA scores (20). The average predictive log score of MLBMA is seen to be lower than that of any

Table 4-4. Average predictive log score and predictive coverage of individual models and MLBMA

Model	Predictive Log Score	Predictive Coverage (%)
Pow0	34.1	86.5
Exp0	35.2	80.8
Exp1	34.0	83.7
MLBMA	31.4	87.5

individual model, indicating that MLBMA is a better predictor than any of these models.

4.3.2 Predictive Coverage

Another measure of model performance is its predictive coverage (Hoeting et al., 1999). This is the percent of measurements D_i^B that fall within a given prediction interval about $\hat{D}_i^B$. In our case, this interval was generated by conducting Monte Carlo simulations of $\log_{10}k$ conditioned on $\mathbf{D}^A$. We used a simulated annealing code (Deutsch and Journel, 1998, p. 183) to allow generation of statistically nonhomogeneous random fields characterized by a power variogram. Figure 4-14a-c show 90% prediction intervals (dashed) defining the 5% and 95% limits of 500 simulations along borehole X2 using individual models with ML parameter estimates conditioned on measurements in the remaining five boreholes. Figure 4-14d shows averages of these intervals over the three models, weighted by their posterior probabilities. The percent of measurements (triangles) lying within these and similar intervals, associated with all six boreholes, defines predictive coverage as listed in Table 4-4. The predictive coverage of MLBMA is larger than that of any individual model, attesting once again to its superior performance.

Figure 4-15 depicts the cumulative distributions of simulated values at two measurement locations in boreholes V2 and Y3 obtained using individual models and MLBMA, while eliminating data from the corresponding boreholes. The measured values are indicated by vertical lines. In both cases the MLBMA distribution is strongly influenced by that of *Pow0* and weakly affected by *Exp1*. Figure 4-16 shows sample predictive variances obtained using individual models and MLBMA at measurement points along each of the two boreholes. Along V2, *Pow0* with a posterior probability of about 83% exerts an overwhelming influence on the predictive variance of MLBMA, which is however lower (closer to those of *Exp0* and *Exp1*). Along Y3, individual models tend to be associated with a somewhat lower predictive variance than MLBMA.

Overall, MLBMA is a more reliable predictor than any individual model, as indicated by its relatively small predictive log score and large predictive coverage.
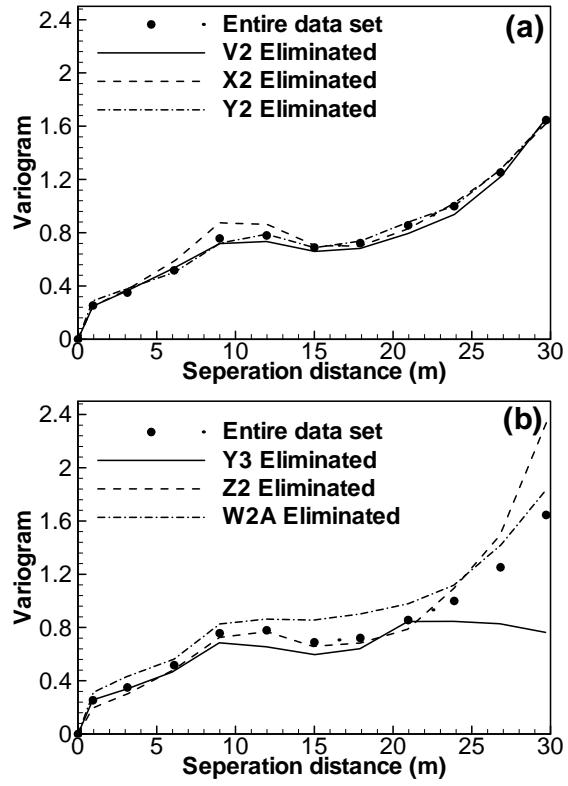
Figure 4-11. Omni-directional sample variograms of all data and all but data from borehole (a) V2, X2, Y2 and (b) Y3, Z2, W2A
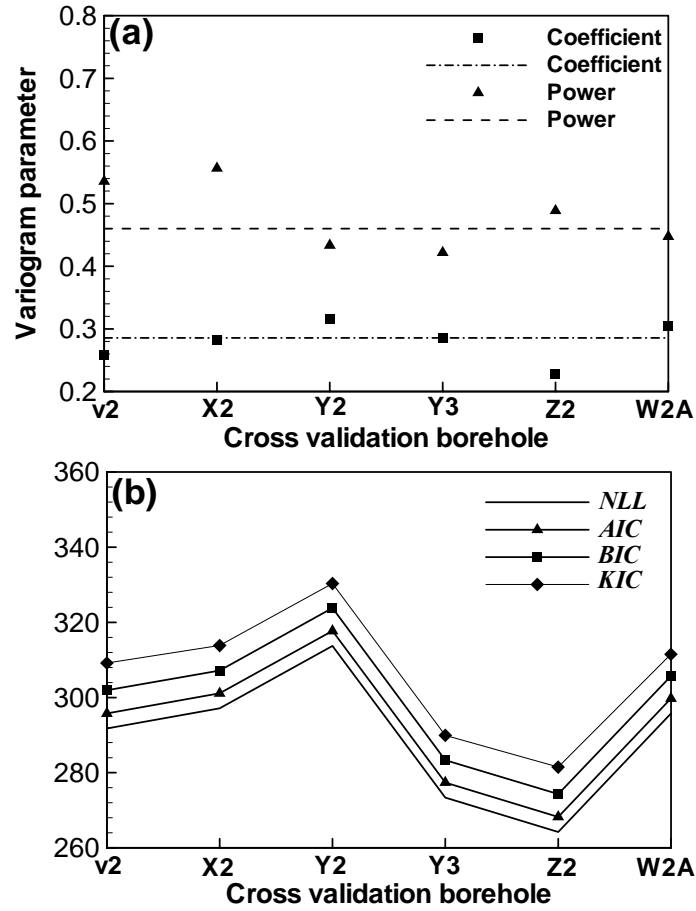
Figure 4-12. Dependence of power variogram (*Pow0*) (a) parameters and (b) quality criteria on data. In (a), symbols designate parameter estimates obtained without data from designated borehole; broken and dashed lines indicate parameters obtained with all data.
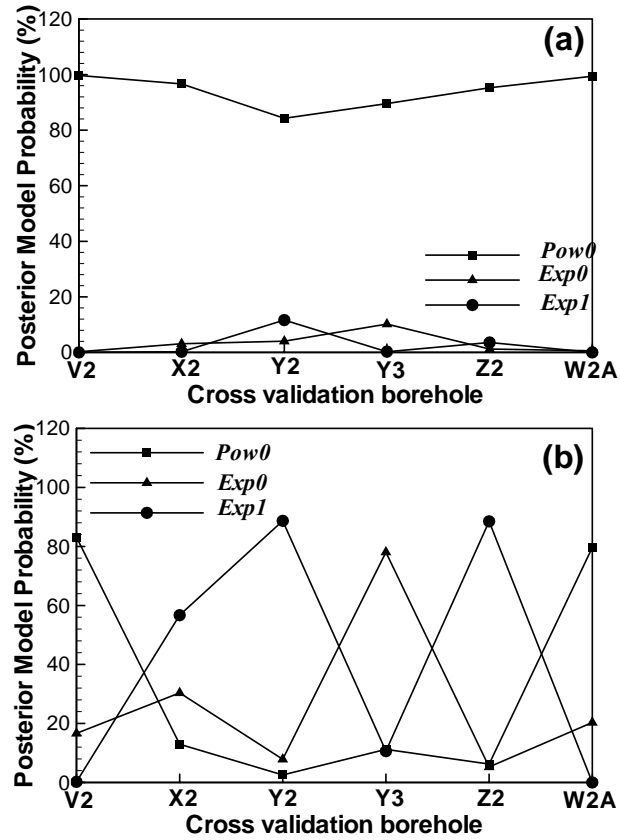
Figure 4-13. Posterior model probabilities based on (a) *BIC* **and (b)** *KIC* **upon eliminating data from designated borehole**
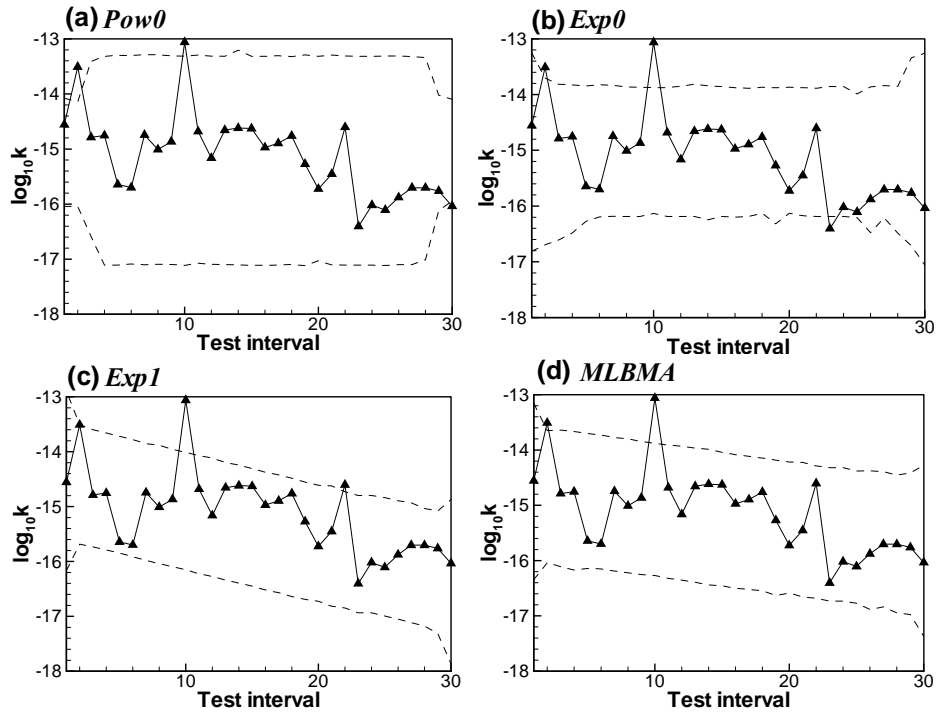
Figure 4-14. 5% (bottom dashed) and 95% (top dashed) limits of simulated prediction interval of $\log_{10}k$ along borehole X2. Triangles designate measured values.
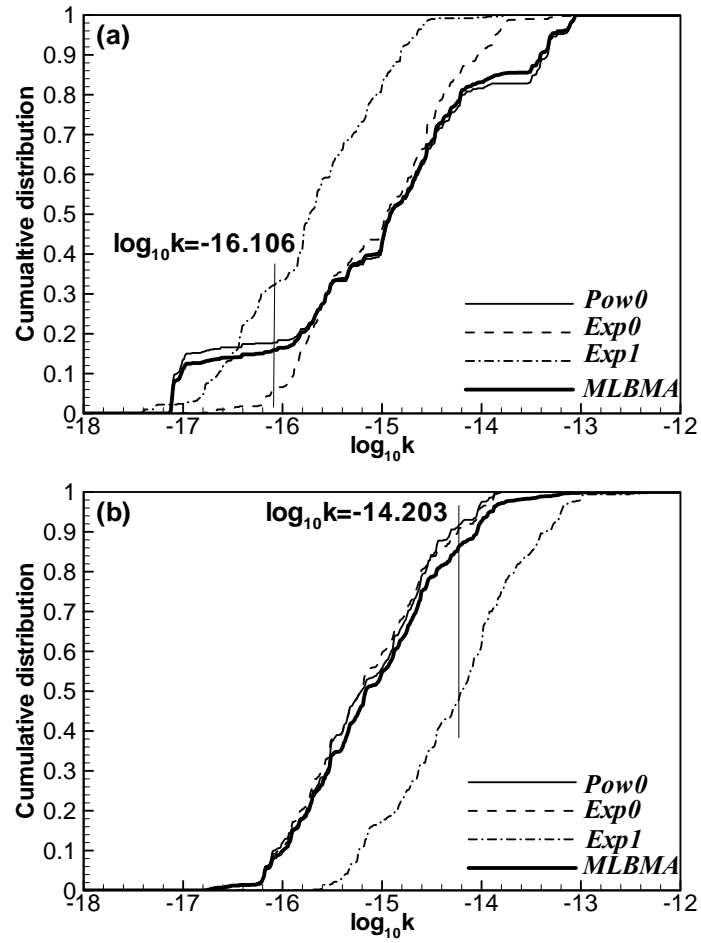
Figure 4-15. Cumulative distribution of simulated $\log_{10}k$ values at a measurement location in borehole (a) V2 and (b) Y3. Vertical line indicates measured value.
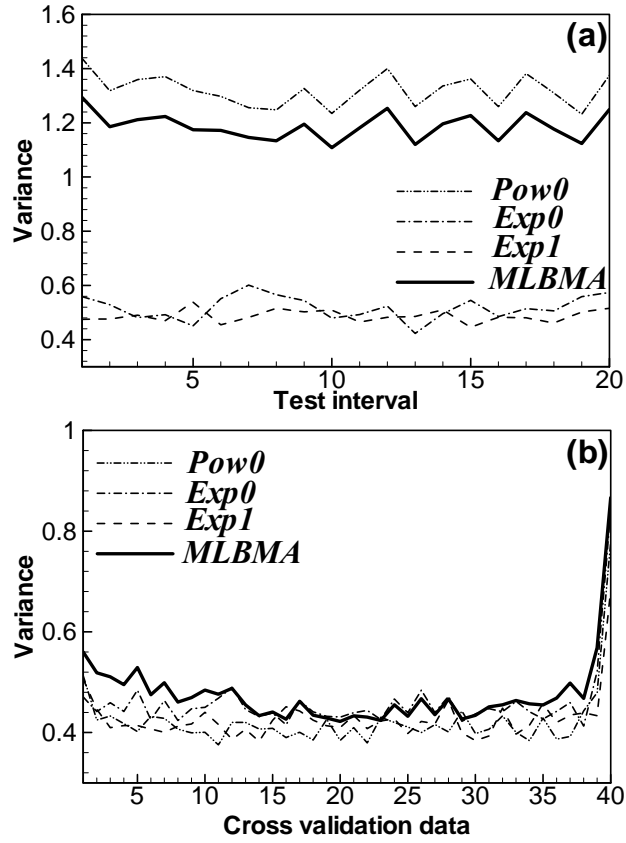
Figure 4-16. Sample variances of $\log_{10}k$ values simulated using various models and MLBMA along borehole (a) V2 and (b) Y3 while eliminating the corresponding data

5   Conclusions

The objective of the research described in this report is the development and application of a methodology for comprehensively assessing the hydrogeologic uncertainties involved in dose assessment modeling. For methodological purposes, uncertainty is classified as being associated with the conceptual-mathematical basis of the model, model parameters, or the scenario to which the model is applied.

This report describes and applies a method to estimate the joint uncertainty in model predictions arising from conceptual model and parameter uncertainties. Analyses of model uncertainty based on a single hydrologic concept are prone to statistical bias (by potential reliance on an invalid model) and underestimation of uncertainty (by under-sampling of the relevant model space). Bias and uncertainty resulting from an inadequate model structure (conceptualization) are often more detrimental to a model's predictive reliability than are suboptimal model parameters.

Bayesian Model Averaging (BMA) provides an optimal but computationally demanding way of combining the predictions of several competing models and assessing their joint predictive uncertainty. The Maximum Likelihood version (MLBMA) of BMA proposed by Neuman (2002, 2003), and described and applied in this report, renders the approach computationally feasible and applicable to real-world hydrologic problems. It applies to deterministic and stochastic models, to complex and simplified models.

Whereas BMA requires specifying a prior distribution for model parameters, MLBMA accepts but does not require such prior information. This is so because, contrary to BMA, MLBMA relies on maximum likelihood model calibration against observational data.

In the most data-limited application, one in which there are no system observations with which to calibrate a model and the only available parameter information is that available from generic databases, model predictions can still be made using prior parameter estimates and model averaging can still be carried out, but only with prior model probabilities. Since the predictions and model probabilities are not conditioned on state variable observations, however, the results are expected to be more uncertain and potentially more biased.

A further benefit of the use of maximum likelihood is that the optimization can yield parameter sensitivity

information. In addition, when the statistical parameters characterizing the parameter and state variable errors are known (i.e., not estimated), maximum likelihood reduces to generalized least squares estimation. In this case, available codes such as PEST and UCODE can be applied.

Prior model probabilities are subjective values reflecting a belief about the relative plausibility of each model based on its apparent consistency with available knowledge and data. Posterior model probabilities are modifications of these subjective values based on an objective evaluation of each model's consistency with available data. Hence, the posterior probabilities are valid only in a comparative, not in an absolute, sense.

MLBMA is based on Kashyap's (1982) information criterion, *KIC*, more commonly used as an optimum decision rule for the ranking of competing models. Like *KIC*, MLBMA favors models which, among a given set of alternatives, are least likely to be incorrect. It honors the principle of parsimony by favoring the least complex among models which, otherwise, fit observational data equally well. Among models of equal complexity, MLBMA favors those exhibiting the best fit. It additionally contains an information term which allows one to consider models of growing complexity as the dataset improves in quantity and quality. Stated otherwise, MLBMA recognizes that when the dataset is limited and/or of poor quality, one should assign relatively low weights to elaborate models with numerous parameters. One should weigh more heavily simpler models with fewer parameters that nevertheless reflect adequately the underlying hydrologic structure and phenomena.

The example application confirms that the non-asymptotic criterion *KIC* is more informative than its asymptotic limit *BIC*, supporting the choice of the former as the basis for MLBMA.

Models considered in MLBMA may have different types and numbers of parameters, but the latter must be estimated and the models weighted based on a single dataset. As an example, to analyze jointly two- and three-dimensional models via MLBMA, a given set of three-dimensional data must be used and either projected onto a two-dimensional plane or averaged in the third dimension for inclusion in the two-dimensional model(s).

Application of MLBMA to alternative geostatistical models of log air permeability variations in unsaturated

fractured tuff has shown it to be a better predictor of spatial variability than any individual model.

To implement MLBMA the following steps are followed.

(1) Postulate alternative conceptual-mathematical models for a site using guidance provided in Neuman and Wierenga (2003).

(2) Assign a prior probability to each model.

(3) Optionally assign prior probabilities to the parameters of each model, using, for example, guidance provided in Meyer and Gee (1999).

(4) Obtain posterior maximum likelihood parameter estimates, and estimation covariance, for each model by inversion (model calibration). In many cases, available codes such as PEST and UCODE can be applied to this step.

(5) Calculate a posterior probability for each model using the model calibration results and the prior model probabilities.

(6) Predict quantities of interest using each model.

(7) Assess prediction uncertainty (distribution, variance) for each model using Monte Carlo or stochastic moment methods.

(8) Weight predictions and uncertainties by the corresponding posterior model probabilities.

(9) Sum the results over all models.

6 References

Abramson, L.R., The philosophical basis for the use of probabilities in safety assessments, *Reliability Engineering and System Safety*, 23:253-257, 1988.

Akaike, H., A new look at statistical model identification, *IEEE Trans. Automat. Contr.*, AC-19, 716-722, 1974.

Akaike H., On entropy maximization principle, in: Krishnaiah P.R. (ed.), *Applications of Statistics*. North Holland, Amsterdam, pp. 27-41, 1977.

Ando, K., Kostner, A., and S.P. Neuman, Stochastic continuum modeling of flow and transport in a crystalline rock mass: Fanay-Augères, France, revisited, in press, *Hydrogeology Journal*, 11(5), 521-535, 2003.

Ayyub, B.M. and R.H. McCuen, *Probability, Statistics, and Reliability for Engineers and Scientists*, Chapman & Hall/CRC Press LLC, Boca Raton, Florida, 2003.

Beven K.J. and J. Freer, Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrology*, 249:11-29, 2001.

Borgonovo, E., G.E. Apostolakis, S. Tarantola and A. Saltelli, Comparison of global sensitivity analysis techniques and importance measures in PSA, *Reliability Engineering and System Safety*, 79:175-185, 2003.

Burnham, K.P. and A.R. Anderson, *Model selection and multiple model inference: a practical information-theoretical approach*, 2nd edition, New York, Springer, 2002.

Carrera J. and S.P. Neuman, Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, *Water Resour. Res.*, 22(2):199-210, 1986a.

Carrera, J. Neuman S.P., Estimation of aquifer parameters under transient and steady state conditions: 3. Application to synthetic and field data. *Water Resour. Res.* 22(2), 228-242, 1986b.

Carrera, J., Medina, A., Axness, C. & Zimmerman, T., Formulations and computational issues of the inversion of random fields. In: *Subsurface Flow and Transport: A Stochastic Approach* (ed. by G. Dagan & S. P.

Neuman), 62-79. Cambridge University Press, Cambridge, United Kingdom, 1997.

Chen, G., W.A. Illman, D.L. Thompson, V.V. Vesselinov, and S.P. Neuman, Geostatistical, type curve and inverse analyses of pneumatic injection tests in unsaturated fractured tuffs at the Apache Leap Research Site near Superior, Arizona, pp. 73-98 in Dynamics of Flow and Transport in Fractured Rocks, edited by B. Faybishenko et al., AGU monograph series, 2000.

Cole, C.R., M.P. Bergeron, C.J. Murray, P.D. Thorne, S.K. Wurstner, and P.M. Rogers, *Uncertainty Analysis Framework – Hanford Site-Wide Groundwater Flow and Transport Model*, PNNL-13641, Pacific Northwest National Laboratory, Richland, Washington, 2001.

Cressie, N., *Statistics of Spatial Data*. John Wiley and Sons, Inc., New York, 1991.

Dagan G. and S.P. Neuman (eds.), *Subsurface Flow and Transport: A Stochastic Approach*, Cambridge University Press, Cambridge, United Kingdom, 1997.

Deutsch, C. V. and A. G. Journel, *GSLIB: Geostatistical Software Library and User's Guide (second edition),* Oxford University Press, New York, 1998.

Doherty, J., *Manual for PEST*, Fifth Edition, Watermark Numerical Computing, Australia, 2002.

Doherty, J., Ground water model calibration using pilot points and regularization, *Ground Water*, 41(2):170-177, 2003.

Draper, D., Assessment and propagation of model uncertainty, *J. Roy. Statist. Soc. Ser. B*, 57(1):45-97, 1995.

Gaganis, P. and L. Smith, A Bayesian approach to the quantification of the effect of model error on the predictions of groundwater models, *Water Resour. Res.* 37(9):2309-2322, 2001.

George, E.I., Comment. *Statist. Sci.* 14(4), 409-412, 1999.

Gomez-Hernandez, J.J., A. Sahuquillo, and J.E. Capilla, Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, I. Theory, *J. Hydrology,* 203:162-174, 1997.

Good, I.J., Rational Decisions, *J. R. Statist. Soc. B,* 57(1), 107-114, 1952.

Guzman, A. G., S. P. Neuman, C. Lohrstorfer, and R. Bassett, Chapter 4, in Validation studies for Assessing Unsaturated Flow and Transport Through Fractured Rock, edited by R.L. Bassett, S.P. Neuman, T.C. Rasmussen, A.G. Guzman, G.R. Davidson, and C.L. Lohrstorfer, pp. 4-1-4-58, NUREG/CR-6203, U.S. Nuclear Regulatory Commission, Washington, D.C., 1994.

Guzman, A. G., A.M. Geddis, M.J. Henrich, C. Lohrstorfer, and S. P. Neuman, Summary of Air Permeability Data From Single-Hole Injection Tests in Unsaturated Fractured Tuffs at the Apache Leap Research Site: Results of Steady-State Test Interpretation, NUREG/CR-6360, U.S. Nuclear Regulatory Commission, Washington, D.C., 1996.

Hannan, E. S., The estimation of the order of ARMA process, *Ann. Stat.*, 1791-1081, 1980.

Helton, J.C., Uncertainty and sensitivity techniques for use in performance assessment for radioactive waste disposal, *Reliability Engineering and System Safety*, 42:327-367, 1993.

Helton, J.C., Treatment of uncertainty in performance assessments for complex systems, *Risk Analysis*, 14:483-511, 1994.

Helton, J.C., Guest editorial: treatment of aleatory and epistemic uncertainty in performance assessments for complex systems, *Reliability Engineering and System Safety*, 54:91-94, 1996.

Hernandez, A.F., S.P. Neuman, A. Guadagnini, and J. Carrera-Ramirez, Conditioning steady state mean stochastic flow equations on head and hydraulic conductivity measurements, 158-162, *Proc. 4th Intern. Conf. on Calibration and Reliability in Groundwater Modelling (ModelCARE 2002)*, edited by K. Kovar and Z. Hrkal, Charles University, Prague, Czech Republic, 2002.

Hernandez, A. F., S. P. Neuman, A. Guadagnini, and J. Carrera, Conditioning mean steady state flow on hydraulic head and conductivity through geostatistical inversion, *Stochastic Environmental Research and Risk Assessment*, 17, DOI10.1007/s00477-003-0154-4, 2003.

Hill, M.C., *Methods and Guidelines for Effective Model Calibration*, U.S. Geological Survey Water-Resources Investigations Report 98-4005, U.S. Geological Survey, Denver, Colorado, 1998.

Hoeksema, R.J. and P. K. Kitanidis, Analysis of the spatial structure of properties of selected aquifers, *Water Resour. Res.*, 21(4), 563-572, 1985.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky, Bayesian model averaging: A tutorial, *Statist. Sci.*, 14(4):382-417, 1999.

Holt, R.M., J.L. Wilson, and R.J. Glass, Spatial bias in field-estimated unsaturated hydraulic properties, *Water Resour. Res.*, 38(12), 1311,doi:10.1029/2002WR001336, 2002.

Isukapalli, S.S., A. Roy and P.G. Georgopoulos, Stochastic response surface methods (SRSMs) for uncertainty propagation: application to environmental and biological systems, *Risk Analysis*, Vol. 18, No. 3, 1998.

James A.L. and C.M. Oldenburg, Linear and Monte Carlo uncertainty analysis for subsurface contaminant transport simulation, *Water Resour. Res.*, 33(11):2495-2508, 1997.

Jefferys, W.H. and J.O. Berger, Ockham's razor and Bayesian analysis, *American Scientist*, 80(1):64-72, 1992.

Kashyap, R.L., Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Trans. Pattern Anal. Mach. Intel. PAMI* 4(2): 99-104, 1982.

Kass, R.E. and A.E.Raftery, Bayes factors, *J. Amer. Statist. Assoc.,* 90(430):773-795, 1995.

Kitanidis P.K. and R. W. Lane, maximum likelihood parameter estimation o hydrologic spatial processes by the Gaussian-Newton method*, J. Hydro.*, 79, 53-71, 1985.

Krishnamurti, T.N., C.M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi. E. Williford, S. Gadgil, and S. Surendran, Multimodel ensemble forecasts for weather and seasonal climate, *J. Climate*, 13(23):4196-4216, 2000.

Kunstmann, H., W. Kinzelbach, and T. Siegfried, Conditional first-order second-moment method and its application to the quantification of uncertainty in groundwater modeling, *Water Resources Research*, 38(4):10.1029/2000WR000022, 2002.

Madigan D. and A.E. Raftery, Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.*, 89(428):1535-1546, 1994.

Martz, H.F. and R.A. Waller, On the meaning of probability, *Reliability Engineering and System Safety*, 23:253-257, 1988.

McKay, M.D., *Evaluating Prediction Uncertainty*, NUREG/CR-6311, U.S. Nuclear Regulatory Commission, Washington, D.C., 1995.

Meyer, P.D., M.L. Rockhold, and G.W. Gee, *Uncertainty Analyses of Infiltration and Subsurface Flow and Transport for SDMP Sites*, NUREG/CR-6565, U.S. Nuclear Regulatory Commission, Washington, D.C., 1997. (http://nrc-hydro-uncert.pnl.gov/)

Meyer, P.D. and G.W. Gee, *Information on Hydrologic Conceptual Models, Parameters, Uncertainty Analysis, and Data Sources for Dose Assessments at Decommissioning Sites*, NUREG/CR-6656, U.S. Nuclear Regulatory Commission, Washington, D.C., 1999. (http://nrc-hydro-uncert.pnl.gov/)

Meyer, P. D. and R. W. Taira, *Hydrologic Uncertainty Assessment for Decommissioning Sites: Hypothetical Test Case Applications*, NUREG/CR-6695, U.S. Nuclear Regulatory Commission, Washington, D.C., 2001. (http://nrc-hydro-uncert.pnl.gov/)

Morgan, M.G., and M. Henrion, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, Cambridge, United Kingdom, 1990.

Mosleh, A., N. Siu, C. Smidts, and C. Lui (eds.), *Model Uncertainty: Its Characterization and Quantification, Proceedings of Workshop I in Advanced Topics in Risk and Reliability Analysis*, NUREG/CP-0138, U.S. Nuclear Regulatory Commission, Washington, D.C., 1994.

Neuman, S.P., Accounting for conceptual model uncertainty via maximum likelihood model averaging, 529-534, *Proc. 4th Intern. Conf. on Calibration and Reliability in Groundwater Modelling (ModelCARE 2002)*, edited by K. Kovar and Z. Hrkal, Charles University, Prague, Czech Republic, 2002.

Neuman, S.P., Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, *Stochastic Environmental Research and Risk Assessment*, 17, DOI10.1007/s00477-003-0151-7, 2003.

Neuman, S. P., and E. A. Jacobson, Analysis of nonintrinsic spatial variability by residual Kriging with application to regional groundwater levels, *Math. Geology*, 16, 491-521, 1984.

Neuman, S.P. and P.J. Wierenga, *A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites*, NUREG/CR-6805, U.S. Nuclear Regulatory Commission, Washington, D.C., 2003.

Poeter, E.P. and M.C. Hill, *Documentation of UCODE, A Computer Code for Universal Inverse Modeling*, U.S. Geological Survey Water-Resources Investigations Report 98-4080, 116 pp., U.S. Geological Survey, Denver, Colorado, 1998.

Press, W.H., S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipe in Fortran 77* (2nd edition), Cambridge University Press, 1992.

Raftery, A. E., Bayesian model selection in structural equation models. In: *Testing Structural Equation Models,* K. Bollen and J. Long (eds.), Sage, Newbury Park, California, pp. 163-180, 1993.

Raftery A.E., D. Madigan, C.T. Volinsky, Accounting for model uncertainty in survival analysis improves predictive performance, in: *Bayesian Statistics,* J. Bernardo, J. Berger, A. Dawid, A. Smith (eds.), Oxford Univ. Press, pp. 323-349, 1996.

Rissanen, J., Modeling by shortest data description, *Automatica*, 14, 465-471, 1978.

Saltelli, A., K. Chan, and E.M. Scott (eds.), *Sensitivity Analysis*, John Wiley & Sons LTD, Chichester, England, 475 pp., 200a.

Saltelli, A., S. Tarantola and F. Campolongo, Sensitivity analysis as an ingredient of modeling, *Statistical Science*, Vol. 15, 4:377-395, 2000b.

Samper, F.J. and S.P. Neuman, Estimation of spatial covariance structures by adjoint state maximum likelihood cross- validation: 1. Theory, *Water Resour. Res.*, 25(3), 351-362, 1989a.

Samper, F.J. and S.P. Neuman, Estimation of spatial covariance structures by adjoint state maximum likelihood cross- validation: 1. Synthetic experiments, *Water Resour. Res.*, 25(3), 363-371, 1989b.

Schwarz, G., Estimating the dimension of a model, *Ann. Stat.*, 6(2), 461-464, 1978.

Taplin, R.H., Robust likelihood calculation for time series, *J. Roy. Statist. Soc. Ser. B*, 55:829-836, 1993.

Tiedeman, C.R., M.C. Hill, F.A. D'Agnese, C.C. Faunt, Methods for using groundwater model predictions to guide hydrogeologic data collection,

with application to the Death Valley regional groundwater flow system, *Water Resour. Res.*, 39(1), 1010, doi:10.1029/2001WR001255, 2003

Volinsky C.T., D. Madigan, A.E. Raftery, R.A. Kronmal, Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. *J. Roy. Statist. Soc. Ser. C* 46, 433-448, 1997.

Vesselinov, V.V., Numerical Inverse Interpretation of pneumatic tests in unsaturated fractured tuffs at the Apache Leap Research Site, Ph. D. Dissertation, the University of Arizona, Tucson, Arizona, 2000.

Wang, W., S.P. Neuman, T. Yao, and P.J. Wierenga, Simulation of large-scale field infiltration experiments using a hierarchy of models based on public, generic, and site data, *Vadose Zone Journal*, 2:297-312, 2003.

Winkler, R.L., Model uncertainty: probabilities for models?, in Mosleh, A., N. Siu, C. Smidts, and C. Lui (eds.), *Model Uncertainty: Its Characterization and Quantification, Proceedings of Workshop I in Advanced Topics in Risk and Reliability Analysis*, NUREG/CP-0138, U.S. Nuclear Regulatory Commission, Washington, D.C., 1993.

Winkler, R.L., Uncertainty in probabilistic risk assessment, *Reliability Engineering and System Safety*, 54:127-132, 1996.

Ye, M., S.P. Neuman, and P.D. Meyer, Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.* (in review), 2004.

Zhang, D., *Stochastic Methods for Flow in Porous Media*, Academic Press, 2001.

Zimmerman, D.A., G. de Marsily, C.A. Gotway, M.G. Marietta, C.L. Axness, R.L. Beauheim, R.L. Bras, J. Carrera, G. Dagan, P.B. Davies, D.P. Gallegos, A. Galli, J. Gomez-Hernandez, P. Grindrod, A. L. Gutjahr, P.K. Kitanidis, A.M. Lavenue, D. McLaughlin, S.P. Neuman, B.S. RamaRao, C. Ravenne, and Y. Rubin, A comparison of seven geostatistically-based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow, *Water Resour. Res.*, 34(6):1373-1413, 1998.

Zio, E. and G.E. Apostolakis, Two methods for the structured assessment of model uncertainty by experts in performance assessments of radioactive waste repositories, *Reliability Engineering and System Safety*, 54:225-241, 1996.

Appendix A. Distribution Coefficients, K_d, and Associated Uncertainty in Dose Assessment Modeling for Decommissioning Analyses

A.1 Introduction

Preliminary or screening dose assessments conducted as part of decommissioning analyses are typically conducted using generic input parameter values. Three examples of codes that are used for this purpose are DandD, RESRAD and MEPAS (Meyer and Gee, 1999). In a recent study, a hypothetical decommissioning test case was used to conduct an uncertainty analysis for two of these codes (DandD v. 1.0 and RESRAD v. 6.0) (Meyer and Taira, 2001). Uranium was used as one of the contaminants of interest. In this case, it was determined that the distribution coefficient was one of the most critical parameters for determining dose.

Because the distribution coefficient is an important source of uncertainty in dose assessment modeling, it is important to have a good understanding of what contributes to uncertainty in the distribution coefficient itself. The distribution coefficient or K_d is an empirical model for the description of partitioning of a contaminant between the soil/sediment and the solution in contact with the soil/sediment and is defined as follows:

$$K_d = C_{ads}/C_{aq} \qquad \text{(A-1)}$$

where C_{ads} is the concentration of the contaminant of interest adsorbed to the solid phase (moles/g) and C_{aq} is the concentration of the contaminant in the aqueous phase (moles/mL). This model assumes that the partitioning of the contaminant between the two phases is in equilibrium and is linear. A significant advantage of the K_d model is its simplicity both for its numerical application in transport codes as well as the relative ease of its experimental measurement. For these reasons, the K_d model is the most widely used adsorption model in hydrologic transport codes for risk assessment calculations. This simplicity and ease of use also make this approach one of the most widely misused models for describing contaminant adsorption. This is particularly true for systems that have highly variable geochemical conditions. Some of the primary factors that can lead to large variation in K_d values include non-linear adsorption, solid/aqueous partitioning conditions that are controlled or influenced by solubility and/or redox conditions, slow reaction kinetics, spatial variability in the solution chemistry or solid phase mineralogy, temporal changes in solution chemistry, and heterogeneities in the physical properties of the aquifer materials.

A.2 Background

A.1.1 Contaminant Adsorption onto Natural Mineral Surfaces

Adsorption, accumulation at the solid-water interface, is one of the primary processes controlling the transport of dissolved contaminants in the vadose zone and groundwater. Adsorption occurs as atoms, molecules, and ions exert forces on each other at this solid-water interface. Adsorption reactions are discussed primarily in terms of intermolecular interactions that occur between the solutes and solid phases (Stumm and Morgan 1996). These interactions include:

1) *Surface complexation reactions* (surface hydrolysis and the formation of coordinative bonds at the surface between metal cations, anions, and surface binding sites).

2) *Electrostatic interactions* at the surfaces, extending over longer distances than chemical forces.

3) *Hydrophobic expulsion* of hydrophobic substances (this includes nonpolar organic solutes), which are usually only sparingly soluble in water and tend to reduce their contact with water and seek relatively nonpolar environments, thus accumulating on solid surfaces and becoming adsorbed on organic sorbents.

4) *Adsorption of surfactants* (molecules that contain both a hydrophobic and a hydrophilic moiety). Interfacial tension and adsorption are intimately related through the Gibbs adsorption law. In simple terms, this law indicates that substances that reduce surface tension will tend to adsorb at interfaces.

5) *Adsorption of polymers* and of polyelectrolytes (humic substances and proteins in particular). This is a rather general phenomenon in natural waters and soil systems that has far-reaching consequences for the interaction of particles with each other

and on the attachments of colloids (and bacteria) to surfaces.

The process in which chemicals become associated with solid phases is often referred to as *sorption*, especially when one is not sure whether one is dealing with *ad*sorption (onto a two-dimensional surface) or *ab*sorption into a three-dimensional matrix.

In addition to the nature of the solid phase, the chemical properties of the solution in contact with the solid phase will have a substantial effect on its adsorption characteristics. For example, pH will have a major influence on the degree of surface hydrolysis, which in turn affects the nature and extent of surface charge. Ionic strength will affect the electrostatic nature of the surface and therefore the electrostatic interactions that can occur. In addition to these effects, the adsorption process itself will change the nature of the surfaces of the solid phase and will influence further adsorption.

The chemical properties of the solution in contact with the solid phase will also affect adsorption as a result of interactions between dissolved species. For example, many metal ions form complexes with major anions in solution. The formation of these complex species can have a major influence on the charge and geometry of the original ion and as a result, significantly alter the sorptive properties of the species of interest. A special case of complex formation is hydrolysis. Hydrolysis is the formation of complexes with hydroxide ion and is a strong function of pH. Ionic strength can be an important factor that affects the activity of all dissolved ions, and as a result, the extent of complex formation. Eh can also have a large influence on adsorption by altering the oxidation state of the contaminant and/or the adsorbent.

A.1.2 Empirical Approaches to Adsorption Modeling

As indicated previously, the linear equilibrium adsorption isotherm or K_d model is an empirical approach that assumes the adsorption of a solute increases linearly with increasing concentration of a solute. As a result of the empirical nature of the K_d model, it cannot represent the individual contributions of different uptake mechanisms. In addition, the K_d model cannot recognize a maximum sorption limit. In actuality, there are a finite number of sorption sites and, as a result, sorption will reach a practical upper limit.

Despite the shortcomings of the K_d model, it can provide an accurate description of adsorption under certain conditions. The K_d model generally works well for trace concentrations of un-ionized hydrophobic organic compounds; however, application to ionic inorganic contaminants is more limited. Appropriate use of the K_d approach for modeling adsorption of ionic species is generally limited to species that have very simple chemistry and site conditions where the groundwater solution chemistry and mineralogy of the aquifer material are quite constant and homogeneous. This is generally an unusual occurrence, particularly at contaminated waste sites.

In addition to the linear equilibrium adsorption isotherm, several other more complex empirical adsorption models are available. The Freundlich isotherm (Freundlich, 1926) is a nonlinear equilibrium adsorption model defined by the relationship:

$$C_{ads} = K_{Fr}(C_{aq})^n \qquad (A-2)$$

where C_{ads}, and C_{aq} are defined as in Eq. (A-1) and K_{Fr} and n are empirical coefficients. For the special case where n = 1, Eqs. A-1 and A-2 are identical. A plot of log C_{ads} versus log C_{aq} should result in a straight line with a slope of n and an intercept of log K_{Fr}. As with the linear adsorption isotherm model, an adsorption maximum cannot be represented with the Freundlich isotherm.

An empirical adsorption model that accounts for an upper limit to adsorption is the Langmuir isotherm (Langmuir, 1918). This model was developed for adsorption of gases onto solid surfaces and assumes that all sorption sites are energetically equal. The general form of the Langmuir isotherm (as adapted for adsorption from solution) is:

$$C_{ads} = K_{La}bC_{aq}/(1 + K_{La}C_{aq}) \qquad (A-3)$$

Where b is the maximum adsorption capacity of the substrate (g solute/g adsorbent), and K_{La} is a constant that represents the strength of adsorption of the solute onto the solid (mL/moles). Values for b can be determined for a given data set by plotting C_{aq}/C_{ads} versus C_{ads}. This should yield a straight line with a slope of 1/b and an intercept of $1/K_{La}b$.

A.1.3 Surface Complexation Approach to Adsorption Modeling

Surface complexation models (SCMs) are chemical models that provide a molecular level mechanistic description of adsorption. Analogous to solution complexation, surface complexation models define surface species, chemical reactions, equilibrium constants, mass balances and charge balances that are

based on an equilibrium thermodynamic approach. Surface complexation models constitute a family of models that have many common characteristics and adjustable parameters. The models differ in the structural representation of the solid-solution interface (location of the adsorbing ions and resulting charge). The primary advantage of surface complexation models over empirical models is the ability to account for variable physical-chemical conditions. This is in stark contrast to empirical models, which generally ignore the chemical complexity of the sorption processes and aqueous complexation.

Although surface complexation models are often incorporated directly into complex reactive transport codes, the advantages of the surface complexation models can be exploited using simpler hydrologic dose assessment codes as well. This has important implications because it is these simpler codes that are most frequently used for regulatory decision-making purposes. In most hydrologic dose assessment codes the complex geologic conceptual model is simplified to a relatively simple geologic conceptual representation (Meyer and Gee, 1999). These simplified conceptual models are typically composed of layers or zones of materials that have distinct and homogenous physical (hydrologic), mineralogical, and chemical properties. By making certain assumptions regarding the average or typical chemical and mineralogical characteristics within these different layers or zones, surface complexation models can be used to calculate individual K_d values appropriate for each layer or zone within the conceptual model.

As indicated above, surface complexation models constitute a family of models that have many common characteristics and adjustable parameters. The most frequently used surface complexation models include the Diffuse Layer Model (DLM), the Constant Capacitance Model (CCM), the Triple Layer Model (TLM), and non-electrostatic SCMs. The three surface complexation models (DLM, CCM, and TLM) will be discussed briefly below and the non-electrostatic SCMs will be discussed in the next section.

The DLM is the simplest of the electrostatic SCMs. In the DLM, protonation/deprotonation and adsorption occur in one plane at the surface/solution interface and only those ions specifically adsorbed in this inner "o-plane" contribute to the total surface charge ($\sigma_s = \sigma_o$). Dzombak and Morel (1990) have provided a detailed evaluation of the DLM, including the development of a strong site/weak site conceptual model for the mineral surface. The analysis of Dzombak and Morel (1990) also provides parameters for its application to the

sorption of a number of cationic and anionic species on ferrihydrate.

The CCM model (Schindler et al., 1976) is conceptually similar to the DLM. In contrast to the DLM, the CCM assumes that the charged surface is isolated from the bulk solution by a plane with a constant capacitance C_1 (Farads/m^2), resulting in a linear potential gradient from the charged substrate to the bulk solution. The CCM approached is generally limited to a specific ionic strength because changes in ionic strength require recalculation of C_1. The constant capacitance term is not measureable and as a result is typically applied as an empirical parameter and fit to the data. This has the advantage of providing a better fit to the experimental data, but at the expense of theoretical rigor.

The TLM (Davis et al., 1978; Davis and Leckie, 1978; 1980) is conceptually similar to both the DLM and the CCM. In the TLM; however, the charge/potential relationships of the mineral-water interface are divided into three layers. The TLM approach provides more flexibility to simulate ionic strength effects by representing sorption of background electrolytes and permitting the formation of both inner- and outer-sphere complexes. As a result of its construction, the TLM requires additional parameters beyond those needed for the DLM and CCM. Additional parameters include equilibrium constants K_{Cat} and K_{An} for background electrolyte sorption, and capacitances C_1 and C_2 associated with the areas between the o- and β-planes and β- and d-planes, respectively.

A.1.4 Non-Electrostatic Surface Complexation Models

Although SCM is the most theoretically rigorous approach to modeling contaminant adsorption onto mineral surfaces, application to natural materials remains problematic. SCM adsorption data are generally determined using well-characterized single-phase minerals whose surface properties, such as surface area, site density, and electrostatic correction terms, are readily measured. For most natural soils and sediments, measurement of the site density and electrostatic correction terms of the individual contributing minerals is impractical if not impossible. Natural mineral surfaces in sediments/soils are typically coated with poorly crystalline secondary mineral coatings (Penn et al. 2001, Coston et al., 1995). In general, these coatings make it extremely difficult to quantitatively assess the electrostatic contribution to the free energy of adsorption.

Davis et al. (1998, 2002) recently demonstrated two approaches for modeling adsorption onto natural heterogeneous materials. The two approaches are the Component Additivity approach and the Generalized Composite approach. The Component Additivity approach is based on summing the adsorption of the individual mineral components of the soil or sediment to get the total adsorption of the mixture. Because this modeling approach is based on summing the results from models already calibrated with pure mineral phases, the Component Additivity approach is predictive and does not involve fitting the adsorption data of the natural materials.

In the Generalized Composite modeling approach, the surface of the mineral assemblage is considered too complex to be quantified in terms of the contributions of individual phases to adsorption. Instead the electrostatic terms are omitted and the mass action expressions are described in terms of "generic" surface function groups. The stoichiometry and formation constants for each reaction are evaluated based on their simplicity and goodness of fit to the experimental adsorption data (Davis et al., 2002; Davis et al., 1998). The generic surface sites represent average properties of the sediment/soil rather than specific minerals. Experimental data for site-specific natural materials must be collected over the range of chemical conditions that can be expected in the field. Because of the semi-empirical nature of this approach, the resulting model parameters are not likely to be transferable to other field sites.

These two modeling approaches were compared for U(VI) adsorption by sediments from the Koongarra natural analog site in northwest Australia (Davis et al., 2002, Waite et al., 2000). The Component Additivity approach required eight reactions and used a diffuse double layer electrostatic model. The Generalized Composite approach only needed four surface reactions and did not include an electrostatic model. The model fit to the experimental adsorption data for both approaches was nearly the same, even though the Generalized Composite model had seven model parameters and the Component Additivity model had eleven.

A.2 Sources of K_d Value Uncertainty

The uncertainty associated with any particular K_d value used in a risk assessment can be placed into three major categories:

1) Experimental uncertainty
2) Sorption process chemistry uncertainty

a) variation in solution chemistry
  - complexation
  - competitive adsorption
  - alteration of the adsorption-site chemistry
b) variation in surface adsorption sites
  - mineralogy
  - surface coatings and fracture fillings
3) Uncertainty resulting from scaling of K_d measurements determined in the laboratory to intact sediments/soil in the field
c) effective surface area
  - surface sites in hydrologic contact with moving radionuclides
  - diffusion

The experimental uncertainty is the sum of the errors resulting from measurement errors that occur during the K_d value measurement. This is generally the most easily quantifiable component of the uncertainty and can be determined using statistical methods. Both the uncertainty in the K_d value that results from variation in the sorption process chemistry, and the uncertainty resulting from the scaling of laboratory K_d values to intact sediments/soil in the field, could be considered to be conceptual model uncertainties. This is because, for a particular K_d value, the solution chemistry, sediment/soil mineralogy and surface area per unit weight of the laboratory sample used for the K_d value determination is assumed to be identical to that of the site (or portion of the site) that is being modeled with the reactive transport code. If any of these parameters vary significantly such that they can result in a significant change in the K_d value, then the conceptual model would have to be considered as unrepresentative.

In order to quantify the uncertainty of a reactive transport model resulting from uncertainty in the K_d value, the uncertainties resulting from the sorption process chemistry and the uncertainty from scaling must be quantified.

Quantification of the sorption process chemistry uncertainty can be broken down into two major parts. The first part is quantification of the variation in the solution chemistry and sediment/soil mineralogy within the site being modeled. This is a site characterization task that must be conducted with expert guidance to ensure that measurements of all geochemical parameters that could potentially influence adsorption of the contaminant of interest are made. In addition to the geochemical parameter measurements, spatial frequency of the sample collection is of critical importance for quantification of the geochemical parameter variation.

The second component of the sorption process chemistry uncertainty required to quantify K_d value uncertainty is quantification of the variation in the K_d value as a function of the important geochemical parameters. This must be conducted in the laboratory over the range of values for each important geochemical parameter that occurs within the site of interest.

The uncertainties that result from scaling issues are largely the result of differences in the amount of adsorption sites that are in hydrologic contact with the mobile aqueous phase within the field site, versus that which are accessible to the aqueous phase in the laboratory K_d value determinations. Because the adsorbed phase concentration (C_{ads} in Eq. A-1) of the K_d is given in terms of unit mass, as opposed to unit surface area, any difference between the surface area per unit weight of soil/sediment that occurs in-situ versus that in the laboratory system will result in error.

A.3 Variability in K_d Values and the Impact on Transport Calculations

As indicated earlier K_d values are empirical constants and as a result can be applied with confidence only to conditions that are the same as those under which the value was measured. If the sediment/soil mineralogy or physical properties, solution chemistry, or contaminant loading of the system to be modeled are significantly different than that for which the K_d value was determined, significant error in the estimated transport rates could result. This is because many factors can affect the degree to which a particular contaminant adsorbs to a particular sediment or soil (as discussed above). These factors include: sediment mineralogy and surface area, major ion concentration in solution (complexation and competitive adsorption), pH of the solution, and the concentration of the adsorbent in solution and on the adsorbate. Careful application of expert geochemical knowledge can often significantly reduce the number of significant variables that must be considered for evaluating K_d values. For example, some radionuclides may have a low tendency to form complexes with other major ions in solution or do not interact significantly with certain mineral surfaces.

In the hypothetical test case conducted by Meyer and Taira (2001), a K_d value of 15 was used for uranium. This value is a geometric mean value for loam taken from the compilation by Sheppard and Thibault (1990). A major problem with using mean K_d values from this and similar literature compilations of K_d values for conducting screening calculations is the inherently

large variation in the K_d values. For example, Sheppard and Thibault (1990) report a range in K_d values for uranium of 0.03 to 2200 ml/gm. The reason for this large degree of variability in K_d values is due largely to differences in solution chemistry and soil properties used in the various K_d value determinations included in the compilation. Because no control is placed on these variables during the statistical analysis of the K_d values, the individual impact of these variables is ignored, resulting in the large overall variation observed.

To better illustrate the impact of these values on the calculated mobility of uranium, these K_d values will be converted to retardation factors. The retardation factor is a measure of the ratio of the average linear velocity of water divided by the average linear velocity of the contaminant. The retardation factor can be calculated using the following equation:

$$R_f = 1 + (K_d \, \rho_b)/\theta \qquad (A-4)$$

where, the retardation factor is R_f (unitless), ρ_b (kg/m^3) is the bulk density, and θ (m^3/m^3) is the volumetric water content. By assuming a bulk density of 1.86 kg/m^3 and a volumetric water content of 0.30 m^3/m^3, equation 1 can be simplified to:

$$R_f = 1 + 6.2K_d \qquad (A-5)$$

Using the range of K_d values for uranium reported by Sheppard and Thibault (1990), the range in retardation factors is calculated to be 1.2 to 14,000. This range in retardation factors illustrates that, for the reported range of K_d values, uranium has the potential to vary from being essentially unretarded ($R_f = 1$ indicates the contaminant moves with the water or no adsorption occurs) to being essentially immobile (strongly adsorbed), depending upon the conditions encountered.

There are several factors that account for this large variation in adsorption potential. These factors include the highly variable adsorption potential of different minerals for uranium, and the strong influence of pH and carbonate concentration of uranium adsorption. For example, Turner et al. (2002) illustrate uranium K_d data for silica, montmorillonite, and clinoptilolite as a function of pH (in equilibrium with atmospheric CO_2). From this data, it can be seen that for silica at pH 8 the typical K_d value is 5 ml/gm. As the pH decreases to between 6.5 and 6.0, the K_d for silica peaks at 50. As the pH decreases further to pH 4 the K_d decreases to about 0.3. In contrast, K_d values for montmorillonite are much higher. At pH 8 the K_d is approximately 300. As the pH decreases to between 6.5 and 6.0 the K_d for montmorillonite peaks at 10,000. As the pH decreases

further to pH 4 the K_d decreases to about 300. These relationships are illustrated in Figure A-1.

It is clear from these illustrations that the variability in K_d values as a result of large heterogeneities in site-specific mineralogy and solution chemistry could result in highly variable adsorption behavior that could potentially result in significant error when compared to modeling results determined with a single generic K_d value.

The magnitude of variation illustrated for uranium K_d values could also be expected for other radionuclides commonly encountered at NRC decommissioning sites. Specific examples are C-14 and possibly Tc-99 and Sr-90. Most of the other radionuclides commonly encountered at NRC decommissioning sites (Cs-137, Co-60, Ni-63, Am-241, Pu-238,-239,-241, Eu-152, Nb-94, and Cm-243) are strongly adsorbing under typical conditions and even large variability in their K_d values is not likely to result in large differences in dose uncertainty. H-3 is not adsorbing with a K_d value of zero with little uncertainty. This suggests that the greatest degree of uncertainty in dose models results from uncertainty of K_d values for a limited number of radionuclides.

A.4 Determination of K_d Values and Associated Uncertainly

Experimental determination of site-specific K_d values is likely to remain the most common method for characterizing adsorption in risk assessment models at most sites in the near term. Geochemical reasoning and thermodynamic modeling can provide valuable guidance and support for the experimental determination of K_d values and how they vary with solution chemistry and mineralogy. In some cases, surface complexation models can be used to estimate K_d values as a function of solution chemistry and mineralogy. This approach has been demonstrated by a number of researchers to support performance assessments at major radioactive waste disposal sites that have significant resources to devote to such efforts (Davis et al., 2002; Turner et al., 2002). This approach is currently gaining acceptance as the best compromise between comprehensive scientific defensibility and practical application. It is expected that this approach for determining input sorption parameters for more routine risk and performance assessment modeling efforts will become increasingly utilized as the database of thermodynamic sorption models increases. This approach typically requires a significant amount of site-specific geochemical characterization.

A.4.1 Systematic Approach for Determination of K_d Values and Associated Uncertainty

A systematic approach for determining K_d values and associated uncertainty for use in dose assessment modeling at specific sites is outlined below in general terms. The first step in this approach is to collect all site-specific characterization data that is available that
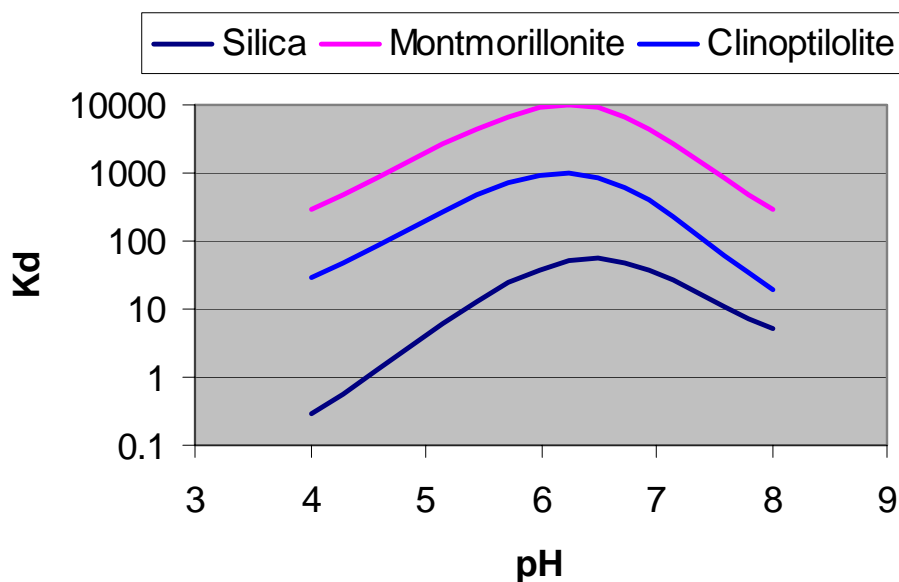

Figure A-1. Variability in uranium K_d as a function of mineral and pH. Based on data from Turner et al. (2002)

may be useful for estimating adsorption of the contaminants of interest. This could include aqueous phase chemical data (contaminant concentrations, major ion data, Eh, and pH), aquifer material mineralogy, mineral surface coatings, stratigraphy, and spatial and temporal variability of these geochemical parameters. This information can be used to guide the selection of K_d values from generic compilations of K_d values or from other adsorption data available in the literature that could be used to calculate K_d values (such as surface complexation model data). If the uncertainty of the K_d value estimates determined in this process is acceptable, no further K_d value refinement is necessary. If the uncertainty of the K_d value estimates determined in this process is too high or if the available characterization data and/or available adsorption data for the contaminants of concern is not adequate, then a more detailed geochemical analysis must be conducted. As part of the geochemical analysis, site-specific characterization needs would be determined and the requirements and scope of an adsorption study to develop site-specific K_d values as a function of important geochemical parameters would be outlined. The site-specific characterization work may involve an iterative process where early characterization results can be used to determine and guide further characterization needs.

A.4.2 Determination of Uranium K_d Values and Associated Uncertainty with Iterative Refinement to Maximize Cost Effectiveness

A brief outline will be provided here to illustrate how this methodology can be applied to a specific contaminant. In this case, uranium has been selected for illustrative purposes because it is a major contaminant of concern for a number of decommissioning sites and uranium has complex adsorptive behavior that ranges from non-adsorbing to highly adsorbing, depending on geochemical conditions.

The first step to estimating a site-specific K_d value is to compile any available site characterization data that would be useful from a geochemical perspective. This would include solution chemistry data (major cation and anion concentrations, alkalinity measurements, pH, Eh and contaminant concentrations), and mineralogy (texture, major mineral components, clay mineralogy and hydrous metal oxide content). The geochemistry of the contaminant of interest will determine which geochemical parameters are most critical for determination of the K_d value. In the case of uranium, the carbonate concentration has a very large effect on

the adsorption of uranium due to strong complex formation with carbonate. For example, K_d values for uranium (VI) adsorption on ferrihydrite at pH 8 have been shown to decrease by four orders of magnitude as the partial pressure of carbon dioxide gas, pCO_2, increases from its value in air (0.032%) to 1% (Davis et al., 2002). This is an important variation to understand, because pCO_2 in aquifers commonly reaches values of 1-5%, while most laboratory determined K_d values have been determined in equilibrium with air. The carbonate concentration (or pCO_2) can be determined from measurements of pH and alkalinity. So in general, the two most important solution parameters to know for estimating K_d values are pH and alkalinity. Other major ions are of secondary importance, but can influence the speciation of the carbonate system.

After the solution parameters, pH and alkalinity, the next most important geochemical parameter to know for uranium K_d estimation is the mineralogy. The mineralogical information can range from very general descriptions (sand, silt, clay, calcarious, etc.), to very specific such as a complete quantitative mineralogical characterization. This would include the percentages of the major minerals present, clay mineralogy and hydrous metal oxide content. In between these two extremes, one could obtain a semi-quantitative XRD scan that would provide characterization of the major crystalline minerals present.

Once the characterization data have been assembled, this information would be used to find K_d values in the literature or from K_d compilations that best match site conditions. Alternatively, adsorption data determined for pure minerals could be used to calculate K_d values. This could involve the use of surface complexation models and geochemical equilibrium codes combined with adsorption site densities estimated from site characterization data to estimate K_d values for specific geochemical conditions.

Depending on the nature of the site and the adsorption data available in the literature, it may be determined that some limited additional characterization data may significantly reduce the uncertainty of the current K_d estimates. For example, if uranium adsorption data are available in the literature for ferrihydrite and montmorillonite at various pH values and carbonate (pCO_2) concentrations and it is determined that these two minerals are significant components of the aquifer material and are likely to be controlling uranium adsorption, it may then be worthwhile to conduct quantitative measurements of these components on available samples from the site. An additional step that could be taken to narrow K_d value uncertainty even further, would be to conduct an adsorption study using

site aquifer material over a range of parameters appropriate to site conditions. This iterative approach to narrowing the uncertainty of the K_d value may be a sensible approach for addressing dose assessment modeling at sites that initially have little characterization data available. It will also provide a means to balance the contrasting needs of reducing K_d value uncertainty and producing a cost effective performance assessment.

The methodology outlined above for determining a K_d value for uranium can also be used to determine the spatial and temporal variation in the K_d value; however, the spatial and temporal variation in the indicated critical parameters must be known or estimated. In the case of pCO_2, values can increase in groundwater recharge as a result of transport through organic rich horizons where significant decomposition is occurring. This can lead to significant spatial and temporal variation in pCO_2 and therefore uranium retardation.

Significant complications that have not been addressed in this discussion are the fact that K_d values are generally given in units based on adsorption per unit mass. Because adsorption is actually related to the site density of the adsorbent, significant differences in surface area per unit mass of the material used in the adsorption measurements the site material can result in error. For example, Turner et al. (2002) have shown that uranium adsorption onto montmorillonite, clinoptilolite, α-alumina, and quartz have similar K_d values on a specific surface area basis (mL/m^2); however, for K_d values on a mass basis (mL/g), the difference between K_d values for montmorillonite and quartz is about three orders of magnitude at near neutral pH values.

A.5 References

Coston, J.A., C.C. Fuller, and J.A. Davis. Pb^{2+} and Zn^{2+} Adsorption by a Natural Aluminum-Bearing and Iron-Bearing Surface Coating on an Aquifer Sand. *Geochim. Cosmochim. Acta,* 59(17): 3535-3547, 1995.

Davis, J. A., G.P. Curtis, and J.D. Randall, *Application of Surface Complexation Modeling to Describe Uranium(VI) Adsorption and Retardation at the Uranium Mill Tailings Site at Naturata, Colorado,* NUREG/CR-6820, U.S. Nuclear Regulatory Commission, Washington, DC, 2003.

Davis, J.A., T.E. Payne, and T. D. Waite. Simulation the pH and pCO2 Dependence of Uranium(VI) Adsorption by a Weathered Schist with Surface Complexation Models. In: *Geochemistry of Soil Radionuclides,* (ed. P.C. Zhang and P.V. Brady), SSSA Special Publication Number 59, pp. 61-86, Soil Science Society of America, Madison, Wisconsin, 2002.

Davis, J. A., J.A. Coston, D.B. Kent, and C.C. Fuller. Application of the Surface Complexation Concept to Complex Mineral Assemblages. *Environ. Sci. Technol.* 32 :2820-2828, 1998.

Davis, J.A. and J.O. Leckie. Surface Ionization and Complexation at the Oxide/Water Interface 3 - Adsorption of Anions." *Journal of Colloid Interface Science.* 74:32-43, 1980.

Davis, J.A. and J.O. Leckie. Surface Ionization and Complexation at the Oxide/Water Interface 2 - Surface Properties of Amorphous Iron Oxyhydroxide and Adsorption of Metal Ions. *Journal of Colloid Interface Science.* 67:90-107, 1978.

Davis, J.A., R.O. James, and J.O. Leckie. Surface Ionization and Complexation at the Oxide/Water Interface 1 - Computation of Electrical Double Layer Properties in Simple Electrolytes." *Journal of Colloid Interface Science.* 63:480-499, 1978.

Dzombak, D.A., and F.M.M. Morel, *Surface Complexation Modeling, Hydrous Ferric Oxide,* John, Wiley and Sons, New York, 1990.

Freundlich, H. *Colloid and Capillary Chemistry.* Methuen, London, 1926.

Langmuir, D. The adsorption of gases on plane surfaces of glass, mica, and platinum. *Jour. Amer. Chem. Soc.* 40: 1361-1403, 1918.

Meyer, P.D. and G.W. Gee. *Information on Hydrologic Conceptual Models, Parameters, Uncertainty Analysis, and Data Sources for Dose Assessments at Decommissioning Sites.* NUREG/CR-6656, PNNL-13091, U.S. Nuclear Regulatory Commission, Washington, DC, 1999.

Meyer, P.D. and R.Y. Taira. *Hydrologic Uncertainty Assessment for Decommissioning Sites: Hypothetical Test Case Applications.* NUREG/CR-6695, PNNL-13375, U.S. Nuclear Regulatory Commission, Washington, DC, 2001.

Penn, R.L., C. Shu, H. Z. Xu, and D.R. Veblen. Iron Oxide Coatings on Sand Grains from the Atlantic Coastal Plain: High-Resolution Transmission Electron Microscopy Characterization. *Geology*, 29(9):843-846, 2001.

Schindler, P.W., B. Furst, R. Dick, and P.U. Wolf. "Ligand Properties of Surface Silanol Groups." *Journal of Colloid Interface Science.* 55:469-475, 1976

Sheppard, M.L. and D.H. Thibault. "Default Soil Solid/Liquid Partition Coefficients, Kds, for Four Major Soil Types: A Compendium." *Health Physics* 59(4):471-478, 1990.

Stumm, W., and J.J. Morgan. *Aquatic Chemistry.* Wiley, New York, 1996.

Turner, D.R., F.P. Bertetti, and R.T. Pabalan. Role of Radionuclide Sorption in High-Level Waste Performance Assessment: Approaches for the Abstraction of Detailed Models. In *Geochemistry of Soil Radionuclides,* (ed. P.C. Zhang and P.V. Brady), SSSA Special Publication Number 59, pp. 211-252, Soil Science Society of America, Madison, Wisconsin, 2002.

Waite, T.D., J.A. Davis, B.R. Fenton, and T.E. Payne. Approaches to Modeling Uranium(VI) Adsorption on Natural Mineral Assemblages. *Radiochim. Acta* 88:687-693, 2000.