Local Analysis of Prameter Covariances Resulting from the Calibration of an Overparameterized Water Quality Model

U. Callies and M. Scharfe

GKSS Research Center, Max-Planck-Str. 1, 21502 Geesthacht, Germany
E-mail: ulrich.callies@gkss.de

Abstract: Mechanistic water quality simulation models are important tools for supporting environmental management decisions. Possibly the most severe problem with the usage of mechanistic models is that in most cases they cannot be fully identified from data due to model overparameterization. The calibration of overparameterized models results in covariances among model parameters, the neglect of which may lead to a significant overestimation of model output uncertainty. We discuss principal component analysis (PCA) of the posterior parameter error covariance matrix as a tool for the identification and proper representation of parameter covariances. Our study deals with a water quality model specifically designed to support the interpretation of algae biomass observations at one single station (Weir Geesthacht) on the Elbe river in Germany.

1. INTRODUCTION

The motivation for our modelling activity has been to test a hypothesis according to which observed negative correlations between temperature and chlorophyll *a* concentrations in summer at station Geesthacht on the Elbe River might indicate algae growth being limited by lack of silica. For this specific purpose a relatively simple model has been designed and fitted to observed chlorophyll *a* concentrations. However, even this simple model turns out not to be identifiable from the data used. One obvious reason why some model parameters or parameter combinations are poorly determined by the data, is that the observations have all been collected at the same location and do therefore not resolve profiles along the river. A possibly more important explanation, however, is that observations of one single state variable cannot disentangle details of the mechanistic processes. Different parameterizations linking external forcing (radiation, discharge, temperature) to model output (algal biomass concentrations) can be similarly effective in reproducing the data.

Being not parsimonious in the light of the existing data (i.e. being overparameterized) is a property of most detailed mechanistic models. Our study is intended to illustrate an approach for coping with this situation by explicitly accounting for parameter interaction structures that result from model overparameterization. Specific combinations of parameters may be much less uncertain than the individual parameters they are made up by. The opposite is also true: Some parameters may be collectively more uncertain than any of the individual parameters. PCA of the posterior parameter error covariance matrix allows to discriminate combinations of model parameters that are effectively controlled by the data from those parameter combinations that are irrelevant for model counterparts of observations. Often results from PCA can be interpreted in terms of the basic mechanisms represented in the model. We illustrate the method for the example from water quality modelling.
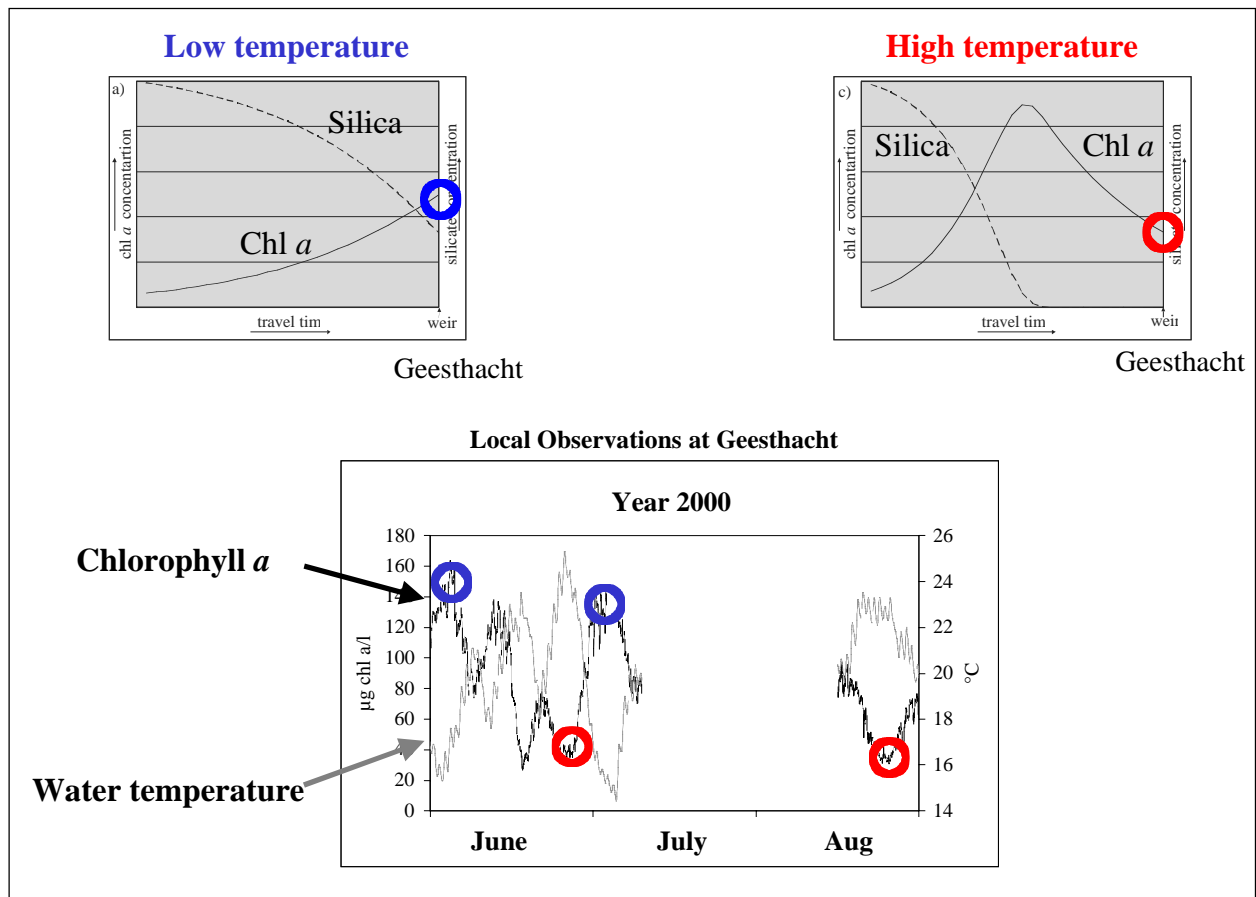
Figure 1: Suggested explanation for negative correlations between water temperature and chlorophyll *a* concentrations observed during certain summer periods at station Geesthacht. Upper panels: Modelled temporal evolution of chlorophyll *a* and silica concentrations in individual water packages assuming low and high temperature, respectively. Local time series can be produced by storing the final points of all trajectories that arrive at Geesthacht. Bottom panel: Observations at Weir Geesthacht. Temperature observations have been shifted by an estimated response time of two days.

2. METHODOLOGY

2.1. The Model

Our model involves the representation of two different algae species, green algae and diatoms, both of which are known to significantly contribute to the total amount of algal biomass in the river Elbe. Only growth of diatoms, however, depends on the availability of the nutrient silica. The general concept is to consider a series of individual water bodies travelling downstream towards station Weir Geesthacht. Starting from unrealistically low values, meaningful concentrations of algal biomass to be compared with observations evolve only by the end of the water parcel's journey after exponential growth over a limited period of

travel time (cf. Figure 1). This travel time is parameterized empirically as a function of river discharge.

At the beginning of its journey each water package is initialised by the same concentration of silica. Diatoms are assumed to cease growing and to start decaying as soon as this initial reservoir of silica has been used up (cf. upper right panel of Figure 1). The higher growth rates are the earlier the diatom maximum occurs and the more pronounced it is. If growth rates are large enough so that all available silica is assimilated already upstream of Geesthacht, further increasing growth rates (i.e. more favourable growth conditions) imply decreasing diatom populations at the end of the particle's journey. Thus, assuming that diatom growth rates increase with temperature this provides a plausible explanation for negative correlations between water temperature and chlorophyll *a* concentrations at station Geesthacht. The assumed relationship between Lagrangian trajectories of individual water packages and local observations at station Geesthacht is outlined in Figure 1.

For each of the two species an individual balance equation describes changes of algal biomass concentration, C_{alg}, with time, t, as a function of temperature dependent rates of algal growth, μ, respiration, ρ, and loss, σ:

$$\frac{d\,C_{alg}}{d\,t} = \left[\mu(T) - \rho(T) - \sigma(T)\right]C_{alg} \quad \text{with} \quad \mu(T) = \mu_{max}^{0}\,q_{1}(T)\,F_{light}\,F_{Si} \tag{1}$$

The focus of the sensitivity and uncertainty study reported below is on six model parameters $x_1, \ldots, x_6$ that enter parameterizations of the algae growth rate, μ. Two parameters are the optimum growth rates μ_{max}^{0} for green algae and diatoms, respectively, that occur when algae growth is not limited by lack of light or nutrients. Next we consider for each algae species the parameter k_{light} that specifies the light intensity, at which the algae growth rate reaches 71% of its maximum possible value. The light limitation factor F_{light} in Eq. (1) results from vertically averaging over the efficient water depth, D, which is another model input parameter that has been selected for our study:

$$F_{light} = \frac{1}{D}\int_{0}^{D}\frac{I\,e^{-\lambda z}}{\sqrt{k_{light}^{2} + I^{2}\,e^{-2\lambda z}}}\,dz \tag{2}$$

Here I denotes the radiation intensity at the water surface and λ the total light attenuation coefficient due to the presence of mineral compounds and algal self shading.

The last parameter we chose is the fraction of silica, f_{Si}, in the cells of diatoms that governs the consumption rate of silica (with concentration C_{Si}) by the diatoms with concentration $C_{alg,d}$ and growth rate μ_d:

$$\frac{d\,C_{Si}}{d\,t} = -\,f_{Si}\,\mu_{d}\,C_{alg,d} \tag{3}$$

The concentration of silica determines the size of the *Michaelis-Menten* type growth limitation factor, F_{Si}, in Eq. (1) with a half saturation constant K_S (relevant only for diatoms):

$$F_{Si} = \frac{C_{Si}}{K_{S} + C_{Si}} \tag{4}$$

All model parameters have been tuned manually to obtain reasonable simulations of chlorophyll *a* and silica concentrations. The resulting reference parameter values are shown in Table 1 together with rough estimates of prior parameter uncertainties.

To run the model external forcing represented by water temperature, discharge and radiation must be specified as a function of time. Hourly temperature observations are part of the data set from Weir Geesthacht. Daily observations of discharge at station Neu-Darchau (Elbe-km 536) are available from ARGE ELBE (http://www.arge-elbe.de). Hourly mean values of global radiation were provided by the GKSS Research Centre Geesthacht which is located few kilometres upstream of the weir.

Table 1: Reference parameter values and assumed uncertainties; Optimised values for 2000

	Name	Ref. Value	Range	StD	Value 2000	Unit
x_1	$k_{light,\,g}$	20	[10-30]	5.77	17.7	W/m^2
x_2	$k_{light,\,d}$	14	[10-18]	2.31	14	W/m^2
x_3	$\mu^0_{max,\,g}$	1.65	[1.3-2.0]	$2.02\ 10^{-1}$	1.63	1/d
x_4	$\mu^0_{max,\,d}$	1.6	[1.2-2.0]	$2.31\ 10^{-1}$	1.62	1/d
x_5	f_{Si}	0.2	[0.1-0.3]	$5.77\ 10^{-2}$	0.2	mgSi /mgC
x_6	D	2.35	[1.2-3.5]	$6.64\ 10^{-1}$	2.08	m

A comparison of observations and corresponding model results is presented in Figure 2. Considering the very simple model approach with water temperature, discharge and solar radiation being the only time dependent model inputs (no time dependent initial values or sources/sinks) the model reproduces a reasonable amount of observed variability. Note that also the observed very low levels of silica during summer are reflected by the model calculations.

For the sensitivity and uncertainty study reported below we assume that only chlorophyll *a* data were available for model calibration. Silica data are considered as an independent option for checking the appropriateness of the model mechanism.

2.2. Sensitivity and Uncertainty Analysis

Let $\vec{x}$ denote the vector being made up by those model input parameters, x_i, which are to be adjusted by fitting the model to data. If $\vec{m}$ denotes the vector of model outputs, m_t, at times t, a quadratic loss function, J_{obs}, may be used to assess the differences between model output and observations, d_t, scaled by an assumed observational error, σ_{obs}. In the case of model overparameterization the minimum of the loss function will be not well-defined due to small curvatures in certain directions. To make the optimisation problem well-posed we complement the loss function by a second component that penalizes all deviations of the actual parameter vector $\vec{x}$ from the reference vector, $\vec{x}_0$, which is specified in Table 1:

$$J = J_{obs} + J_{prior} = \frac{1}{2}\sum_{t=1}^{N} \frac{\left[m_t(\vec{x}) - d_t\right]^2}{\sigma_{obs}^2} \;+\; \alpha\frac{1}{2}\left(\vec{x} - \vec{x}_0\right)^T \mathbf{V}_{prior}^{-1}\left(\vec{x} - \vec{x}_0\right) \tag{5}$$
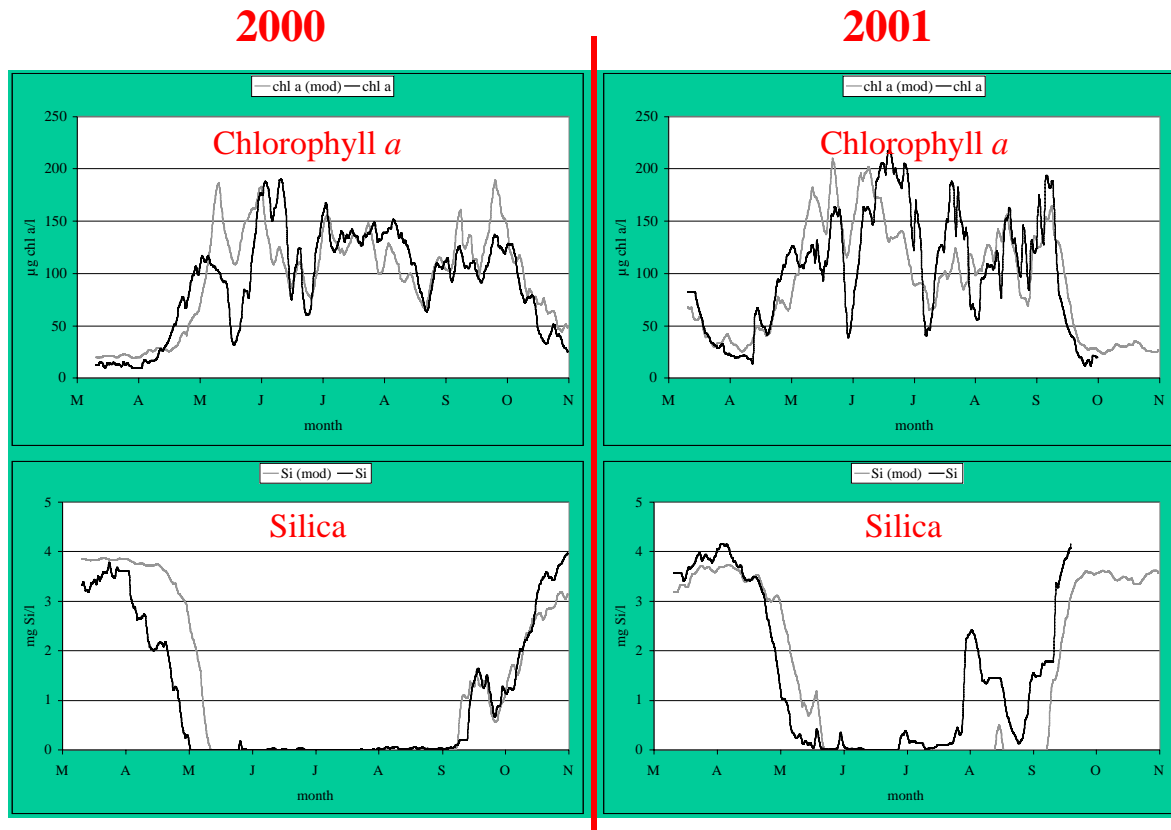
Figure 2: Observations (black) and corresponding model predictions for chlorophyll *a* (upper panels) and silica (lower panels) at station Geesthacht for two different years.

The diagonal prior parameter covariance matrix, $\mathbf{V}_{prior}$, is introduced to remove differing physical dimensions of the six parameters by proper scaling. For the present study we decided to measure changes of parameter values in terms of multiples of their estimated prior uncertainty (cf. Table 1). The scalar α enables one to adjust the overall sizes of the data and the penalty term relative to each other. Note that the size of the data term and therefore the appropriate choice of α must be a function of the number of observations. In the following we choose $\alpha=100$ for $N=78$ data points comprising about one observation every third day between March and October 2000.

The relevance of parameter covariances for a moderately non-linear model's fit to data can be analysed by examining the curvature of the loss function at its minimum. Figure 3 illustrates the general idea. Directions of high curvature allow for only weak parameter variations without getting into conflict with the data, whereas in directions with low curvature parameters can be changed significantly without much affecting the value of the loss function. A complete description of the loss function's local curvature is provided by the Hessian matrix containing all second derivatives of the loss function. Principal component analysis (PCA) of the Hessian matrix allows to discriminate those directions in the 6-dimensional parameter space that can be calibrated by the available data from other directions, along

which effects of parameter changes tend to compensate each other. For a linear model the inverse Hessian can be identified as being the posterior parameter error covariance matrix [3].
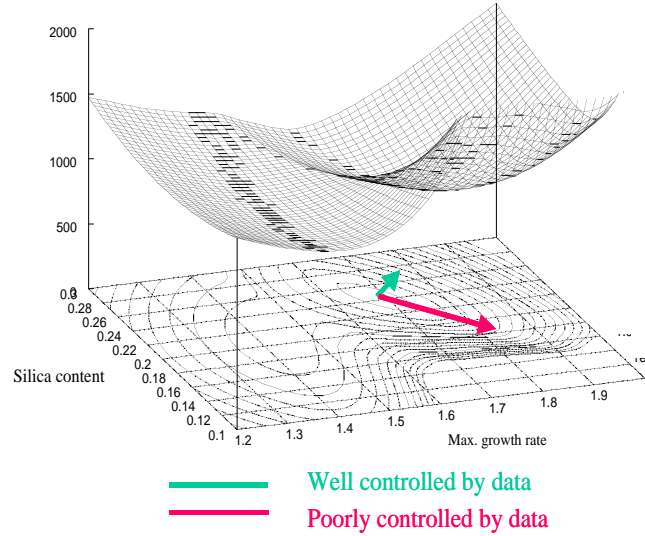

Figure 3: Loss function for variations of only two parameters; eigenvectors of the Hessian matrix defining directions of minimum and maximum curvature, respectively.

For the analysis reported below it is convenient to incorporate both the diagonal prior covariance matrix $\mathbf{V}_{\text{prior}}$ and the amplitude factor α into a coordinate transformation to non-dimensional parameters $\vec{x}'$:

$$\vec{x} = \left(\frac{1}{\alpha} \mathbf{V}_{\text{prior}} \right)^{1/2} \vec{x}' \tag{6}$$

For a linear model the loss function is quadratic. For a non-linear model the loss function should still be nearly quadratic in the vicinity of its minimum so that the following formula provides a good approximation of the Hessian and the inverse of the posterior parameter covariance matrix, $\mathbf{V}'_{\text{post}}$, [3]:

$$\left(\mathbf{V}'_{\text{post}} \right)^{-1} = \frac{\partial^2 J}{\partial \vec{x}'^2} = \frac{1}{\sigma_{\text{obs}}^2} \sum_{t=1}^{N} \left(\frac{\partial m_t}{\partial \vec{x}'} \right)^{\text{T}} \left(\frac{\partial m_t}{\partial \vec{x}'} \right) + \mathbf{I} \tag{7}$$

A programming environment for the solution of control problems (Integrated Modelling and Analysis System (IMAS, [1]) including an automatic differentiation tool for computer programs has been used for an analytic calculation of the Jacobian matrix $\partial \vec{m} / \partial \vec{x}$ and for a gradient based minimization (Quasi-Newton algorithm) of the loss function. The inclusion of prior knowledge into Eq. (7) renders the inversion of the Hessian matrix possible. The relative

strength of background knowledge has been chosen sufficiently small to make the background term negligible for modes being well controlled by the data.
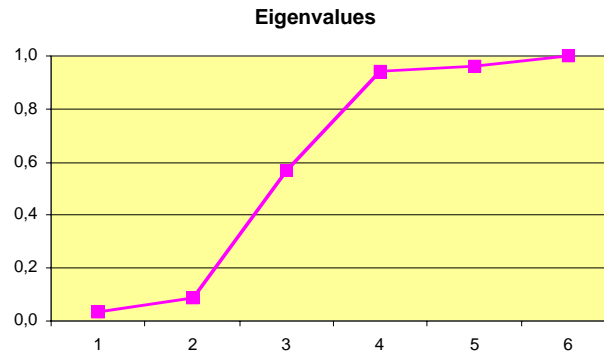
Eigenvalues

Figure 4: Eigenvalues of the posterior parameter covariance matrix analysed for year 2000.

3. RESULTS

The model has been fitted to data from the year 2000. The optimized parameter values listed in Table 1 do not deviate far from the reference values that were obtained as by careful manual model calibration. Figure 4 depicts the eigenvalues of the analysed posterior parameter error covariance matrix, which represent variances along directions in the parameter space being defined by the associated eigenvectors. According to Eq. (7) all eigenvalues of $\mathbf{V}'_{\text{post}}$ must be smaller than one as a consequence of the particular way of scaling. It turns out that only the first two eigenvalues are small enough to conclude that prior uncertainties have been overwritten by observational evidence. Three eigenvalues remain virtually unaffected by the data and stay close to the prior value one. One eigenvalue is influenced by both data and prior knowledge. Thus, PCA of the posterior parameter covariance matrix suggests that only two degrees of freedom in the six-dimensional parameter space are clearly relevant for a successful reproduction of chlorophyll *a* observations. It should be kept in mind, however, that these results are influenced by the way, in which parameters have been scaled. Scaling introduces some subjectivity but is a necessary prerequisite for PCA, which needs homogeneous physical dimensions [2].

A crucial question is whether or not the two parameter combinations, which are well controlled by the data, can be interpreted in terms of mechanistic processes represented in the model. Principal components are artificially defined variables and therefore do not necessarily have a physical interpretation. The two upper panels in Figure 5 depict the eigenvectors that correspond with the two smallest eigenvalues in Figure 4. To provide a better understanding of the mechanism of parameter calibration the bottom panel of Figure 5 presents an analysis of model prediction errors caused by uncertainties in the space of principal components. If the model is assumed to be linear, uncertainties of model parameters can be propagated independently and their effects on model output can be superimposed to each other.

It turns out that the two leading principal components that jointly explain about 89% of model output variance affect model predictions (i.e. are identifiable from the data) in distinct time intervals. This can be explained considering the different signs that diatom related

loading coefficients of the two eigenvectors have. According to both eigenvectors a strong (positive) impact on predicted chlorophyll *a* concentrations can be achieved by an intensification of the maximum growth rate of green algae, x_3, together with a reduction of their demand of light, x_1. However, according to the first eigenvector, which dominates model output uncertainty during periods when silica is lacking (cf. Figure 2), values of the corresponding parameters for diatoms, x_4 and x_2, are changed in the opposite directions. This reduces the overall diatom growth rate but nevertheless results in an increase of the final diatom biomass concentration (cf. upper right panel in Figure 1). If silica concentrations are sufficiently high, calibration of the second instead of the first eigenmode becomes crucial, which treats both algae species symmetrically.
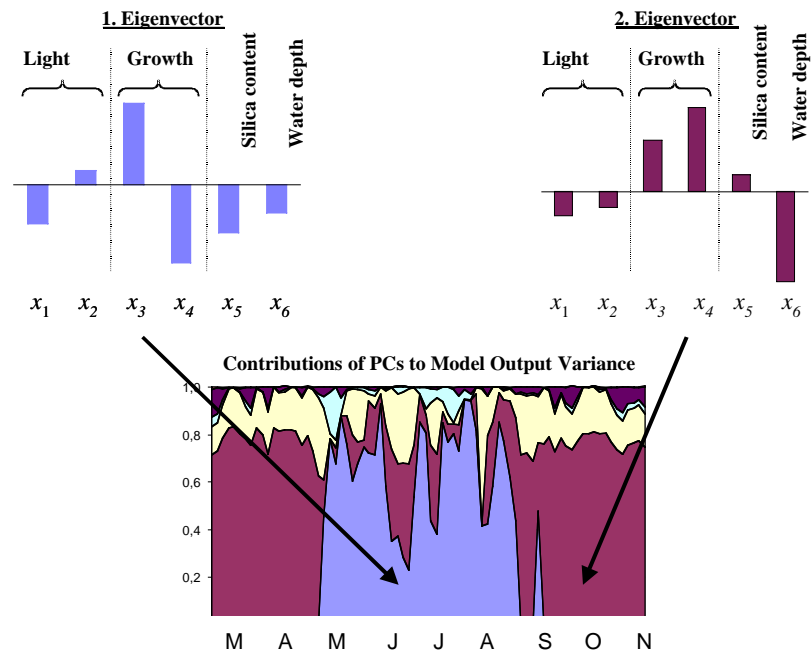

Figure 5: Cumulative plot of the relative contributions of the six principal components (amplitudes of the eigenvectors of the posterior parameter error covariance matrix) to model output variability. About 80% of model output variance can be attributed to the first two principal components.

The results of the local uncertainty analysis could be used for the implementation of a more systematic model calibration procedure. It suffices to calibrate the two or three leading principal components and to assign arbitrary values to the others. To illustrate the idea we performed a Monte Carlo experiment with 10,000 simulations. For each model run values of the six selected model parameters were drawn independently. Then those parameter combinations, which gave rise to the most successful model simulations (i.e. lowest values of the loss function), were projected onto the eigenvectors of the local parameter error covariance matrix. These projections, being standardized to have zero mean and unit standard deviation, are depicted in Figure 6. Projections onto the first three vectors are connected by solid lines. Projections onto the first two eigenvectors show small variability while anomalies

of the other projections may have large values even for the very best simulations. This proves that the choice of these amplitudes has little impact on the model's fit to data.
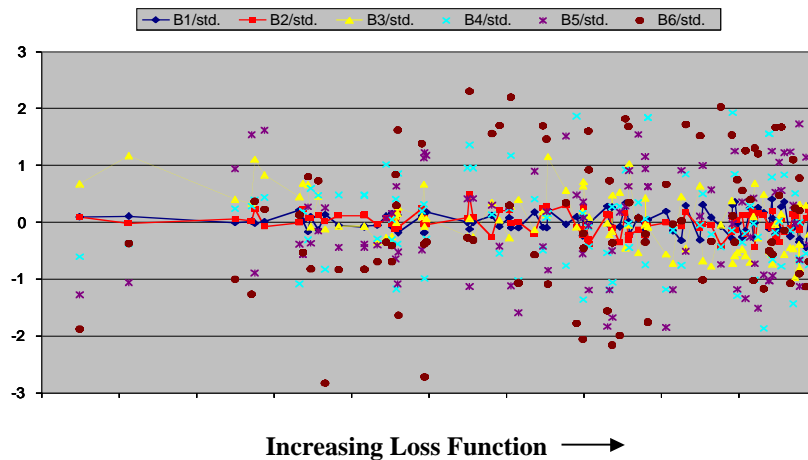

Figure 6: Monte Carlo generated parameter vectors giving rise to successful simulations are projected onto the six eigenvectors of the local parameter covariance matrix at the loss function's minimum. All projections are scaled by their respective standard deviations. Projections onto the first three eigenvectors are connected by solid lines.

4. DISCUSSION

Posterior parameter correlations reflect the fact that in the process of model calibration changing the values of different parameters had similar effects on model counterparts of the data. Taking into account such parameter interaction structures strongly mitigates uncertainties in model predictions [6]. PCA of the posterior parameter error covariance matrix gives a clear picture of how many degrees of freedom are really controlled by data. It helps to identify the nature of a model's overparameterization and its dependence on the kind of data being available for model calibration. The artificial new input parameters (principal components) do not necessarily have a physical meaning. In the present example, however, two parameter patterns being controlled by data could be related to the discrimination between different algae species in the model.

Vajda and Turányi [5] applied PCA for optimally reducing the mechanism of chemical reactions based, however, on a response function which measures model output variability but makes no reference to observed data. Using a small number of sensitive principal components as new independent model input parameters, thereby implicitly taking into account model parameter interactions, could much facilitate the adaptation of a model to new applications.

When a model is linear, uncertainties of model parameters can be propagated independently and their effects on model output can be superimposed. If a model is significantly non-linear, there are obvious limitations for a local sensitivity analysis, and a more general global method may be needed. As long as the local method is applicable, however, one can take advantage of the very basic definition of sensitivity in terms of the slope of model output at a given point in the parameter space [4]. Signs of model output

sensitivities are available rather than their pure sensitivity strengths, which facilitates an interpretation in mechanistic terms.

In the present application it turned out that the relevance of eigenmodes changed when the modelled concentration of silica dropped to zero. Accordingly any parameter change, which affects the modelled lengths of periods with lack of silica, will have an impact on the posterior parameter covariance matrix. This indicates the limitations of linear uncertainty analysis applied to the present example. Depending on a model concept with or without the inclusion of silica, model parameters must be calibrated differently, possibly giving rise to chlorophyll *a* simulations of similar quality. To consult observations of silica is the only way to resolve this ambiguity with regard to model formulation.

5. ACKNOWLEDGEMENTS

REFERENCES

1. A. Rhodin, U. Callies and D. P. Eppel. IMAS-Integrated Modeling and Analysis System, Release 1.0. Report GKSS 96/E/28, GKSS Research Centre, D-21494 Geesthacht, Germany, 1996.
2. A. Tarantola. *Inverse Problem Theory.* Elsevier, Amsterdam, 1987.
3. W. C. Thacker. The role of the Hessian matrix in fitting models to measurements. *J. Geophys. Res.,* 94, 6177-6196, 1989.
4. T. Turányi and H. Rabitz. Local Methods. In A. Saltelli, K. Chan and E. M. Scott, editors, *Sensitivity Analysis*, 81-99, Wiley, Chichester, 2000.
5. S. Vajda, S. and T. Turányi. Principal component analysis for reducing the Edelson-Field-Noyes model of the Belousov-Zhabotinsky reaction, *J. Phys. Chem.*, 90, 1664-1670, 1986.
6. G. van Straten. Maximum likelihood estimation of parameters and uncertainty in phytoplankton models. In M. B. Beck and G. van Straten, editors, *Uncertainty and forecasting of water quality,* 157-171. Springer, Berlin, 1983.