Multi-Scale Assessment of Prediction Uncertainty in Coupled Reactive Transport Models

Ming Ye (mye@fsu.edu)

Department of Scientific Computing, Florida State University, Tallahassee, FL 32306

Introduction

This final project report summarizes research activities, findings, and deliverables obtained at the Florida State University (FSU) during the funding period of 2009 – 2013. The goal of this project is: (1) to seek fundamental understanding of the factors causing and controlling predictive uncertainty in groundwater reactive transport modeling, and (2) to develop interdisciplinary tools for quantifying and reducing the predictive uncertainty. Uncertainty analysis in this project is focused on parametric uncertainty of individual models and model uncertainty between alternative models. The FSU project team was in charge of developing methods for uncertainty quantification and reduction. The methods have been applied to quantify predictive uncertainty in uranium reactive transport modeling at the Naturita Site.

During this project period, we published twelve peer-reviewed journal articles (one is under revision and one under review) and four conference proceedings. In particular, the paper of Lu et al. (2012) published in Water Resources Research was selected as the Editor's Highlight for its new insights into faster computation of uncertainties. We presented our research results in multi-disciplinary conferences such as geology, mathematics, geochemistry, and civil engineering. The publication on the SIAM News (Hill et al., 2012) introduced our interdisciplinary research results to members of the Society of Industrial and Applied Mathematics (SIAM). A master student (Geoffery Miller) and a doctoral student (Dan Lu) graduated with support of this project. The doctoral student received the 2012 Student Travel Fellowship and presented her research at the 2012 annual PI meeting; she is working as a post-doc at the Oak Ridge National Laboratory. We also collaborated with colleagues in the Los Alamos and Oak Ridge National Laboratories to help their research using the methods developed in this project. The collaboration produced one journal article (Dai et al., 2012) and one conference proceeding (Shi et al., 2012).

Scientific Questions and Project Findings

This project explored the scientific questions below related to various aspects of uncertainty analysis in coupled groundwater reactive transport models.

Question 1: How does uncertainty behave for groundwater reactive transport models?

This question was tackled by using the surface complexation model developed by Kohler et al. (1996). The model, denoted as C4 hereinafter, has two functional groups. The weak site, S_1OH , is associated with one reaction:

and the strong site, S_2OH , with two reactions:

where S1 and S2 denote the adsorption sites. The uncertain parameters are the three equilibrium constants, logK1, logK2, and logK3 of the three reactions, respectively, and the fractions of the two sites. Since the fractions sum to one, the fraction of the strong site, logSite, is explicitly treated as an uncertain variable. Among the seven column experiments conducted by Kohler et al. (1996) to understand uranium reactive transport, three experiments, denoted as Expt_1, Expt_2, and Expt_8, were used in the analysis. They were performed under different chemical conditions of pH and initial uranyl concentration shown in Table 1. A sensitivity analysis (using both local and global techniques) revealed that logK1 is the most sensitive parameter, followed by logSite, logK2, and logK3 in the descending order of sensitivity. More details of the sensitivity analysis are given in Figure 4 of Appendix A, the manuscript of Shi et al. (2013) submitted for publication in Water Resources Research.

Experiment	pH	[U(VI)] [μ M]
1	4.26 ± 0.02	1.0
2	3.90 ± 0.02	1.0
8	4.22 ± 0.02	50.0

Table 1. Conditions for column experiments 1, 2, and 8.

For the four random parameters, numerous random samples were sampled. For each sample, the three experiments (Expt_1, Expt_2, and Expt_8) were simulated, and the sum of squared weighted residuals (differences between observed and corresponding simulated uranium concentrations) was calculated to measure the overall discrepancy (or goodness-of-fit) between observations and corresponding model simulations. Figure 1 plots the sum of squared weighted residual (SSWR) with the two most sensitive parameters (i.e., logK1 and logSite) identified in this project. The peak marked in Figure 1 corresponds to the best model fit. Note that the magnitude of SSWR is represented by color. More details of constructing the response surface can be found in Appendix A, the manuscript of Shi et al. (2013) submitted for publication in Water Resources Research.

The dramatic fluctuations of SSWR indicate that a slight change of parameter values may cause significant change in the model fit. When the model is used for prediction, uncertainty in model parameters may cause significant predictive uncertainty. The predictive uncertainty needs to be quantified, when the model predictions are used for decision-making.

The SSWR response surface is significantly different from that of linear models, which is always smooth with one minimum value corresponding to optimum parameter optima. The fluctuation and multiple peaks (and valleys) shown in Figure 1 suggests that there are multiple parameter optima and that find the global optima is difficult in that a search algorithm may be trapped at local optima.

Figure 1. Two-dimensional surface of $-\log_{10}SSWR$ when $\log K1$, $\log Site$, and $\log K2$ vary simultaneously (the least important parameter, $\log K3$, is fixed at its global optima). $SSWR$ stands for sum of squared weighted residuals, a measure of model fit. The peak marked in the figure corresponds to the global optimum. This figure was published in Shi et al. (2013).

Question 2: What cause the uncertainty in groundwater reactive transport modeling?

It is critical to identify the sources of predictive uncertainty so that the uncertainty can be reduced if more resources are available to conduct more research and to collect more data. The following three sources of uncertainty were identified this study:

- (1) Uncertainty in model parameters,
- (2) Uncertainty in model structures, and
- (3) Uncertainty in experimental data.

Parametric Uncertainty: Taking model C4 as an example, Table 2 lists three sets of parameter values used to simulate Expt_8. Figure 2a shows that, although the three parameter sets are dramatically different, the simulated breakthrough curves corresponding to the three parameter sets are essentially the same, and they all match well the experimental data. The first parameter set is the global optimum of the parameters given in Appendix A of Shi et al. (2013), and the simulated concentrations of S_1OHUO_2OH , S_2OHUO_2OH , and S_2OUO^{2+} (the products of the three reactions) are plotted in Figure 2b. In the second parameter set, while $\log K1$ and $\log Site$ values are fixed to their optimum values, the values of $\log K2$ and $\log K3$ are decreased and increased, respectively. As shown in Figure 2c, the breakthrough curve of S_1OHUO_2OH does not change because $\log K1$ and $\log Site$ are fixed at their optimum values. However, the simulated concentrations of S_2OHUO_2OH are essentially zero due to decrease of $\log K2$, and those of S_2OUO^{2+} increase due to increase of $\log K3$. Similarly, when $\log K2$ increases but $\log K3$ decreases as in the third parameter set, simulated concentrations of S_2OHUO_2OH increase but those of S_2OUO^{2+} becomes zero, as shown in Figure 2d. Because of the competition between the two reactions associated with the strong site, multiple parameter sets give similar breakthrough curves that are reasonably close to the experimental data. In other words, for the individual

experiment, if measurements of $\text{S}_2\text{OHUO}_2\text{OH}$ and $\text{S}_2\text{OUO}_2^{2+}$ are not available, one cannot remove the uncertainty in $\log K_2$ and $\log K_3$. We will present this research finding in the 2013 annual meeting of the American Geophysical Union (AGU), and are preparing a manuscript to publish this research finding together with other findings.

Parameter Set	$\log K_1$	$\log K_2$	$\log K_3$	$\log \text{Site}$
Case 1	-4.95	-3.23	1.27	-1.62
Case 2	-4.95	-4.8	1.6	-1.62
Case 3	-4.95	-2.4	0.1	-1.62

Table 2. Three parameters sets that give similar model fit to experimental data of Expt_8.

Figure 2. (a) Observed and simulated breakthrough curves using the three parameter sets listed in Table 2. (b-d) Simulated breakthrough curves of the three species for the three cases listed in Table 2. While the simulated $\text{S}_2\text{OUO}_2\text{OH}$ and S_2OUO_2 concentrations are dramatically different for parameter sets 2 and 3, their sums are similar. The competition between the reactions and the substitution effect is one of the sources of parametric uncertainty.

Model Uncertainty: Model uncertainty refers to the situation that there are multiple models supported by available data and information. This is prevalent in groundwater reactive transport modeling. Table 3 lists four alternative models considered in Kohler et al. (1996) to simulate uranium (VI) adsorption. Three of them (C4 – C6) can well simulate the three experiments (Expt_1, Expt_2, and Expt_8). In particular, the simulations of Expt_8 of the four models are visually identical, as shown in Figure 3. It is therefore uncertain to use which model to simulate uranium adsorption. The tradition is to select the best model and discard the other models. Given the model uncertainty, this practice is inappropriate, as it disregard model uncertainty and may

lead to biased model predictions and/or underestimation of predictive uncertainty. More details of quantifying the model uncertainty can be found in Appendix B, the paper of Lu et al. (2013) published in Water Resources Research.

Model	Reactions	Estimated parameter	
C3	$S_1OH + UO_2^{2+} + H_2O = S_1OUO_2OH + 2H^+$	logK1	
	$S_2OH + UO_2^{2+} = S_2OUO_2^+ + H^+$	logK2	logSite
C4	$S_1OH + UO_2^{2+} + H_2O = S_1OUO_2OH + 2H^+$	logK1	
	$S_2OH + UO_2^{2+} + H_2O = S_2OUO_2OH + 2H^+$	logK2	logSite
	$S_2OH + UO_2^{2+} = S_2OUO_2^+ + H^+$	logK3	
C5	$S_1OH + UO_2^{2+} + H_2O = S_1OUO_2OH + 2H^+$	logK1	
	$S_2OH + UO_2^{2+} + H_2O = S_2OUO_2OH + 2H^+$	logK2	logSite
	$S_2OH + UO_2^{2+} = S_2OUO_2^+ + H^+$	logK3	
	$S_3OH + UO_2^{2+} + H_2O = S_3OUO_2OH + 2H^+$	logK4	
C6	$S_1OH + UO_2^{2+} + H_2O = S_1OUO_2OH + 2H^+$	logK1	
	$S_2OH + UO_2^{2+} + H_2O = S_2OUO_2OH + 2H^+$	logK2	logSite
	$S_2OH + UO_2^{2+} = S_2OUO_2^+ + H^+$	logK3	
	$S_3OH + UO_2^{2+} = S_3OUO_2^+ + H^+$	logK4	

Table 3. Four alternative surface complexation models and parameters to be estimated.

Figure 3. Observed and simulated breakthrough curves for Experiment 8. The four alternative models give the same predictions that match well the experimental data. This figure was published in Lu et al. (2013), which is included in Appendix B of the report.

Data Uncertainty: While it is well known in a qualitative sense that uncertainty depends on data used for uncertainty quantification, the dependence has not been studied quantitatively, especially for groundwater reactive transport modeling. The dependence was investigated with the aid of Monte Carlo techniques, which can explore the entire parameter space by running the model with numerous parameter values. Figure 4 plots the response surface of the negative log objective function ($-\log_{10}SSWR$) with the two most influential parameters ($\log K1$ and $\log Site$) based on individual experimental data and all the data. Figures 4a – 4c show that, for the individual experimental data, there are multiple minima, i.e., the multiple peaks in the figure along the valley shown in the subplots. This renders it difficult to find global optimum of model parameters using either global or local optimization methods. However, Figure 4d shows that, for the data of all the three experiments, the global optimum becomes distinct and can be easily obtained. From the uncertainty perspectives, the parametric uncertainty is the smallest when all the data are used for uncertainty quantification. How to combine data of different experiments is the key to uncertainty reduction. This scientific understanding provides guidelines of designing experiments for parameter estimation and uncertainty reduction. We will present this research at the 2013 annual meeting of the American Geophysical Union, and are preparing a manuscript to publish this research.

Figure 4. Contours of negative log objective function ($-\log_{10}SSWR$) based on data of (a) Experiment 1, (b) Experiment 2, and (c) Experiment 8 and (d) the three experiments combined. Parameters $\log K1$ and $\log Site$ vary but $\log K2$ and $\log K3$ are fixed at their global optima. This figure shows uncertainty reduction after using all the three experiments.

In the rest of this report, we will summarize the methods developed and evaluated in this project for quantifying parametric and model uncertainty in groundwater reactive transport modeling. Subsequently, we present a data-worth analysis developed for uncertainty reduction.

Question 3: How to quantify parametric uncertainty of groundwater reactive transport modeling?

Quantifying parametric uncertainty of individual models is the foundation of quantifying model uncertainty. Among the various techniques of uncertainty quantification, the most widely used ones are the regression methods and Bayesian methods. For quantification of parametric uncertainty, the regression methods use the confidence intervals, while the Bayesian methods use the credible intervals. After evaluating which methods are more appropriate for groundwater reactive transport modeling, we answered the following questions from both theoretical and practical viewpoints:

What are the relations between the confidence intervals of regression methods and credible intervals of Bayesian methods?

If model residuals are Gaussian, a theoretical study conducted by Lu et al. (2012) found that:

- 1) While the regression and Bayesian methods are theoretically different, the two kinds of intervals are identical for linear and linearized nonlinear models.
- 2) For nonlinear models, nonlinear confidence and credible intervals can be numerically identical, if the following conditions are satisfied: (1) parameter confidence regions (defined using approximate likelihood method) and parameter credible regions (estimated using MCMC realizations) are numerically identical, and (2) predictions are a smooth, monotonic function of the parameters. Both occur if intrinsic model nonlinearity is small.
- 3) While the conditions of Gaussian errors and small intrinsic model nonlinearity are violated by many environmental models, heuristic tests using analytical and numerical models suggest that linear and nonlinear confidence intervals can be useful approximations of uncertainty even under significantly non-ideal conditions.
- 4) In the context of epistemic model error for a complex synthetic nonlinear groundwater problem, the linear and nonlinear confidence and credible intervals for individual models performed similarly enough to indicate that the computationally frugal confidence intervals can be useful in many circumstances.

The above conclusions were demonstrated using various numerical models, ranging from simple analytical expressions to synthetic groundwater flow models with realistic complexity. Figure 5 shows an example for three groundwater flow models with homogeneous (HO), zonal (3Z), and heterogeneous (INT) hydraulic conductivity. For the three models with different levels of complexity, the linear and nonlinear confidence intervals and the nonlinear credible intervals are similar. For the most complex models, the intervals all include the true value of stream change (due to pumping) marked by the red line. The computational cost for the nonlinear credible interval is the highest, requiring about half millions of model runs. The computational cost for the linear confidence interval is the lowest, only requiring about 100 model runs. The linear confidence interval is thus favored, considering its low computational cost.

Figure 5. Linear and nonlinear confidence intervals and nonlinear credible intervals (≤ 106 , ≤ 1594 , and 420,000 model runs, respectively) for prediction of streamflow change. The true value is marked by the red line. The three kinds of intervals are similar. This figure was published in Lu et al. (2012), which is included in Appendix C of this report.

To further evaluate the above findings, two numerical studies were conducted for an unsaturated flow model (Shi et al., 2012) and a groundwater reactive transport model (Shi et al., 2013); the latter two models have higher model nonlinearity than the groundwater flow models of Lu et al. (2012). These two extra numerical studies reached the same conclusions as the theoretical study of Lu et al. (2012) for groundwater models. Figure 6 plots the probability density functions of the four parameters of model C4 discussed above. This figure shows that the parameter distribution obtained from the regression methods are similar to those of Bayesian methods, which explains similarity between the confidence and credible intervals.

Figure 6. Parameter distributions of the four parameters of model C4 obtained using the regression methods (blue lines) and DREAM_(zs) algorithm (histograms). The red lines indicate the global optimum of sum of square weighed residuals. The distributions are based on the assumption of Gaussian residuals.

Given the similarity between the regression confidence intervals and Bayesian credible intervals and considering the low computational cost of the regression methods, we used the regression methods for quantifying parametric uncertainty of groundwater reactive transport modeling in the rest of the project.

Question 4: How to quantify model uncertainty of groundwater reactive transport modeling?

Model uncertainty is quantified in this project using the Bayesian model averaging (BMA) method. Instead of making predictions using a single model, BMA first develops alternative models based on available data and knowledge. Plausibility of these models is evaluated quantitatively by model calibration. Implausible models are discarded, if they cannot yield comparable simulations of the observations. For the remaining models, their relative probabilities are estimated as model averaging weights that are used to calculate weighted average of the predictions of the plausible models. The key issue in the model averaging method is the estimation of model averaging weights, and it is the focus of this project.

There are two kinds of methods for estimating the model averaging weights: the Monte Carlo methods and the maximum likelihood methods. Theoretically speaking, the maximum likelihood methods can be viewed as an approximation of the Monte Carlo methods. The approximation is necessary because the Monte Carlo methods require hundreds of thousands of model evaluation, which is always computationally impractical to groundwater reactive transport modeling. However, accuracy of the approximation is still unknown. Therefore, we first investigate the question below in this project:

How accurate are the model averaging weights evaluated using the maximum likelihood methods in comparison with those obtained using the Monte Carlo methods?

The maximum likelihood approximations of the model averaging weights are evaluated using the Bayesian information criterion (BIC) and the Kashyap information criterion (KIC), which have been widely used in groundwater modeling (Ye et al., 2004, 2008; Poeter and Hill, 2007). The Markov chain Monte Carlo (MCMC) method was used to estimate the model averaging weights as the reference for performance evaluation of BIC and KIC. The evaluation of accuracy of the maximum likelihood approximation was conducted using a synthetic study in Lu et al. (2011). Computational cost of the models in the synthetic study is relatively low and thus allows us to investigate various aspects of the problem. In particular, the evaluation was conducted using replicate datasets so that the conclusions are more comprehensive than using a single data set. Key findings of this synthetic study are as follows:

1. KIC leads to more accurate results than BIC for approximating model averaging weights. Although KIC reduces asymptotically to BIC, KIC provides consistently more reliable indications of model quality for a range of sample sizes; BIC may select inferior models.
2. KIC based on observed Fisher information matrix (FIM) is more reliable than KIC based on an approximation of expected FIM computed using a single sample, the difference between the two increasing with diminishing sample size.

3. Estimates of model probability associated with KIC (observed FIM based) are close to but not identical to corresponding MCMC-based estimates. However, MCMC simulation may require a much larger computational effort.
4. Fisher information term in KIC allows (a) imposing a more severe penalty on models having greater complexity (number of parameters) and (b) differentiating more accurately between models of disparate structures than is possible with other criteria that do not include a Fisher term.

Based on these results, we used KIC for evaluating model averaging weights in the rest of the project for groundwater reactive transport models.

U(VI) Surface Reaction	Site	log K	Site fraction (f)
$S_1OH + UO_2^{2+} + H_2O = S_1OUO_2OH + 2H^+$	weak site	-4.9748	0.967979
$S_2OH + UO_2^{2+} + H_2O = S_2OUO_2OH + 2H^+$	strong site	-3.4547	0.031819
$S_2OH + UO_2^{2+} = S_2OUO_2^+ + H^+$	strong site	0.6113	
$S_3OH + UO_2^{2+} + H_2O = S_3OUO_2OH + 2H^+$	stronger site	-1.1926	0.0002
$S_4OH + UO_2^{2+} = S_4OUO_2^+ + H^+$	strongest site	2.8388	0.000002

Table 4. True surface complexation reactions and parameters of the true model in the synthetic study. Total site density used in this model is 1.3E-03 M/L. Summation of site fraction is one.

A particular challenge to evaluating model averaging weights for groundwater reactive transport models is the spatial and/or temporal correlation in residuals, which are the difference between observations and corresponding model simulations. The residuals consist of measurement errors and model errors. While it is reasonable to assume that the measurement errors are uncorrelated in space and time, model errors are always correlated in space and time. For example, if a model cannot simulate concentration at one time, it is likely that it cannot give a good simulation of the next time. The correlation however is always ignored when calculating model averaging weights, and its impact has not been explored. We thus investigated the following question:

What is the impact of disregarding the residual correlation and how to incorporate the correlation in the evaluation of model averaging weights?

This problem was addressed using synthetic and real-world data of groundwater reactive transport models (Lu et al., 2013). While only temporal correlation was considered, the conclusions of this study should be applicable to spatial correlation. The results of the synthetic study are briefly reported here, and more results can be found in Lu et al. (2013), which is included in Appendix B of the report. The synthetic study considers a true surface complexation model listed in Table 4, and four alternative models listed in Table 3. The true model was used to

generate synthetic data with the different geochemical conditions of the six experiments of Kohler et al. (1996), three of which are listed in Table 1. The synthetic data, after being corrupted with measurement errors, were used for model calibration and predictive analysis.

We used the synthetic data of Expt_1, Expt_2, and Expt_8 for the calibration, which was performed using the regression method implemented in UCODE_2005. The residuals of breakthrough curves are plotted in the left panel of Figure 7; they are random, and temporal correlation is not observed. This is confirmed by the auto-correlation functions (ACFs) plotted in the right panel of Figure 7. However, for the alternative models, the residual correlation is apparent, as shown in the left panel of Figure 8 for alternative model C5, the best model. This correlation is artificial and caused by error in model structure. When the correlation was disregarded, model C5 receives 100% model averaging weight, which is physically unreasonable, because the true model is a combination of models 5 and 6. In other words, models 5 and 6 should have similar model averaging weights.

Figure 7. Plot of residuals of the true model for (a1) Expt_1, (a2) Expt_2, and (a3) Expt_8. No temporal correlation is observed in the figures. This is also confirmed by the corresponding sample auto-correlation functions (ACFs) plotted in figure (b1)-(b3), because all ACFs of lag larger than zero are smaller than the threshold values marked by the blue lines. The ACF of lag zero by definition is one, because it is auto-correlation of data themselves. The results were published in Lu et al. (2013), which is included as Appendix B of the report.

Figure 8. Plots of residuals (left panel) and residuals-AR(p) (right panel) of model C5 for (a1-b1) Experiment 1, (a2-b2) Experiment 2, and (a3-b3) Experiment 8 in the synthetic study. The results were published in Lu et al. (2013), which is included as Appendix B of the report.

To incorporate the residual correlation into the evaluation of model averaging weights, our solution is to build the correlation into the covariance matrix of the regression methods so that the residuals can be properly weighted. The correlation is described using the time series methods. After the residual correlation is considered by properly weighting, the temporal correlation is removed, as shown in the right panel of Figure 8. After the residual correlation is considered, models C5 and C6 have probabilities of 60% and 40%, respectively. This is reasonable based on understanding of the true reactive transport model and the alternative models. In other words, using the proper weighting resolved the problem that a single model received 100% model averaging weight.

The difference in the estimates of model averaging weights dramatically affects predictive performance of model averaging. If model C5 has model probability of 100%, model averaging cannot be performed, and model predictions are based on a single model. When the model averaging weights are partitioned into the two most plausible models (C5 and C6), model averaging can be performed, and predictive performance of model averaging is superior to that of individual models. This is shown in Figure 9 that plots the prediction of synthetic data of Expt_3 (black dots) using individual models (blue lines) and model averaging (red lines). For the individual models, its uncertainty bound (95% confidence interval) cannot cover all the data. However, the uncertainty bound of model averaging is able to cover all the data, indicating improvement of predictive performance. The improvement is reflected not only in the increase of predictive coverage but also in reduction of predictive bias (results not shown).

Figure 9. Predictions (solid lines) and their 95% linear confidence intervals (dashed lines) of Experiment 3 (black dots) based on single models (blue lines) and model averaging (red lines) in the synthetic study. Predictive performance of model averaging is superior to that of individual models. The results were published in Lu et al. (2013), which is included as Appendix B of the report.

Question 5: How to reduce predictive uncertainty by collecting data of maximum value of information or data-worth?

As discussed above, data uncertainty is one source of predictive uncertainty. In other words, predictive uncertainty can be reduced by collecting more data (or data with more information). The challenging question is how to collect the data of maximum value of information, i.e., achieving the maximum uncertainty reduction by collecting the minimum amount of data. Collaborating with Shlomo Neuman in the University of Arizona, we developed a Bayesian framework that uses a multimodel approach (i.e., model averaging) to optimum value-of-information or data-worth analysis (Neuman et al., 2012). A unique contribution of this method is that both parametric and model uncertainty is considered. Another uniqueness of this method is that, using the total law of covariance, it rigorously proves uncertainty reduction resulted from collecting more data. The firm theoretical basis makes this method general, not specific to certain kinds of problems.

The Bayesian approach of data-worth analysis was first implemented for a synthetic geostatistical problem in two space dimensions (Neuman et al., 2012). It was extended to a real-world case of log air permeability data ($\log k$) from unsaturated fractured tuff at the Apache Leap Research Site (ALRS) near Superior, Arizona (Lu et al., 2012). Only the major results of the

real-world application are included in this report. More details of this study can be found in Appendix D of this report. Spatially distributed $\log k$ data were obtained from 184 pneumatic injection tests along 6 boreholes at the site (Figure 10). We considered the following two “cross-validation” cases:

- (1) CV I: $\log k$ measured in W2a, Y3 and Z2 play the role of existing data **D**, boreholes X2 and Y2 are the sites of potential new data **C**₁ and **C**₂, the goal being to predict $\log k(\Delta)$ along V2;
- (2) CV II: $\log k$ measured in W2a, X2 and Y2 play the role of **D**, boreholes V2 and Z2 are the sites of potential new data **C**₁ and **C**₂, the goal being to predict $\log k(\Delta)$ along Y3.

Given that in each case one has funds to measure $\log k$ in only one borehole, which among boreholes X2 and Y2 should one conduct such measurements in case CV I, and which among V2 and Z2 in case CV II?

The issue of data-worth is compounded by uncertainty about the correct geostatistical model and parameters to be employed in each case. We considered three alternative variogram models: exponential (Exp0), spherical (Sph0), and power (Pow0). Model parameters are coefficients of the variograms. Mean and covariance of Δ are estimated using kriging methods. The objective function is reduction of predictive uncertainty of Δ , i.e., to select, among two competing locations, the borehole location with larger uncertainty reduction. This was first done using true data. Based on the actual variance reductions listed in Table 5, boreholes X2 and V2 are selected as the optimum target for data collection for Case I and Case II, respectively. Assuming the true data are unknown, they are generated based on **D** and used to estimate variance reductions. The borehole locations selected by the estimated variance reductions are the same as those by the actual reductions. Due to its mathematical rigorous, this method is applicable to general environmental problems, including those of groundwater reactive transport modeling.

Figure 10. Spatial locations of 184 $\log_{10}k$ data used in the proof-of-concept study. This figure is published in Lu et al. (2012), which is included in Appendix D of this report.

	CV I		CV II	
	C ₁ (X2)	C ₂ (Y2)	C ₁ (V2)	C ₂ (Z2)
Actual	2.31	1.01	3.11	2.84
Predicted	0.59	0.42	1.79	0.84

Table 5. Actual and predicted trace variance reductions in two cross-validation cases, CV I and CV II. The actual and predicted uncertainty reductions suggest the same boreholes (highlighted in red) as the optimum target for data collection in the two cases. This result is published in Lu et al. (2012), which is included in Appendix D of this report.

List of Publications

Peer-Reviewed Journal Articles

1. Shi, X. (*post-doc*), **M. Ye**, G.P. Curtis, G.L. Miller, P.D. Meyer, M. Kohler, S.B. Yabusaki, and J. Wu (2013), Assessment of parametric uncertainty for surface complexation modeling of uranium reactive transport, *Water Resources Research*, Under Revision.
2. Lu, D. (*student*), **M. Ye**, M.C. Hill, E.P. Poeter, and G.P. Curtis (2013), Integration of Markov chain Monte Carlo simulation into UCODE for Bayesian uncertainty analysis, *Environmental Modeling and Software*, Under Review.
3. Lu, D. (*student*), **M. Ye**, P.D. Meyer, G.P. Curtis, X. Shi, X.-F. Niu, and S.B. Yabusaki (2013), Effect of error covariance structure on estimation of model averaging weights and predictive performance, *Water Resources Research*, 49, 1-20, doi:10.1002/wrcr.20441.
4. Lu, D. (*student*), **M. Ye**, and M.C. Hill (2012), Analysis of Regression Confidence Intervals and Bayesian Credible Intervals for Uncertainty Quantification, *Water Resources Research*, 48, W09521, doi:10.1029/2011WR011289.
5. Shi, X. (*post-doc*), **M. Ye**, S. Finsterle, and J. Wu, Comparing nonlinear regression and Markov chain Monte Carlo methods for assessment of predictive uncertainty in vadose zone modeling, *Vadose Zone Journal*, 11(4), doi:10.2136/vzj2011.0147.
6. Lu, D. (*student*), **M. Ye**, S.P. Neuman, and L. Xue (2012), Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs, *Advances in Water Resources*, 35, 69-82, DOI: 10.1016/j.advwatres.2011.10.007.
7. Neuman, S.P., L. Xue, **M. Ye**, and D. Lu (2012), Bayesian analysis of data-worth considering model and parameter uncertainties, *Advances in Water Resources*, 36, 75-85, doi:10.1016/j.advwatres.2011.02.007.
8. Dai, Z., A. Wolfsberg, P. Reimus, H. Deng, E. Kwicklis, M. Ding, D. Ware, and **M. Ye** (2012), Identification of sorption processes and parameters for radionuclide transport in fractured rock, *Journal of Hydrology*, 414-415, January, 220-230, doi/10.1016/j.jhydrol.2011.10.035.
9. Lu, D. (*student*), **M. Ye**, and S.P. Neuman, Dependence of Bayesian model selection criteria and Fisher information matrix on sample size, *Mathematical Geosciences*, 43, 971-993, DOI 10.1007/s11004-011-9359-0.

10. **Ye, M.**, D. Lu, S. P. Neuman, and P. D. Meyer (2010), Comment on “Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window” by Frank T.-C. Tsai and Xiaobao Li, *Water Resources Research*, 46, W02801, doi:10.1029/2009WR008501.
11. **Ye, M.**, P.D. Meyer, Y.-F. Lin, and S.P. Neuman (2010), Quantification of model uncertainty in environmental modeling, *Stochastic Environmental Research and Risk Assessment*, DOI 10.1007/s00477-010-0377-0 (preface to a special issue).
12. **Ye, M.** (2010), MMA: A computer code for multi-model analysis, *Ground Water*, 48(1), 9 – 12, doi: 10.1111/j.1745-6584.2009.00647.x (invited software review).

SIAM News

Hill, M.C., D. Kavetski, M. Clark, **M. Ye**, and D. Lu (2012), Uncertainty quantification for environmental models, *SIAM News*, 45(9), November Issue.

Conference Proceedings

1. Shi, X., G. Tang, **M. Ye**, W.-M. Wu, C.S. Criddle, J. Parker, D. Watson, S.C. Brooks, F. Zhang, and J. Wu (2012), Uncertainty quantification for biogeochemical models of ethanol-stimulated uranium (VI) reduction, XIX International Conference on Water Resources, CMWR 2012, University of Illinois at Urbana-Champaign, June 17-22,
2. **Ye, M.**, D. Lu, S.P. Neuman, and L. Xue (2011), Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs, Proc. Intern. Conf. Groundwater: Our source of Security in an Uncertain Future, Geol. Soc. South Africa and Intern. Assoc. Hydrogeologists, Pretoria, South Africa, Sept. 19 - 21.
3. Neuman, S.P., L. Xue, **M. Ye**, and D. Lu (2011), Multimodel Assessment of the Worth of Data Under Uncertainty, Waste Management Symposium, February 28 -March 3, 2011, Phoenix, Arizona.
4. Lu, D., M.C. Hill, and **M. Ye** (2011), Analysis of regression and Bayesian predictive uncertainty measure. In R.M. Maxwell, E.P. Poeter, M.C. Hill, and C. Zheng (Ed.), *MODFLOW and More 2011: Integrated Hydrologic Modeling*. Integrated Groundwater Modeling Center, Colorado School of Mines, 454 – 458.

Master Thesis:

Geoffery L. Miller (2010), Parametric Uncertainty Analysis of Uranium Transport Subsurface Complexation Models, Department of Scientific Computing, Florida State University, Tallahassee, FL.

Doctoral Dissertation:

Lu, Dan (2013), Assessment of Conceptual Model Uncertainty of Groundwater Flow and Solute Transport, Department of Scientific Computing, Florida State University, Tallahassee, FL.

Conference Abstracts

1. **Ye, M.**, D. Lu, and S.P. Neuman (2012), Dependence of Bayesian model selection criteria and Fisher information matrix on sample size, Annual meeting of the Geological Society of America, November 3 – 7, 2012, Charlotte, NC.

2. Lu, D., **Ye, M.**, M.C. Hill, and G.P. Curtis (2012), Integration of Markov Chain Monte Carlo simulation into UCODE for Bayesian uncertainty analysis, Annual meeting of the Geological Society of America, November 3 – 7, 2012, Charlotte, NC.
3. Curtis, G.P., Lu, D., **Ye, M.**, and M. Kohler (2012), Calibrations and predictions of uranium transport in small scale field experiments, Annual meeting of the Geological Society of America, November 3 – 7, 2012, Charlotte, NC.
4. **Ye, M.**, D. Lu, S.P. Neuman, and L. Xue (2012), Multimodel Bayesian Analysis of Data-Worth: Theory and Applications, XIX International Conference of Computational Methods in Water Resources, June 17 -21, 2012, Urbana-Champaign, IL.
5. Curtis, G.P., **M. Ye**, M. Kohler, M. Hay, P.D. Meyer, S.B. Yabusaky, and J. Davis (2011), Evaluation of prediction uncertainty of uranium transport in tracer test experiments, AGU Fall Meeting, December 5-9, San Francisco, CA.
6. Hill, M.C., D. Kavetski, M. Clark, **M. Ye**, M. Arabi, D. Lu, L. Foglia, and S. Mehl (2011), The “how” of environmental modeling: toward enhanced transparency and refutability, AGU Fall Meeting, December 5-9, San Francisco, CA.
7. Shi, X., **M. Ye**, G.P. Curtis, D. Lu, P.D. Meyer, S.B. Yabusaky, J. Wu (2011), Assessment of parametric uncertainty in groundwater reactive transport modeling using Markov chain Monte Carlo techniques, AGU Fall Meeting, December 5-9, San Francisco, CA.
8. Xue, L., S.P. Neuman, **M. Ye**, and D. Lu (2011), Multimodel Bayesian optimization of data collection schemes, AGU Fall Meeting, December 5-9, San Francisco, CA.
9. **Ye, M.**, D. Lu, S.P. Neuman, and P.D. Meyer (2011), Effect of data correlation on model weights, AGU Fall Meeting, December 5-9, San Francisco, CA.
10. Lu, D., **M. Ye**, G.P. Curtis, P.D. Meyer, and S.B. Yabusaky (2011), Effect of temporal residual correlation in model weights, AGU Fall Meeting, December 5-9, San Francisco, CA.
11. **Ye, M.**, X. Shi, and G.P. Curtis (2011), Identifying Multimodal, Non-Gaussian Parameter Distributions in Groundwater Reactive Transport Modeling, Cross-Pacific Conference on Environmental and Water Resources, September 3 – 5, Orlando, FL.
12. Lu, D., M.C. Hill, and **M. Ye** (2011), Analysis of regression and Bayesian predictive uncertainty measure, MODFLOW and More 2011 Conference, June 5 – 8, 2011, Golden, CO.
13. **Ye, M.**, D. Lu, S.P. Neuman, and L. Xue, (2011), Multimodel Assessment of the Worth of Data Under Uncertainty, International Conference on Groundwater: Our Source of Security in an Uncertain Future, September 19 – 21, 2011, Pretoria, South Africa.
14. Neuman, S.P., L. Xue, **M. Ye**, and D. Lu (2011), Multimodel Assessment of the Worth of Data Under Uncertainty, Waste Management Symposium, February 28 – March 3, 2011, Phoenix, Arizona.
15. **Ye, M.**, D. Lu, G.P. Curtis, P.D. Meyer, and S. Yabusaki (2010), Effect of Temporal Residual Correlation on Estimation of Model Averaging Weights, AGU Fall Meeting, December 13-17, San Francisco, California.
16. Lu, D., M.C. Hill, and M. Ye (2010), A Controlled Experiment for Investigating Uncertainty Measures in Groundwater Flow Modeling, AGU Fall Meeting, December 13-17, San Francisco, California.
17. Miller, G.L., D. Lu, **M. Ye**, G.P. Curtis, B.S. Mendes, and D. Draper (2010), Assessment of Parametric Uncertainty using Markov Chain Monte Carlo Methods for Surface

- Complexation Models in Groundwater Reactive Transport Modeling, AGU Fall Meeting, December 13-17, San Francisco, California.
18. Neuman, S.P., **M. Ye**, L. Xue, D. Lu (2010), Multimodel Bayesian Analysis of the Worth of Data, AGU Fall Meeting, December 13-17, San Francisco, California.
 19. Curtis, G.P., **M. Ye**, M. Kohler, P.M. Fox, and J.A. Davis (2010), Evaluating Prediction Uncertainty of Uranium Transport in Small Scale Tracer Tests, AGU Fall Meeting, December 13-17, San Francisco, California.
 20. **Ye, M.**, D. Lu, G. Miller, G.P. Curtis, P.D. Meyer, and S.B. Yabusaki (2010), Assessment of predictive uncertainty in coupled groundwater reactive transport modeling, Goldschmidt 2010 Conference, June 13-18, Knoxville, Tennessee.
 21. **Ye, M.** (2010), Groundwater reactive transport modeling under uncertainty, Joint Mathematics Meeting, January 13 – 16, 2010, San Francisco, California.
 22. **Ye, M.** and D. Lu (2009), On Model Selection Criteria and Model Complexity, 2009 AGU annual meeting, San Francisco, California.
 23. Neuman, S.P. and **M. Ye** (2009), Assessing and optimizing the worth of information under model, parameter and data uncertainties, 2009 AGU annual meeting, San Francisco, California.
 24. **Ye, M.**, G.P. Curtis, P.D. Meyer, and S.B. Yabusaki (2009), Multimodel analysis and averaging using UCODE_2005 and MMA with an application of assessing geochemical model uncertainty at the Naturita Site, 2009 ISCEMEM Annual meeting, October 13 – 15, Rockville, Maryland.
 25. **Ye, M.** (2009), Groundwater model calibration and uncertainty quantification, 2009 SIAM Annual Meeting, July 6 – 10, Denver, Colorado.
 26. Curtis, G.P., **M. Ye**, P.D. Meyer, S.B. Yabusaki, and D.R. Rodriguez (2009), Multiscale assessment of prediction uncertainty in multicomponent reactive transport models, DOE-ERSP 4th Annual PI Meeting, April 20 – 23, Lansdowne, Virginia.

References:

- Kohler, M., G.P. Curtis, D.B. Kent, and J.A. Davis (1996), Experimental investigation and modeling of uranium (VI) transport under variable chemical conditions, *Water Resources Research*, 32(12), 3539-3551.
- Poeter, E. P., and M. C. Hill (2007), MMA: A Computer Code for Multi-Model Analysis, U.S. Geol. Surv. Tech. Methods, TM6-E3, 113 pp.
- Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum Likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, 40, W05113, doi:10.1029/2003WR002557.
- Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resour. Res.*, 44, W03428, doi:10.1029/2008WR006803.

Appendix A

This appendix contains the manuscript of Shi et al. (2013), which is under revision for publication in Water Resources Research. This research helps answer the two questions below explained in detail in the report:

Question 1: How does uncertainty behave and affect simulations of groundwater reactive transport models?

Question 2: What cause the uncertainty in groundwater reactive transport modeling?

1
2
3 **Assessment of Parametric Uncertainty for Surface Complexation Modeling**
4 **of Groundwater Uranium Reactive Transport**
5

6 Xiaoqing Shi^{1,2}, Ming Ye^{2,*}, Gary P. Curtis³, Geoffery L. Miller²,
7 Philip D. Meyer⁴, Matthias Kohler³, Steve Yabusaki⁴, and Jichun Wu¹
8

9 ¹School of Earth Sciences and Engineering, Nanjing University, Nanjing 210093, China

10 ²Department of Scientific Computing, Florida State University, Tallahassee, FL 32306, USA

11 ³U.S. Geological Survey, Menlo Park, CA 94025, USA

12 ⁴Pacific Northwest National Laboratory, Richland, WA 99352, USA
13

14 *Corresponding author: mye@fsu.edu, Phone: 850-644-4587, Fax: 850-644-0098
15
16

Abstract

The validity of using Gaussian assumptions for model residuals in uncertainty quantification of a groundwater reactive transport model was evaluated in this study. The least-square regression methods explicitly assume Gaussian residuals, and the assumption leads to Gaussian likelihood functions, model parameters, and model predictions. While the Bayesian methods do not explicitly require the Gaussian assumption, Gaussian residuals are widely used. This paper shows that the residuals of the reactive transport model are non-Gaussian, heteroscedastic, and correlated in time; characterizing them requires using the formal generalized likelihood function developed by Schoups and Vrugt [2010]. When the formal generalized likelihood function is used in the Bayesian uncertainty quantification, the resulting parameter distributions are non-Gaussian, and the corresponding predictive performance is significantly improved in comparison with predictive performance of least-square regression and Bayesian methods that use Gaussian likelihood function. The differential evolution adaptive metropolis (DREAM_(ZS)) algorithm of Markov chain Monte Carlo (MCMC) simulations is a robust tool for quantifying uncertainty in groundwater reactive transport models. Using the formal generalized likelihood function in DREAM_(ZS) is computationally more expensive than using the Gaussian likelihood function. For the groundwater reactive transport model considered in this study, the parametric uncertainty is caused by competing geochemical reactions and a substitution effect of adsorption between the reactions. Although this study was conducted as a numerical experiment, the uncertainty analysis may help select appropriate likelihood functions, improve model calibration, and reduce predictive uncertainty in other groundwater reactive transport and environmental modeling in a broader sense.

1. Introduction

Groundwater reactive transport modeling is a vital tool for analyzing and managing the subsurface environment [Davis et al., 2004a; Steefel et al., 2005; Scheibe et al., 2008]. Because model predictions are inherently uncertain, accurate quantification of predictive uncertainty is necessary to avoid poorly managed and overly costly remediation and monitoring. Predictive uncertainty is caused by propagation of various sources of uncertainty in model structures, parameters, driving forces, and data used in modeling [Neuman, 2003; Wagner and Gupta, 2005; Srinivasan et al., 2007; Meyer et al., 2007; Clark et al., 2011; Refsgaard et al., 2012; Gupta et al., 2012; Tartakovsky, 2013]. Parametric uncertainty is due to imperfect knowledge of model parameters, initial and boundary conditions, and/or driving forces. Model structure uncertainty may be manifested by different plausible conceptualizations and/or mathematical descriptions of the transport processes (e.g., the traditional advection-dispersion model versus other alternative models, as discussed in Srinivasan et al. [2007] and Tang et al. [2009]) and the geochemical reaction processes (e.g., different formulations of surface complexation models developed for simulating uranium adsorption in batch experiments [Davis et al., 2004b], column experiments [Kohler et al., 1996], and tracer experiments [Davis et al., 2004b; Curtis and Davis, 2006]). Quantification of model uncertainty can be pursued using either model selection methods [e.g., Kohler et al., 1996; Davis et al., 2004b; Matott and Rabideau, 2008a] or model averaging methods [e.g., Neuman, 2003; Poeter and Anderson, 2005; Ye et al., 2004, 2008, 2010a,b]. These methods require quantifying parametric uncertainty so that model selection or model averaging is conducted with consideration of parametric uncertainty, instead of relying on a single parameter value or a small set of discrete parameter values.

This study was focused on quantification of parametric uncertainty in groundwater reactive transport modeling, with particular attention paid to uranium reactive transport

modeling using surface complexation models. In comparison with parametric uncertainty analysis in groundwater flow and nonreactive transport modeling [see review articles of Neuman, 1973; Neuman and Yakowitz, 1979; Yeh and Yoon, 1981; Yeh, 1986; Kitanidis, 1986; Beck, 1987; Carrera, 1993; Liu and Gupta, 2007; Matott et al., 2009], quantification of parametric uncertainty in reactive transport modeling has its own challenges [Leavitt et al., 2011]. In particular, reactive transport models are nonlinear with respect to the parameters due to nonlinear reaction equations and coupling of reactions and biogeochemical processes [Cabaniss, 1999; Mugunthan et al., 2005]. The nonlinearities lead to complex surfaces of objective functions (e.g., least squares or likelihood) with multiple local minima, as shown in Matott and Rabideau [2008b]. For such objective function surfaces, gradient-based parameter estimation algorithms (e.g., the Gauss-Levenberg-Marquardt method implemented in UCODE_2005 [Poeter et al., 2005] and PEST [Doherty, 2010]) may perform poorly, for example, being terminated at a local minimum. In addition, due to the nonlinearities, the probability distributions of model parameters and outputs may be non-Gaussian with multiple modes and/or long tails [Cabaniss, 1997; Denison and Carnier-Laplace, 2005; Leavitt et al., 2011]. This questions accuracy of uncertainty quantification methods that either explicitly [Srinivasan et al., 2007; Liu et al., 2008; Tartakovsky et al., 2009] or implicitly [Dai and Samper, 2004] assume Gaussian parameters of groundwater reactive transport models. A literature review of Lu et al. [2012] showed that the regression confidence intervals based on the Gaussian assumption of model parameters may differ from the Bayesian credible intervals that do not require the Gaussian assumption. However, the numerical studies of Lu et al. [2012] and Shi et al. [2012] for flow models showed that discrepancy between the two kinds of intervals is small. For groundwater reactive transport models, it is unknown to what extent the Gaussian assumption affects accuracy of uncertainty quantification, and this was one problem tackled in this study.

The Markov Chain Monte Carlo (MCMC) methods may be a solution to the above challenges of model nonlinearity and parameter non-Gaussianity [e.g., Marshall et al., 2004; Gallagher and Doherty, 2007; Smith and Marshall, 2008; Vrugt et al., 2008a, 2009a,b]. As a Bayesian technique, MCMC can be directly applied to nonlinear models without any approximation of model linearization, and it does not require any assumptions on the form of the parameter probability distributions. Instead, it estimates the posterior parameter distributions and takes into account model nonlinearity in the estimation. Smith and Marshall [2008] demonstrated that MCMC techniques are able to identify multimodal parameter distributions of hydrologic models. Building on the Differential Evolution-Markov Chain (DE-MC) method of ter Braak [2006], Vrugt et al. [2008a, 2009a] developed the Differential Evolution Adaptive Metropolis (DREAM) algorithm; its efficiency in estimating multimodal non-Gaussian parameter distributions was demonstrated in Vrugt et al. [2009b] for synthetic simple test problems. While DREAM appears to be a promising tool, it is unknown whether DREAM can resolve the above challenges in groundwater reactive transport modeling, because it has not been used in reactive transport modeling. The application of DREAM to groundwater flow and transport modeling in Keating et al. [2010] suggested that DREAM may not be practical to high-dimensional and computationally expensive models, because it requires tens to hundreds of thousands of model runs. While this problem was resolved in Laloy and Vrugt [2012] using the multi-try DREAM_(ZS) algorithm, applications of DREAM to groundwater modeling, especially reactive transport modeling, are still limited.

Another challenge to quantification of parametric uncertainty is the selection of likelihood functions of parameters or equivalently probability distributions of residuals (difference between observations and corresponding model simulations). While the likelihood functions are used directly in Bayesian uncertainty quantification or as the objective functions in maximum likelihood and regression theories, exact expressions of likelihood functions are

unknown. Beven [2008] advocated using informal likelihood functions (also known as generalized likelihood function or likelihood measures) that are not necessarily related to the residual probabilities. Rubin et al. [2010] and Over et al. [2013] used a non-parametric likelihood function in the method of anchored distributions developed to quantify model and parametric uncertainty. Zhang et al. [2013] used informal exponential likelihood function in their Bayesian uncertainty analysis for groundwater reactive transport modeling. Formal likelihood functions are used more widely than informal ones, and Gaussian likelihood function is the most popular one. However, it may not be always appropriate to characterize the residual probability, and other likelihood functions are needed. For example, Chen et al. [2010] used t distribution because its heavier tails are more robust than Gaussian distribution to fit residuals of surface seismic refraction data. Schoups and Vrugt (2010) developed the formal generalized likelihood function that uses the skew exponential power (SEP) distribution, and it is probably the most inclusive formal likelihood function to describe residuals that are correlated, heteroscedastic, and non-Gaussian. While the formal generalized likelihood function has been used for hydrologic models, it is unknown whether it is appropriate to characterize residual distributions of groundwater reactive transport models, which was explored in this study.

To address the above unresolved questions in parametric uncertainty quantification for groundwater reactive transport modeling, a numerical study was conducted for a Surface Complexation Model (SCM) developed by Kohler et al. [1996] to simulate uranium reactive transport in column experiments. The SCM is ideal for this numerical exploration and comparative study for the following two reasons: (1) the SCM has only four unknown parameters and is thus not high-dimensional, and (2) the SCM is not computational expensive because each forward model run takes about 3 - 5 minutes using a single processor. The following three kinds of parametric uncertainty analysis were performed:

- (1) Regression-based uncertainty quantification implemented in UCODE_2005. The quantification uses approximate linearization of nonlinear models and assumes Gaussian residuals; the Gaussian assumption leads to asymptotic Gaussian distributions of model parameters and predictions.
- (2) Bayesian uncertainty quantification implemented in DREAM. The quantification does not require either linearizing nonlinear models or Gaussian parameters, but uses Gaussian likelihood function.
- (3) The same Bayesian uncertainty quantification of (2) but uses the formal generalized likelihood function (not the Gaussian likelihood function).

The results of the three kinds of uncertainty analysis were compared to investigate:

- (1) Whether the parameter distributions of the three methods are different. As shown below, both theoretical analysis and our numerical results indicate that, if the Gaussian likelihood function is used and the parameter distributions of regression and Bayesian methods are uni-modal, the distributions are similar.
- (2) Whether the confidence intervals of regression methods and the credible intervals of the Bayesian methods (using the two different likelihood functions) are different. A cross-validation study was conducted to evaluate predictive performance of the confidence and credible intervals. As shown below, the credible interval based on the formal generalized likelihood function outperforms the confidence and credible intervals based on the Gaussian likelihood function.

The rest of the paper is organized as follows. Section 2 provides necessary details about the theories of nonlinear regression and Bayesian uncertainty analysis. Section 3 describes the SCM, the local optimization and uncertainty quantification using UCODE_2005, and analysis of the residuals of UCODE_2005, followed by MCMC simulation, its convergence, and comparison of regression-based and Bayesian quantification of parametric and predictive

uncertainty. In Section 4, the MCMC results are used to understand sensitivity of the SCM parameters and geochemical reasons of parametric uncertainty. Major findings of this study are summarized in Section 5.

2. Regression and Bayesian Methods for Parametric Uncertainty Analysis

A brief overview of the regression and Bayesian techniques is given in this section to make this paper self-contained; detailed discussions of the methods are referred to Draper and Smith [1981], Hill and Tiedeman [2007], Box and Tiao [1992], Gelman and Rubin [1992], and Draper [2007]. A thorough comparison (theoretical and numerical) of the confidence intervals of nonlinear regression methods and credible intervals of Bayesian approaches in the context of groundwater flow modeling is referred to Lu et al. [2012].

2.1 Nonlinear regression methods

The statistical model used for regression-based nonlinear model calibration and uncertainty quantification can be expressed as

$$\mathbf{y} = f(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \quad (1)$$

where $\mathbf{y}$ is a vector of n observations, f is a nonlinear model with respect to its parameters, $\boldsymbol{\beta}$ (a vector of p model parameters), and $\boldsymbol{\varepsilon}$ is a vector of n errors. The errors are random and assumed to follow a multivariate normal distribution, $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{C}_\varepsilon)$. The variance-covariance matrix, $\mathbf{C}_\varepsilon$, is often represented as $\mathbf{C}_\varepsilon = \sigma^2 \boldsymbol{\omega}^{-1}$, where $\boldsymbol{\omega}$ is a $n \times n$ known weight matrix and σ^2 is a scalar, possibly unknown but can be estimated. A common practice in nonlinear regression is to linearize the model by expanding it in a Taylor series and retaining the first two terms, i.e., $f(\hat{\mathbf{b}}) \approx f(\boldsymbol{\beta}) + \mathbf{X}_\beta(\hat{\mathbf{b}} - \boldsymbol{\beta})$, where $\hat{\mathbf{b}}$ is the estimate of $\boldsymbol{\beta}$ obtained by minimizing the generalized least-squares objective function,

$$S(\mathbf{b}) = [\mathbf{y} - f(\mathbf{b})]^T \boldsymbol{\omega} [\mathbf{y} - f(\mathbf{b})], \quad (2)$$

188 and $\mathbf{X}_\beta = \left[\frac{\partial f}{\partial \mathbf{b}} \right]_{\mathbf{b}=\beta}$ is a $n \times p$ sensitivity matrix [Seber and Wild, 2003, p.23-24]. The objective
 189 function is also referred to as sum of squared weighted residuals (*SSWR*), where residuals are
 190 defined as the difference between observed and corresponding simulated values. The
 191 parameter estimates, $\hat{\mathbf{b}}$, follow asymptotically the multivariate normal distribution [Seber
 192 and Wild, 2003, p.24; Hill and Tiedeman, 2007, p.398],

$$193 \quad \hat{\mathbf{b}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}_\beta^T \boldsymbol{\omega} \mathbf{X}_\beta)^{-1}), \quad (3)$$

194 because of the normality assumption of the errors. If the nonlinear prediction function $g(\hat{\mathbf{b}})$
 195 is also approximated to the first order by $g(\hat{\mathbf{b}}) \approx g(\boldsymbol{\beta}) + \mathbf{Z}_\beta^T (\hat{\mathbf{b}} - \boldsymbol{\beta})$, then $g(\hat{\mathbf{b}})$ follows
 196 asymptotically the normal distribution

$$197 \quad g(\hat{\mathbf{b}}) \sim N(g(\boldsymbol{\beta}), \sigma^2 \mathbf{Z}_\beta^T (\mathbf{X}_\beta^T \boldsymbol{\omega} \mathbf{X}_\beta)^{-1} \mathbf{Z}_\beta), \quad (4)$$

198 where $\mathbf{Z}_\beta$ is the prediction sensitivity vector $\mathbf{Z}_\beta = \left[\frac{\partial g}{\partial \boldsymbol{\beta}} \right]_{\boldsymbol{\beta}=\hat{\mathbf{b}}}$ [Seber and Wild, 2003, p.192].

199 Thus, the $(1-\alpha) \times 100\%$ linear confidence interval of $g(\boldsymbol{\beta})$ is

$$200 \quad g(\hat{\mathbf{b}}) \pm t_{1-\alpha/2}(n-p) [s^2 \mathbf{Z}_\beta^T (\mathbf{X}_\beta^T \boldsymbol{\omega} \mathbf{X}_\beta)^{-1} \mathbf{Z}_\beta]^{1/2}. \quad (5)$$

201 where $t_{1-\alpha/2}(n-p)$ is a t statistic with significance level α and degrees of freedom $n-p$, and

$$202 \quad s^2 = (\mathbf{y} - f(\hat{\mathbf{b}}))^T \boldsymbol{\omega} (\mathbf{y} - f(\hat{\mathbf{b}})) / (n-p) \quad (6)$$

203 is the estimate of σ^2 . In practice, the sensitivity matrices $\mathbf{X}_\beta = \left[\frac{\partial f}{\partial \mathbf{b}} \right]_{\mathbf{b}=\beta}$ and $\mathbf{Z}_\beta = \left[\frac{\partial g}{\partial \mathbf{b}} \right]_{\mathbf{b}=\beta}$

204 are approximated by $\hat{\mathbf{X}}_{\hat{\mathbf{b}}} = \left[\frac{\partial f}{\partial \mathbf{b}} \right]_{\mathbf{b}=\hat{\mathbf{b}}}$ and $\hat{\mathbf{Z}}_{\hat{\mathbf{b}}} = \left[\frac{\partial g}{\partial \mathbf{b}} \right]_{\mathbf{b}=\hat{\mathbf{b}}}$, respectively, with $\boldsymbol{\beta}$ replaced by its

205 estimate, $\hat{\mathbf{b}}$, obtained from the nonlinear regression [Seber and Wild, 2003, p.191].

206 Estimating the *nonlinear* confidence intervals does not require linearizing the model. The
 207 intervals are determined as the minimum and maximum of model predictions intersecting

with a confidence region of model parameters [Vecchia and Cooley, 1987]. The region is defined as the set of parameter values whose corresponding objective function values satisfy [Christensen and Cooley, 1999; Cooley, 2004]

$$S(\mathbf{b}) \leq S(\hat{\mathbf{b}}) \left[\frac{1}{n-p} t_{\alpha/2}^2 (n-p) + 1 \right]. \quad (7)$$

This parameter region contains the true model parameter with approximate probability of $(1-\alpha) \times 100\%$ for the errors, ϵ , defined in equation (1). Estimating the nonlinear confidence interval requires the following assumptions: (1) the model accurately represents the system, (2) model predictions, $g(\boldsymbol{\beta})$, are sufficiently monotonic, (3) there is a single minimum in the objective function, and (4) the residuals are multivariate normal distributed to obtain a valid critical value, and (5) model intrinsic nonlinearity is small [Cooley and Naff, 1990; Hill and Tiedeman, 2007; Lu et al., 2012]. If the assumptions are not satisfied, equation (7) may not define the objective function value associated with the designated $1-\alpha$ confidence level and the nonlinear intervals may be inaccurate.

Calculation of the linear and nonlinear confidence intervals relies on $\hat{\mathbf{b}}$ obtained by minimizing equation (2). The optimization is conventionally conducted in groundwater modeling using the Gauss-Marquardt-Levenberg (GML) method, which has been implemented in several popular codes of automated calibration including PEST [Doherty, 2005], UCODE_2005 [Poeter et al, 2005], iTOUGH2 [Finsteler, 2007; Finsteler and Zhang, 2011]. However, the gradient-based GML method may not be able to find the global minimum of the objective function; it may be terminated in regions of local minima [Chen et al., 2008]. Whether the UCODE-2005 parameter estimates are global optima was investigated in this study using global optimization methods such as DREAM_(ZS) and other techniques discussed in Section 3.

2.2 Bayesian method

Different from the nonlinear regression methods that treat model parameters as deterministic variables but parameter estimates as random variables, the Bayesian methods treat model parameters themselves as random variables and characterize parametric uncertainty using their posterior distribution obtained from the Bayes' theorem,

$$p(\boldsymbol{\beta} | \mathbf{y}) = \frac{p(\mathbf{y} | \boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{y})} = \frac{p(\mathbf{y} | \boldsymbol{\beta})p(\boldsymbol{\beta})}{\int p(\mathbf{y} | \boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}}, \quad (8)$$

where $p(\boldsymbol{\beta} | \mathbf{y})$ is the posterior parameter distribution conditioned on data $\mathbf{y}$, $p(\boldsymbol{\beta})$ is prior distribution, and $p(\mathbf{y} | \boldsymbol{\beta})$ is the likelihood function that will be discussed in detail in Section 2.3. The posterior distributions are always obtained numerically using MCMC techniques, which do not require linearizing the models and are able to infer non-Gaussian posterior distributions. After the posterior parameter distributions are obtained, the predictive uncertainty is quantified by first drawing parameter samples from the distribution and then executing the prediction model with the samples. The $(1 - \alpha) \times 100\%$ credible interval is determined via

$$\int_l^u p(g(\boldsymbol{\beta}) | \mathbf{y}) dg(\boldsymbol{\beta}) = 1 - \alpha, \quad (9)$$

where $p(g(\boldsymbol{\beta}) | \mathbf{y})$ is posterior distribution of $g(\boldsymbol{\beta})$ conditioned on data $\mathbf{y}$. In this study, the credible interval limits, l and u , are determined using the equal-tailed method as [Casella and Berger, 2002]

$$p(g(\boldsymbol{\beta}) \leq l | \mathbf{y}) = p(g(\boldsymbol{\beta}) \geq u | \mathbf{y}) = \alpha / 2. \quad (10)$$

If $\alpha = 0.5$, the limits correspond to the 2.5% and 97.5% percentiles of model predictions. Other methods of estimating the credible intervals (e.g., highest posterior density interval) are also available [Box and Tiao, 1992; Chen and Shao, 1999; Casella and Berger, 2002], but were not used in this work.

Efficiently and accurately estimating $p(\mathbf{y} | \boldsymbol{\beta})$ is a daunting challenge to MCMC techniques,

especially when $p(\mathbf{y}|\boldsymbol{\beta})$ is multimodal. Recently, Vrugt et al. [2008a, 2009a,b] proposed the DREAM_(ZS) algorithm, which runs multiple Markov chains from different starting points in parallel for global search in the parameter space, which makes it possible to search multiple parameter regions corresponding to multiple modes. This is explained quantitatively by the way of DREAM_(ZS) generating the candidate parameter point. Different from other MCMC techniques that generate candidate points using covariance matrix or the adaptive covariance, DREAM_(ZS) generates candidate points by directly using the current location of the multiple chains via [Vrugt et al., 2009b]

$$\mathbf{z}^i = \mathbf{x}_{t-1}^i + \gamma(\mathbf{x}_{t-1}^{r_1} - \mathbf{x}_{t-1}^{r_2}) + \mathbf{e} \quad r_1 \neq r_2 \neq i \quad (11)$$

where $\mathbf{z}^i$ is the candidate point, $\mathbf{x}^i$, $\mathbf{x}^{r_1}$, and $\mathbf{x}^{r_2}$ are different Markov chains, γ is a scale factor, and $\mathbf{e}$ is random error following normal distribution with small variance. Using equation (11) increases the possibility of direct jumps between different modes. In addition, the scale (γ) and orientation of the proposal distribution ($\mathbf{x}_{t-1}^{r_1} - \mathbf{x}_{t-1}^{r_2}$) are automatically tuned during the evolution to the posterior distribution. Self-adaptive randomized subspace sampling is used to further improve efficiency of the sampling. More details of DREAM_(ZS) are referred to Vrugt et al. [2008a, 2009a,b].

2.3 Gaussian and formal generalized likelihood functions

The Bayesian methods can embrace any likelihood function, and a widely used one is the Gaussian likelihood function,

$$p(\mathbf{y} | \boldsymbol{\beta}) = \frac{1}{(2\pi)^{p/2} |\mathbf{C}|} \exp \left[-\frac{1}{2} (\mathbf{y} - f(\boldsymbol{\beta}))^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - f(\boldsymbol{\beta})) \right], \quad (12)$$

where $\boldsymbol{\Sigma}$ is the covariance function of residuals, $\mathbf{r} = \mathbf{y} - f(\boldsymbol{\beta})$. Analytical expressions of the posterior distribution are in general not available except for special cases (e.g., Woodbury and Ulrich, 2000; Hou and Rubin, 2005). For the Gaussian likelihood, Box and Tiao [1992, p. 428]

showed that, when using non-informative prior distributions of β and Σ , the posterior parameter distribution follows

$$p(\beta | y) \propto [S(\beta)]^{-n/2}, \quad (13)$$

where n is number of observations, and

$$S(\beta) = (y - f(\beta))^T (y - f(\beta)), \quad (14)$$

is the sum of squared residuals (SSR), the objective function of ordinary least square methods.

Equations (13) and (14) are the connection between the shape of posterior distribution and the surface of SSR objective function. For example, multiple minima on a least-square objective function may correspond to multiple modes of the parameter density function. This is illustrated in Figure 1 for a single parameter following bimodal distribution (Figure 1a). As shown in Figure 1b, parameter samples corresponding to the high peak of the distribution may produce model simulations $f(\beta)$ close to observation, y ; model simulations produced by samples corresponding to the low peak deviate more from the observation; model simulations corresponding to other parts of the distribution deviate even more from the observation. As a result, Figure 1c shows that the objective function has two minima, the global minimum corresponding to the high peak and the local minimum to the low peak. In line with this, if the probability density function of the parameter has multiple modes, the objective function will have multiple minima. The relations between the modes of parameter distributions and minima of objective functions are proved by Box and Tiao [1992, p.489] for linear models. Although such a proof does not exist for nonlinear models, if multiple minima are observed during regression-based model calibration, it is an indicator that model parameters may be non-Gaussian.

The formal generalized likelihood function of Schoups and Vrugt [2010] was considered in this study, because it explicitly considers heteroscedasticity, temporal correlation, and

non-Gaussianity in residuals that were found in this study (see Section 3.3). The formal generalized likelihood function characterizes the residual, r_t , using the following model [Schoups and Vrugt, 2010]:

$$\Phi_p(B)r_t = \sigma_t a_t \text{ with } a_t \sim SEP(0,1,\xi,\theta) \quad (15)$$

where $\Phi_p(B) = 1 - \sum_{i=1}^p \phi_i B^i$ is an autoregressive polynomial with p autoregressive parameter, ϕ_i , B is the backshift operator, $B^i r_t = r_{t-i}$, σ_t is standard deviation at time t , and a_t is an i.i.d. random error with zero mean and unit standard deviation, described by a skew exponential power (SEP) density defined below with parameters ξ and θ to account for non-Gaussianity. Following Schoups and Vrugt [2010], heteroscedasticity was explicitly accounted for using the linear model

$$\sigma_t = \sigma_0 + \sigma_1 E_t \quad (16)$$

where E_t is the mean simulation, and the coefficients, σ_0 and σ_1 , are determined as nuisance parameters during the MCMC simulation. Residual non-Gaussianity was accounted for using the $SEP(0,1,\xi,\theta)$ density function [Schoups and Vrugt, 2010]

$$p(a_t | \xi, \theta) = \frac{2\sigma_\xi}{\xi + \xi^{-1}} \omega_\xi \exp \left[-c_\theta |a_{\xi,t}|^{2/(1+\theta)} \right] \quad (17)$$

where $a_{\xi,t} = \xi^{-\text{sign}(\mu_\xi + \sigma_\xi a_t)} (\mu_\xi + \sigma_\xi a_t)$, and values for the nuisance parameters, μ_ξ , σ_ξ , c_θ , and ω_θ , are computed as a function of skewness parameter, ξ , and kurtosis parameter, θ , as described in Schoups and Vrugt [2010]. The MCMC simulation was conducted using the DREAM_(zs) code developed by Vrugt [2009] that includes the formal generalized likelihood function as a choice.

3. Results

This section starts from introducing the surface complexation model (SCM) and its

uncertainty and nonlinearity in Section 3.1. Parameter estimation and parameter uncertainty analysis conducted using UCODE_2005 are discussed in Section 3.2, followed by the Bayesian parametric uncertainty analysis conducted using DREAM_(ZS) based on the Gaussian and formal generalized likelihood functions in Section 3.3. The results of uncertainty quantification (e.g., the 95% linear and nonlinear confidence intervals and the 95% credible intervals) are compared in Section 3.4 for two cases of cross-validation to evaluate accuracy and precision of the intervals.

3.1. Geochemical model, parametric uncertainty, and model nonlinearity

In order to study uranium transport and test potential applicability of surface complexation modeling, Kohler et al. [1996] conducted eight column experiments in a well-characterized U(VI)-quartz-fluoride column system. The breakthrough curves of U(VI) exiting the column over the course of several pore volumes of water showed the retardation effect due to uranium adsorption on the quartz surface. The uranium adsorption was simulated in Kohler et al. [1996] using the SCMs, in which uranium is absorbed on surface hydroxyl functional groups according to chemical reactions, for example,

where S_jOH represents a surface hydroxyl functional group. The formation constant, K , of (18) is defined as

$$K_j = (S_jOUO_2OH)(H^+)^2 / (S_jOH)(UO_2^{2+}) \quad (19)$$

where the quantities in parenthesis denote the activity of each species (the activity coefficients of the surface species are assumed to be equal to one). A SCM may involve several functional groups with different adsorption affinity measured by K . The number of functional groups that participate in the adsorption reaction and the adsorption affinity of these groups are unknown because of a wide range in bonding environments on the mineral surface. This is significant

contribution to model uncertainty for the SCMs. The distribution in bonding affinity is often approximated by two or three discrete site types [Davis and Kent, 1990]. Kohler et al. [1996] postulated seven alternative SCMs with different numbers of surface functional groups and different reaction stoichiometry. For each of the seven models, the formation constants of each reaction and the fractions of the reaction surface of each functional group were unknown, which is the reason of parametric uncertainty. This study is focused on quantification of parametric uncertainty of an individual model, which is the basis of quantifying model uncertainty in future study.

The assessment of parametric uncertainty was conducted for Model C4 selected by Kohler et al. [1996] as the best model based on a qualitative analysis and the principle of parsimony. This model has two functional groups, S₁OH and S₂OH, hereinafter referred to as weak site and strong site, respectively. The weak site is associated with one reaction

and the strong site with two reactions

Model calibration and uncertainty analysis was performed for the 10-base logarithms of the formation constants of the three reactions (denoted as logK1, logK2, and logK3) and the 10-base logarithm of the fraction of the strong site (denoted as logSite). The fraction of the weak site is not considered explicitly, because the summation of the two site fractions is one. The four calibrated parameters played different roles to determine the shape of the simulated breakthrough curves. logK1, the formation constant for the equilibrium reaction associated with the weak site, strongly influences the climbing limb of the breakthrough curve and the center of mass. The decreasing limb of the breakthrough curve is most sensitive to logK2 and

logK3, the formation constants of the equilibrium reactions associated with the strong site; surface fraction of the strong site is measured by logSite. The parameters were subject to the physical constraints that $\log K1 < \log K2$ and $\log K3 > \log K1 + 3.9$. These bounds were used when sampling parameter realizations for MCMC simulations. Nonlinearity of this model was evaluated using the total nonlinearity and intrinsic nonlinearity measures calculated by UCODE_2005. According to Hill and Tiedeman [2007, p.142-145], a model is effectively linear when these measures are less than 0.09, and highly nonlinear when they are greater than 1.0. For this SCM, the total nonlinearity and intrinsic nonlinearity measures were 0.5 and 0.3, indicating that the model is moderately nonlinear.

3.2. Local optimization using UCODE_2005

The model calibration was conducted by matching simulated uranium concentrations to a total of 120 observations from three experiments denoted as Experiments 1, 2, and 8 with different experimental conditions [Kohler et al., 1996]. By taking $\mathbf{C}_\varepsilon = \sigma^2 \boldsymbol{\omega}^{-1}$ and assuming that the measurement error of each observation is independent, the objective function was

$$SSWR = \sum_{j=1}^{Ngroup} W_j SSWR_j = \sum_{j=1}^{Ngroup} \left[\frac{N}{N_{obs,j}} \sum_{i=1}^{Nobs,j} \frac{\omega_{ji} (Y_{ji} - \hat{Y}_{ji})^2}{\sigma^2} \right] \quad (23)$$

where $Ngroup = 3$ is the number of experiments, $SSWR_j$ is the objective function of each experiment, $W_j = N / N_{obs,j}$ is used to balance different numbers of observations in the individual experiments ($N = 120$ being the total number of observations and $N_{obs,j}$ being the number of observations for the j -th experiment, 39, 32, and 49 for Experiments 1, 2, and 8, respectively), ω_{ji} is the inverse of the variance of measurement error for the i -th observation of the j -th experiment, σ^2 is the error variance that can be estimated via equation (6), and Y_{ji} and $\hat{Y}_{ji}$ are observed and simulated concentration, respectively. The computer code RATEQ

[Curtis, 2005] was used for the forward model simulation. The forward model execution was the same to that of Kohler et al [1996] except that new thermodynamic data of Guillaumont et al. [2003] were used in this study. A complete simulation of the three experiments took 3 – 5 minutes on the supercomputer of the Florida State University.

UCODE_2005 was used for the local optimization to minimize the *SSWR* of equation (23). The initial parameter values are listed in Table 1, and they were close to those obtained by Kohler et al. [1996] who used a now outdated set of thermodynamic data. The parameter ranges needed for the local optimization were determined empirically around the initial values from the geochemical point of view. The parameters estimates using UCODE_2005 are listed in Table 1, and they are almost identical to those obtained using the global optimization methods of Shuffled complex evolution (SCE-UA) [Duan et al., 1992, 1993, 1994] and Multi-start Broyden-Fletcher-Goldfarb-Shanno (BFGS) method implemented in DAKOTA [Adams et al., 2010] (Table 1). As a result, the parameter set corresponding to local-UCODE is defined as “global optimization parameter” and used hereinafter in the regression-based uncertainty quantification below to evaluate the linear and nonlinear confidence intervals.

It should be noted that finding the global optimum by UCODE_2005 depends on the initial values (Table 1). Figure 2 plots the $-\log_{10}SSWR$ surface obtained from 180,000 models runs by varying the three most influential parameters ($\log K_1$, $\log Site$, and $\log K_2$) identified in the sensitivity analysis discussed below (the fourth parameter, $\log K_3$, was fixed at its optimum value). The figure shows a mixture of closely grouped peaks and valleys. For a gradient-based algorithm, if the initial parameter guesses are far from the global optimum, they are likely trapped at local optima, as observed in the results of the multi-start BFGS (not shown). The complex response surface also poses challenges to global optimization methods. Without fine tuning of certain parameters of SCE-UA and multi-start BFGS, they cannot find the global optima listed in Table 1.

3.3. Residuals of UCODE_2005

Figure 3 plots the observed and simulated breakthrough curves for the three experiments (Figures 2, a-1 – a-3) and the weighted observations and simulations (Figure 2b). While the two figures suggest that the model-fit of UCODE_2005 is acceptable, Gaussian likelihood function may not be accurate to characterize the residuals. Figure 2c shows that the residuals are on the order of 10^{-1} and significantly larger than the measurement errors that are on the order of 10^{-3} , indicating that model error dominates over measurement error. The residual variance in general increases with the simulated concentrations, a sign of heteroscedasticity. The residual histogram (Figure 2d) has a sharp peak that cannot be characterized by the assumed Gaussian density function. In addition, the residuals are correlated in time, because the partial autocorrelation function (PACF) at lags 2-4 is beyond the 95% confidence interval, as shown in Figure 2e for Experiment 1 as an example (the temporal correlation was observed for all the three experiments). The formal generalized likelihood function is sufficient to characterize these three characteristics of heteroscedasticity, peaked density function, and temporal correlation.

3.4. MCMC simulation using Gaussian and formal generalized likelihood functions

The MCMC simulations for the Gaussian and formal generalized likelihood functions were conducted by running three chains in parallel. Uniform distributions were used as the prior parameter distributions, and the lower and upper bounds of the four model parameters are listed in Table 1. Convergence of the MCMC simulation was monitored using the scale reduction factor ($\hat{R}$) defined by Gelman et al. [2004, p331-333]. According to the rule of thumb that convergence is attained after $\hat{R}$ is less than 1.2. Figure 4 shows that the MCMC simulation using the Gaussian likelihood reached convergence after 3,000 MCMC runs. This is confirmed by the well mixing of the three chains plotted in Figure 5. A total of 15,000

model executions were conducted, and the first 5,000 samples were discarded as the burn-in period. The remaining samples were used for the uncertainty quantification discussed below.

For the MCMC simulation using the formal generalized likelihood function, it is more difficult to reach convergence. In this simulation, in addition to the four parameters of the SCM, the parameter set includes the two variance parameters (σ_0 and σ_1 in equation 16), two shape parameter (ξ and θ in equation 17), and two autocorrelation coefficients (ϕ_1 and ϕ_2 in equation 15). The order of autoregressive model, $AR(p)$, was determined using the method described in Lu et al. [2013]. A total of 75,000 MCMC runs were conducted. Figure 6 shows that all the $\hat{R}$ values became smaller than 1.2 after 43,000 MCMC samples but decreased after 60,000 samples. Therefore, the first 60,000 samples were discarded, and the remaining 15,000 samples were used for the uncertainty quantification discussed below. The larger burn-in period may be attributed to the augmented parameter set, and more investigation is warranted in a future study. Figure 7 of chain evolution shows the difference of burning the first 43,000 and 60,000 samples. When the burn-in period is 43,000 samples, logK3 and logSite can take values of 0.8 and -1.4, respectively; these values disappear when the burn-in period is 60,000 samples. The uncertainty quantification discussed below was based on the more reasonable burn-in period of 60,000 samples.

3.5. Comparison of parametric uncertainty

The parameter distributions obtained from the regression and Bayesian methods are compared for the four model parameters (logK1, logK2, logK3, and logSite). Figure 8 plots the Gaussian density function obtained using the UCODE_2005 results, i.e., the means and variances of the density functions are respectively the parameter estimates and estimation variances evaluated using the regression theories. Figure 8 also plots the histograms of the MCMC samples obtained using the two likelihood functions. The histograms of the MCMC

samples using the Gaussian likelihood function also have the Gaussian shape. The modes of the histograms are almost identical to the means of the Gaussian density functions, i.e., the UCODE_2005 parameter estimates (Table 1). The standard deviations of the MCMC samples are slightly larger than those of the UCODE_2005 results. The similarity indicates that, if Gaussian likelihood function is used, the posterior parameter distributions of Bayesian methods are similar to the Gaussian parameter distributions of the regression methods. While this was approved for linear models (equation (13) adapted from Box and Tiao [1992]), it is interesting to observe that the same is true for the moderately nonlinear SCM in this study. However, it should be noted that, since the residuals are non-Gaussian (as shown in Figure 3 and discussed in Section 3.3), the Gaussian likelihood is invalid. Therefore, the resulting parameter distributions are inappropriate for quantifying predictive uncertainty, as shown in Section 3.6 below.

Using the formal generalized likelihood function substantially changes the global parameter optima and posterior parameter distributions. The global optima corresponding to the formal generalized likelihood function are listed in Table 1, and they are substantially different from those obtained using UCODE_2005 and DREAM_(ZS) with the Gaussian likelihood function. Figure 8 shows that the histograms of the MCMC samples based on the Gaussian likelihood are dramatically different from those based on the formal generalized likelihood. The histogram of logK1 is unimodal, but the mode is different from that based on the Gaussian likelihood function. The histogram of logK2 based on the formal generalized likelihood appears to be a uniform distribution, spreading significantly wider than that based on the Gaussian likelihood. The histogram of logK3 based on the formal generalized likelihood is somewhat bimodal; the two modes are close but substantially different from that of the histogram based on the Gaussian likelihood. The same is true for the histograms of logSite. These results indicate that parametric uncertainty is larger when the formal

generalized likelihood function is used. The same was reported in Schoups and Vrugt [2010], and they attributed this to the following three reasons: (1) the heteroscedastic error model results in larger uncertainty; (2) considering residual correlation reduces information content in the data and leads to larger uncertainty; and (3) the SEP distribution has heavier tails than the Gaussian distribution. There may be specific reasons for the SCM, investigation of which is warranted in a future study.

Using the formal generalized likelihood function also substantially changes the parameter correlation. Comparing Figure 9 based on the Gaussian likelihood function with Figure 10 based on the formal generalized likelihood function shows that, except the correlation between $\log K_3$ and $\log \text{Site}$, the correlation between any two model parameters decreases when the formal generalized likelihood function is used. The weaker correlation may be attributed to the interaction between the model parameters and the parameters of the generalized likelihood function.

3.6. Comparison of predictive uncertainty

In a cross-validation manner, the SCM was used to predict Experiment 5 with 41 observations and Experiment 7 with 170 observations. The two experiments were not used for the model calibration; their chemical conditions are different from those of Experiments 1, 2, and 8 used for model calibration, and so are the breakthrough curves. Predictive uncertainty was quantified using the 95% confidence intervals (linear and nonlinear) estimated using UCODE_2005 and the 95% credible intervals estimated using the $\text{DREAM}_{(\text{ZS})}$ samples based on the Gaussian and formal generalized likelihood functions.

The MCMC results using the formal generalized likelihood function are the best for predicting the data of Experiments 5 and 7. Figures 11a and 12a show that the 95% linear and nonlinear confidence intervals are too narrow to include the observed breakthrough data. In addition, the mean predictions are biased toward delayed breakthrough. Figures 11b and 12b

show that, although the 95% credible intervals based on the Gaussian likelihood function are wider than the linear and nonlinear confidence intervals, the mean predictions are still biased. The mean predictions of the MCMC simulation based on the formal generalized likelihood function are close to the observations, and the corresponding 95% credible intervals cover significantly more observations than the other three sets of intervals (Figures 11b and 12b). These results indicate that using the formal generalized likelihood function greatly improves model prediction and uncertainty quantification. Therefore, it should be used for quantifying predictive uncertainty of groundwater reactive transport modeling, despite that its corresponding MCMC simulation is most computationally expensive.

4. Discussion

The MCMC results were used to investigate parameter sensitivity and geochemical reasons of parametric uncertainty in the SCM.

4.1. Parameter sensitivity

Parametric sensitivity was evaluated using the reduction of the entropy of the prior and posterior distributions. The entropy is defined as $-\sum_{i=1}^N p_i \log p_i$ (Papoulis [1991], p533-537), where p is parameter density function and N is number of bins on a histogram. Large entropy corresponds to large uncertainty. Table 2 lists reduction of the entropy from the prior to the posterior parameter distributions. Considering that uncertainty reduction is the largest for the most sensitive parameters, the parameter sensitivity is in the order of $\log K1 > \log \text{Site} > \log K2 > \log K3$ for the Gaussian likelihood function but $\log K1 > \log \text{Site} > \log K3 > \log K2$ for the formal generalized likelihood function. In other words, $\log K1$ (formation constant of the reaction in equation (20) associated with the weak site) and $\log \text{Site}$ (surface fraction of the strong site) are the most and second most important parameters. The two least important parameters are $\log K2$ and $\log K3$ of the reactions in equations (21) and (22) associated with

the strong site. This order of importance is physically reasonable because the multisite adsorption model causes a self-sharpening front. For the SCM, $\log K_1$ strongly influence the initial and mean arrival of uranium; in contrast, $\log K_2$ and $\log K_3$ determine adsorption to the strong site which controls the extent of tailing shown by low concentrations on the breakthrough curve. These low concentrations and the competitive adsorption onto a single site lead to a small sensitivity to the two strong site parameters. It is reasonable that, when the formal generalized likelihood function was used, $\log K_2$ became the least influential parameter, because its histogram is uniform for the formal generalized likelihood function but Gaussian-like for the Gaussian likelihood function.

The sensitivity ranking is similar to that based on the local and hybrid of local and global techniques. The local sensitivity was conducted using UCODE_2005 to calculate, for each of the four model parameters, the composite scaled sensitivity (CSS) defined in [Hill and Tiedeman, 2007]. Smaller CSS values indicate smaller parameter sensitivity. The hybrid of local and global sensitivity was conducted using the Morris One-At-a-Time (MOAT) method implemented in DAKOTA software [Adams et al., 2010]. The method evaluates elementary effect (a local sensitivity) at a number of discrete points in parameter space. The mean elementary effect, μ , measures importance of a parameter on the model output; a high mean value indicates large overall importance. A high standard deviation, σ , of the elementary effect suggests that the parameter is either interacting with other parameters or has a high nonlinear effect on the output. Figure 13 shows that the order of sensitivity based on CSS and MOAT is the same as that based on the Gaussian likelihood function (Table 2). Given the high computational cost of MCMC simulation, it is computationally efficient to conduct the local and MOAT sensitivity analysis first to select the most influential parameters for MCMC simulations.

4.2. Geochemical reasons of parametric uncertainty

Table 3 lists four sets of model parameter values to demonstrate the parametric uncertainty. Figure 14a shows that the two sets of parameter values in Cases 1 and 2 (taken from the MCMC samples) give visually identical simulated breakthrough curves that match well to the observations of Experiment 2. The same is observed in Figure 15a for the two sets of parameter values in Cases 3 and 4 (adjusted manually) for simulating Experiment 8. The reason for this uncertainty is explained below from the geochemical point of view.

Recalling that the weak site (S_1OH) of the SCM is associated with the reaction of equation (20) and the strong site (S_2OH) with the two reactions of equations (21) and (22), uranyl adsorption on the strong site can be substituted by the sorption on the weak site, and the extent of substitution depends on the model parameter values and experiment chemical conditions. Take Cases 1 and 2 for Experiment 2 as an example. Figures 14b and 14c show that, because of the different parameter values, the concentrations of the uranium surface complexes are significantly different. However, the sums of the three species are the same for the two cases, and the sums are the observed uranyl concentrations. The substitution effect is better illustrated in Cases 3 and 4 by only changing the values of $\log K_2$ and $\log K_3$ of the strong site. Because the values of $\log K_1$ and $\log Site$ are the same, the simulated S_1OHUO_2OH concentrations (associated with the weak site) of Experiment 8 are the same in Cases 3 and 4 (Figures 15b and 15c). However, when $\log K_2$ increases and $\log K_3$ decreases from Case 3 to Case 4, the simulated S_2OHUO_2OH (associated with $\log K_2$) concentrations increase and the simulated $S_2OUO_2^+$ (associated with $\log K_3$) concentrations decrease (Figures 15b and 15c). As a result, the sums of the three species are the same in Cases 3 and 4. These results suggest that, to reduce the parametric uncertainty, it is necessary to obtain concentrations of the individual species, not their summation. Alternatively, one may need to control the geochemical conditions to minimize the substitution effect, which however is beyond the

scope of this study.

5. Conclusions

The following is a summary of the key findings of the uncertainty quantification based on the regression and local optimization methods implemented in UCODE_2005 and Bayesian methods using MCMC simulation implemented in DREAM_(ZS) with the Gaussian and formal generalized likelihood functions:

- (1) The response surface of least-square objective function is complex, with mixture of closely grouped peaks and valleys, due to nonlinearity of groundwater reactive transport models. This poses great challenges to both local and global optimization methods. DREAM_(ZS) is robust to find the global parameter optima.
- (2) In the residuals of UCODE_2005 model calibration, model errors are significantly larger than measurement errors. The formal generalized likelihood function of Schoups and Vrugt [2010] is needed to characterize heteroscedasticity, non-Gaussian density function, and temporal correlation of the residuals.
- (3) When the Gaussian likelihood function is used in the DREAM_(ZS) simulation, the posterior parameter distributions and credible intervals are similar to those of UCODE_2005 based on the Gaussian assumption of likelihood function, model parameters, and model predictions.
- (4) When the formal generalized likelihood function is used in the DREAM_(ZS) simulation, the posterior parameter distributions are non-Gaussian with multiple modes. The parametric uncertainty (measured by the Shannon's entropy) is larger than that of UCODE_2005 and DREAM_(ZS) with the Gaussian likelihood function.
- (5) The credible intervals based on the formal generalized likelihood function are less biased and have larger predictive coverage than the credible intervals based on Gaussian likelihood function and the linear and nonlinear confidence intervals of regression

methods. However, the credible interval based on the generalized likelihood function cannot cover all the cross-validation data, suggesting that, in addition to considering parametric uncertainty, model uncertainty should be considered to further improve model predictions.

(6) While DREAM_(ZS) simulation with the formal generalized likelihood function gives better results as summarized above, it is computationally expensive. This computational cost may be reduced by using computationally efficient surrogate models [Razavi et al., 2012], which may be built using polynomial chaos expansion [Laloy et al., 2013] and sparse grid collocation [Zhang et al., 2013] within the DREAM framework.

(7) For the surface complexation model considered in this study, parametric uncertainty is caused by the competing reactions and the substitutable effect between adsorption on the strong and weak sites. DREAM_(ZS), local, and global sensitivity methods gave the similar order of parameter importance (from most to least importance): $\log K_1 > \log \text{Site} > \log K_3 > \log K_2$ or $\log K_1 > \log \text{Site} > \log K_2 > \log K_3$. When DREAM_(ZS) is not computationally affordable, sensitivity analysis may be conducted using computationally frugal methods such as MOAT.

Although this study was conducted as a numerical experiment for the surface complexation model of uranium reactive transport, it is expected that many conclusions of this study are applicable to other groundwater reactive transport models and many environmental models in a broader sense.

Acknowledgment

This work was supported in part by NSF-EAR grant 0911074, DOE-SBR grant DE-SC0002687, DOE Early Career Award DE-SC0008272, and ORAU/ORNL High Performance Computing Grant. The first author was employed by the Florida State University when conducting part of this research, and he was also supported by the National Nature

Science Foundation of China grants No. 41172206 and 41172207. The authors thank Jasper Vrugt for providing the DREAM_{zs} code, and Bruno Mendes and David Draper for inspiring discussion on comparison of confidence and credible intervals.

References

Adams, B. M., Dalbey, K. R., Eldred, M. S., Gay, D. M., Swiler, L. P., Bohnhoff, W. J., Eddy, J. P., Haskell, K., and P. D. Hough (2010), DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis. Sandia National Laboratories.

Aggarwal, M. and J. Carrayrou (2006), Parameter Estimation for Reactive Transport by a Monte-Carlo Approach, *AIChE Journal*, 52(6), 2281-2289.

Aggarwal, M., M.C.A. Ndiaye, and J. Carrayrou (2007), Parameters estimation for reactive transport: A way to test the validity of a reactive model, *Physics and Chemistry of the Earth*, 32, 518-529.

Beck, B. (1987), Water quality modeling: a review of the analysis of uncertainty. *Water Resources Research* 23 (8), 1393–1442.

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modeling. *Advances in Water Resources* 16, 14–51.

Beven, K., P. J. Smith, and J. E. Freer (2008), So just why would a modeler choose to be incoherent, *J. Hydrol.*, 354, 15 – 32.

Box, E.P., and G.C. Tiao (1992), *Bayesian Inference in Statistical Analysis*, 558 pp., Wiley, New York.

Brooks, S. P., and A. Gelman (1998), General methods for monitoring convergence of

661 iterative simulations, *J Comput Graph Stat*, 7(4), 434-455.

662 Cabaniss, S.E. (1999), Uncertainty propagation in geochemical calculation: nonlinearity in
663 solubility equilibria, *Applied Geochemistry*, 14, 255-262.

664 Carrera, J. (1993), An overview of uncertainties in modelling groundwater solute transport, *J.*
665 *Contam. Hydrol.*, 13: 23-48.

666 Casella, G., and R. L. Berger, (2002). *Statistical inference*. Duxbury Press.

667 Chen, J., S. S. Hubbard, D. Gaines, V. Korneev, G. Baker, and D. Watson (2010), Stochastic
668 estimation of aquifer geometry using seismic refraction data with borehole depth
669 constraints, *Water Resour. Res.*, 46, W11539, doi:10.1029/2009WR008715.

670 Chen, J., A. Kemna, and S. S. Hubbard (2008), A comparison between Gauss-Newton and
671 Markov chain Monte Carlo based methods for inverting spectral induced polarization data
672 for Cole-Cole parameters, *Geophysics*, 73(6), F247-F259.

673 Chu, W., X. Gao, and S. Sorooshian (2010), Improving the shuffled complex evolution
674 scheme for optimization of complex nonlinear hydrological systems: Application to the
675 calibration of the Sacramento soil - moisture accounting model, *Water Resour. Res.*, 46,
676 W09530, doi:10.1029/2010WR009224.

677 Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working
678 hypotheses for hydrological modeling, *Water Resour. Res.*, 47, W09301,
679 doi:10.1029/2010WR009827.

680 Cooley, R. L. (2004), A theory for modeling groundwater flow in heterogeneous media, U. S.
681 Geological Survey Professional Paper 1979.

682 Curtis, G. P. (2005), *Documentation and Applications of the Reactive Geochemical Transport*

683 Model RATEQ. Report NUREG CR-6871, U.S. Nuclear Regulatory Commission,
684 Rockville, MD, 97 p.

685 Curtis, G.P. and Davis, J.A. (2006), Tests of Uranium(VI) Adsorption Models in a Field
686 Setting, Report NUREG CR-6911, U.S. Nuclear Regulatory Commission, Rockville, MD,
687 99 p.

688 Dai, Z., and J. Samper (2004), Inverse problem of multicomponent reactive chemical
689 transport in porous media: Formulation and applications, *Water Resour. Res.*, 40, W07407,
690 doi:10.1029/2004WR003248.

691 Davis, J.A., and D.B. Kent (1990), Surface complexation modeling in aqueous geochemistry:
692 in *Mineral-Water Interface Geochemistry*, M.F. Hochella and A.F. White, eds., *Reviews*
693 *in Mineralogy*, v. 23, p. 177-260.

694 Davis, J.A., D.E. Meece, M. Kohler and G.P. Curtis (2004b), Approaches to surface
695 complexation modeling of uranium(VI) adsorption on aquifer sediments, *Geochim.*
696 *Cosmochim. Acta* 68, pp. 3621-3641.

697 Davis, J.A., S.B. Yabusaki, C.I. Steefel, J.M. Zachara, G.P. Curtis, G.D. Redden, L.J.
698 Criscenti, and B.D. Honeyman (2004a), Assessing conceptual models for subsurface
699 reactive transport of inorganic contaminants, *EOS*, 85(4), November.

700 Denison, F.H. and J. Garnier-Laplace (2005), The effects of database parameter uncertainty in
701 uranium (VI) equilibrium calculations, *Geochimica et Cosmochimica Acta*, 69(9),
702 2183-2191.

703 Doherty, J. (2003), *Manual for the PEST Surface Water Modelling Utilities*. Watermark
704 Numerical Computing, Australia. Available from <http://www.sspa.com/pest>

705 Doherty, J. (2010), *PEST: Model-independent parameter estimation, User Manual*, 5th edition

706 (with slight additions in 2010), Watermark Numerical Computing, Brisbane, Australia.
 707 Draper, N. R., and H. Smith (1981), Applied regression analysis, 2nd ed., John Wiley, New
 708 York.
 709 Draper, D. (2007), Bayesian multilevel analysis and MCMC, in J. de Leeuw and E. Meijer
 710 (Editors), Handbook of Multilevel Analysis, Springer.
 711 Duan, Q. Y., S. Sorooshian, and V. K. Gupta (1992), Effective and Efficient Global
 712 Optimization for Conceptual Rainfall-Runoff Models, Water Resour. Res., 28(4),
 713 1015-1031.
 714 Duan, Q., V.K. Gupta, and S. Sorooshian (1993), Shuffled Complex Evolution Approach for
 715 Effective and Efficient Global Minimization, J. Optim. Theory Appl. 76, 501–521.
 716 Duan, Q. Y., S. Sorooshian, and V. K. Gupta (1994), Optimal Use of the SCE-UA Global
 717 Optimization Method for Calibrating Watershed Models, J. Hydrol., 158(3-4), 265-284.
 718 Finsterle, S. (2007), iTOUGH2 command reference, Rep. LBNL-40041. Lawrence Berkeley
 719 National Laboratory, Berkeley, CA.
 720 Finsterle, S. and Y. Zhang (2011), Solving iTOUGH2 simulation and optimization on
 721 problems using the PEST protocol. Environ. Model. Software, 26:959–968,
 722 doi:10.1016/j.envsoft.2011.02.008.
 723 Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple
 724 sequences, Statistical Sci. 7:457-511.
 725 Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004), Bayesian data analysis.
 726 Chapman & Hall.
 727 Guillamont, R., T. Fanghanel, I. Grenthe, V. Neck, D. Palmer, and M.H. Rand (2003), Update

on the chemical thermodynamics of uranium, neptunium, plutonium, americium and technetium, OECD-NEA, Chemical Thermodynamics Vol. 5, 182-187, Elsevier, Amsterdam.

Gupta, H.V., M.P. Clark, J.A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, Water Resources Research, 48, W08301, doi:10.1029/2011WR011044.

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

Hill, M. C. and C. R. Tiedeman, (2007), Effective Groundwater Model Calibration. John Wiley & Sons, Inc.

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

Hou, Z., and Y. Rubin (2005), On minimum relative entropy concepts and prior compatibility issues in vadose zone inverse and forward modeling, Water Resour. Res., 41, W12425, doi:10.1029/2005WR004082.

Jakeman, A.J. and G.M. Hornberger (1993). How much complexity is warranted in a rainfall-runoff model? Water Resources Research 29 (8), 2637–2649.

Keating, E. H., J. Doherty, J. A. Vrugt, and Q. Kang (2010), Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality, Water Resour. Res., 46, W10517, doi:10.1029/2009WR008584.

Kitanidis, P. K. (1986), Parameter uncertainty in estimation of spatial functions: Bayesian analysis, Water Resour. Res., 22(4), 499-507.

- Kohler, M., G. Curtis, D. Kent, and J. Davis (1996), Experimental investigation and modeling of Uranium(VI) transport under variable chemical conditions. *Water Resources Research* 32, 3539-3551.
- Laloy, E., and J. A. Vrugt (2012), High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing, *Water Resour. Res.*, 48, W01526, doi:10.1029/2011WR010608.
- Laloy, E., B. Rogiers, J. A. Vrugt, D. Mallants, and D. Jacques (2013), Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov Chain Monte Carlo simulation and polynomial chaos expansion, *Water Resour. Res.*, 49, doi:10.1002/wrcr.20226.
- Leavitt, J.J., K.J. Howe, and S.E. Cabaniss (2011), Equilibrium modeling of U(VI) speciation in high carbonate groundwaters: model error and propagation of uncertainty, *Applied Geochemistry*, 26, 2019-2026, doi:10.1016/j.apgeochem.2011.06.031.
- Liu, C., J. M. Zachara, N. P. Qafoku, and Z. Wang (2008), Scale-dependent desorption of uranium from contaminated subsurface sediments, *Water Resour. Res.*, 44, W08413, doi:10.1029/2007WR006478.
- Liu, Y., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, 43, W07401, doi:10.1029/2006WR005756.
- Lu, D., M. Ye, and S.P. Neuman (2011), Dependence of Bayesian model selection criteria and Fisher information matrix on sample size, *Mathematical Geosciences*, 43(8), 971-993, DOI 10.1007/s11004-011-9359-0.
- Lu, D., M. Ye, and M.C. Hill (2012), Analysis of Regression Confidence Intervals and

774 Bayesian Credible Intervals for Uncertainty Quantification, *Water Resour. Res.*, 48,
775 W09521, doi:10.1029/2011WR011289.

776 Lu, D., M. Ye, P. D. Meyer, G. P. Curtis, X. Shi, X.-F. Niu, and S. B. Yabusaki (2013),
777 Effects of error covariance structure on estimation of model averaging weights and
778 predictive performance, *Water Resour. Res.*, 49, 6029–6047, doi:10.1002/wrcr.20441.

779 Majdalani, S., M. Fahs, J. Carrayrou, P. Ackerer (2009), Reactive transport parameter
780 estimation: Genetic algorithm vs. Monte carlo approach. *AIChE Journal*, 55(8): 1959–
781 1968.

782 Marshall, L., D. Nott, and A. Sharma (2004), A comparative study of Markov chain Monte
783 Carlo methods for conceptual rainfall-runoff modeling, *Water Resour. Res.*, 40(2).

784 Massoudieh A., A. Mathew, T. R. Ginn (2008), Column and batch reactive transport
785 experiment parameter estimation using a genetic algorithm, *Computers & Geosciences*,
786 34(1): 24-34.

787 Matott, L. S., J. E. Babendreier, and S. T. Purucker (2009), Evaluating uncertainty in
788 integrated environmental models: A review of concepts and tools, *Water Resour. Res.*, 45,
789 W06421, doi:10.1029/2008WR007301.

790 Matott, L.S., and A.J. Rabideau (2008a), Calibration of complex subsurface reaction models
791 using a surrogate-model approach, *Advances in Water Resources*, 31(2), 1697-1707,
792 doi:10.1016/j.advwatres.2008.08.006.

793 Matott, L.S., and A.J. Rabideau (2008b), Calibration of subsurface batch and
794 reactive-transport models involving complex biogeochemical processes, *Advances in*
795 *Water Resources*, 31(2), 269-286.

796 Meyer, P.D., M. Ye, M.L. Rockhold, S.P. Neuman, K.J. Cantrell (2007), Combined

797 Estimation of Hydrogeologic Conceptual Model, Parameter, and Scenario Uncertainty,
798 NUREG/CR-6940, PNNL-16396, U.S. Nuclear Regulatory Commission, Office of
799 Nuclear Regulatory Research, Rockville, Maryland, USA.

800 Miller, G.L. (2010), Parametric uncertainty analysis of uranium transport surface
801 complexation models. Master Thesis. Florida State University.

802 Mahoney, J.J., S.A. Cadle, and R.T. Jakubowski (2009), Uranyl adsorption onto hydrous
803 ferric oxide-A re-evaluation for the diffuse layer model database, Environ. Sci. Technol.,
804 43(24), 9260-9266.

805 Morris M. D. (1991), Factorial Sampling Plans for Preliminary Computational Experiments,
806 Technometrics, 21(2), pp. 239-245.

807 Mugunthan, P., C. A. Shoemaker, and R. G. Regis (2005), Comparison of function
808 approximation, heuristic, and derivative-based methods for automatic calibration of
809 computationally expensive groundwater bioremediation models, Water Resour. Res., 41,
810 W11427, doi:10.1029/2005WR004134.

811 Neuman, S.P. and S. Yakowitz (1979), Statistical approach to the inverse problem of aquifer
812 hydrology. 1. Theory. Water Resour. Res., 15(4), 845-860,
813 DOI:10.1029/WR015i004p00845.

814 Neuman, S.P. (1973), Calibration of distributed parameter groundwater flow models viewed
815 as a multiple-objective decision process under uncertainty, Water Resources Research,
816 9(4), 1006-1021, DOI: 10.1029/WR009i004p01006.

817 Over, M. W., Y. Yang, X. Chen, and Y. Rubin (2013), A strategy for improved computational
818 efficiency of the method of anchored distributions, Water Resour. Res., 49, 3257–3275,
819 doi:10.1002/wrcr.20182.

820 Pappenberger, F., and K.J. Beven (2006), Ignorance is bliss: or seven reasons not to use
821 uncertainty analysis. *Water Resources Research* 42, W05302,
822 doi:10.1029/2005WR004820.

823 Papoulis, A. (1991), Probability, random variables, and stochastic processes. 3rd edition.
824 McGraw-Hill (New York).

825 Poeter, E., Hill, M., Banta, E., Mehl, S., and Christensen, S. (2005), UCODE2005 and Six
826 Other Computer Codes for Universal Sensitivity Analysis, Calibration, and Uncertainty
827 Evaluation. U.S. Geological Survey.

828 Herman, J.D., J.B. Kollat, P.M. Reed, and T. Wagener (2013), Technical note: Method of
829 Morris effectively reduces the computational demands of global sensitivity analysis for
830 distributed watershed models, *Hydrology and Earth System Sciences*, 10, 4275-4299,
831 doi:10.5194/hessd-10-4275-2013.

832 Razavi, S., B. A. Tolson, and D. H. Burn (2012), Review of surrogate modeling in water
833 resources, *Water Resour. Res.*, 48, W07401, doi:10.1029/2011WR011527.

834 Refsgaard, J.C., S. Christensen, T.O. Sonnenborg, D. Seifert, A.L. Hojberg, L. Trolborg
835 (2012), Review of strategies for handling geological uncertainty in groundwater flow and
836 transport modeling, *Advances in Water Resources*, 36, 36-50.

837 Rojas, R., Feyen, L., and Dassargues, A. (2008), Conceptual model uncertainty in
838 groundwater modeling: Combining generalized likelihood uncertainty estimation and
839 bayesian model averaging. *Water Resources Research* 44, W12418.

840 Rubin, Y., X. Chen, H. Murakami, and M. Hahn (2010), A Bayesian approach for inverse
841 modeling, data assimilation, and conditional simulation of spatial random fields, *Water*
842 *Resour. Res.*, 46, W10523, doi:10.1029/2009WR008799.

843 Scheibe, T.D., P. Meakin, P.C. Lichtner, and D. Zachmann (2008), Science for Problems
844 under the Surface. *SciDAC Review*, 7, 32-41.

845 Schoups, G., and J. A. Vrugt (2010), A formal generalized likelihood function for parameter
846 and predictive inference of hydrologic models with correlated, heteroscedastic, and non -
847 Gaussian errors, *Water Resour. Res.*, 46, W10531, doi:10.1029/2009WR008933.

848 Shi, X., M. Ye, S. Finsterle, and J. Wu (2012), Comparing nonlinear regression and Markov
849 chain Monte Carlo methods for assessment of predictive uncertainty in vadose zone
850 modeling, *Vadose Zone Journal*, 11(4), doi:10.2136/vzj2011.0147.

851 Smith, T. J., and L. A. Marshall (2008), Bayesian methods in hydrologic modeling: A study
852 of recent advancements in Markov chain Monte Carlo techniques, *Water Resour. Res.*, 44,
853 W00B05, doi:10.1029/2007WR006705.

854 Srinivasan, G., D. M. Tartakovsky, B. A. Robinson, and A. B. Aceves (2007), Quantification
855 of uncertainty in geochemical reactions, *Water Resour. Res.*, 43, W12415,
856 doi:10.1029/2007WR006003.

857 Steefel, C. I., S. Carroll, P. H. Zhao, and S. Roberts (2003), Cesium migration in Hanford
858 sediment: a multisite cation exchange model based on laboratory transport experiments, *J.*
859 *Contam. Hydrol.*, 67(1-4), 219-246.

860 Steefel, C.I., D.J. DePaolo, P.C. Lichtner (2005), Reactive transport modeling: an essential
861 too and a new research approach for the earth sciences, *Earth and Planetary Science*
862 *Letters*, 240, 539-558.

863 Sun, Y., C. Tong, D. Duan, T.A. Buscheck, J.A. Blink (2012), Combining simulation and
864 emulation for calibrating sequentially reactive transport system, *Transp. Porous Med.*, 92,
865 509-526.

866 Tang, G., M.A. Mayes, J.C. Parker, X.L. Yin, D.B. Watson, P.M. Jardine (2009), Improving
867 parameter estimation for column experiments by multi-model evaluation and comparison,
868 Journal of Hydrology, 376, 567–578, doi:10.1016/j.jhydrol.2009.07.063.

869 Tartakovsky, D. M., M. Dentz, and P. C. Lichtner (2009), Probability density functions for
870 advective-reactive transport with uncertain reaction rates, Water Resour. Res., 45,
871 W07414, doi:10.1029/2008WR007383.

872 Tebes-Stevens, C. L., Espinoza, F., and Valocchi, A. J. (2001), Evaluating the sensitivity of a
873 subsurface multicomponent reactive transport model with respect to transport and reaction
874 parameters. Journal of Contaminant Hydrology 52, 3-27.

875 Ter Braak, C. J. F. (2006), A Markov Chain Monte Carlo version of the genetic algorithm
876 differential evolution: easy Bayesian computing for real parameter spaces, Stat Comput,
877 16(3), 239-249.

878 Vrugt, J. A., B. A. Robinson, and J. M. Hyman (2009a), Self-adaptive multimethod search for
879 global optimization in real - parameter spaces, IEEE Trans. Evol. Comput., 13(2), 243–
880 259, doi:10.1109/TEVC.2008.924428.

881 Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon
882 (2009b), Accelerating Markov chain Monte Carlo simulation by differential evolution
883 with self-adaptive randomized subspace sampling, Int. J. Nonlinear Sci. Numer. Simul.,
884 10(3), 273–290.

885 Vrugt, J. A., C. J. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008b),
886 Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with
887 Markov chain Monte Carlo simulation, Water Resources Research, 44, W00B09,
888 doi:10.1029/2007WR006720.

889 Vrugt, J. A., P. H. Stauffer, T. Wohling, B. A. Robinson, and V. V. Vesselinov (2008a),
890 Inverse modeling of subsurface flow and transport properties: A review with new
891 developments, *Vadose Zone J.*, 7(2), 843–864, doi:10.2136/vzj2007.0078.

892 Wang, M. and C. Zheng (1996), Parameter estimation for transient and steady-state flow
893 models using genetic algorithms, in *ModelCARE 96: Calibration and Reliability in*
894 *Groundwater Modeling*, K. Kavar and P. van de Heijde, eds., IAHS Publ. 237: 21-30,
895 International Association of Hydrological Sciences.

896 Wagener, T. and H.V. Gupta (2005), Model identification for hydrological forecasting under
897 uncertainty, *Stoch. Environ. Res. Risk Assess.*, 19: 378-387.

898 Woodbury, A. D. and T. J. Ulrych (2000). A full-Bayesian approach to the groundwater
899 inverse problem for steady state flow. *Water Resources Research* 36(8): 2081-2093.

900 Ye M, Neuman SP, Meyer PD (2004), Maximum Likelihood Bayesian averaging of spatial
901 variability models in unsaturated fractured tuff, *Water Resour. Res.*, 40, W05113,
902 doi:10.1029/2003WR002557.

903 Ye, M., Meyer, P. D., and Neuman, S. P. (2008), On model selection criteria in multimodel
904 analysis. *Water Resources Research* 44, W034228.

905 Ye, M., D. Lu, S. P. Neuman, and P. D. Meyer (2010a), Comment on “Inverse groundwater
906 modeling for hydraulic conductivity estimation using Bayesian model averaging and
907 variance window” by Frank T.-C. Tsai and Xiaobao Li, *Water Resour. Res.*, 46, W02801,
908 doi:10.1029/2009WR008501.

909 Ye, M., K.F. Pohlmann, J.B. Chapman, G.M. Pohll, and D.M. Reeves (2010b), A
910 model-averaging method for assessing groundwater conceptual model uncertainty,
911 *Ground Water*, doi:10.1111/j.1745-6584.2009.00633.x.

912 Yeh, W. W-G and Y.S. Yoon (1981), Aquifer parameter identification with optimum
 913 dimension in parameterization, *Water Resour. Res.*, (3), 664–672.

914 Yeh, W. W.-G. (1986), Review of Parameter Identification Procedures in Groundwater
 915 Hydrology: The Inverse Problem, *Water Resour. Res.*, 22(2), 95–108,
 916 doi:10.1029/WR022i002p00095.

917 Zhang, G., D. Lu, M. Ye, M. Gunzburger, and C. Webster (2013), An adaptive sparse-grid
 918 high-order stochastic collocation method for Bayesian inference in groundwater reactive
 919 transport modeling, *Water Resour. Res.*, 49, doi:10.1002/wrcr.20467.

920 Zhao, J., T.D. Scheibe, R. Mahadevan (2011), Model-based analysis of the role of biological,
 921 hydrological, and geochemical factors affecting uranium bioremediation, *Biotechnol.*
 922 *Bioeng.*, 108, 1537-1548.

923 Zheng, C. and Wang, P. P. 1999. MT3DMS, A modular three-dimensional multi-species
 924 transport model for simulation of advection, dispersion and chemical reactions of
 925 contaminants in groundwater systems; documentation and users guide. U.S. Army
 926 Engineer Research and Development Center Contract Report, Vicksburg, MS.

927

928
929
930
931

932
933
934

935

936
937
938
939
940
941

942

943

944

945

946

947

948

949

950

951

952

LIST OF TABLES

Table 1. Parameter estimates obtained using UCODE_2005, DAKOTA, SCE-UA, and MCMC simulation (with Gaussian and formal generalized likelihood functions). The lower (LB) and upper (UB) bounds of the parameter ranges are also listed.

Parameter Name	UCODE Initial Values	UCODE Estimates	LB	UB	DAKOTA	SCE-UA	MCMC (Gaussian)	MCMC (Generalized)
logK1	-5.0	-4.866	-6.5	-3.5	-4.865	-4.864	-4.864	-4.946
logK2	-3.0	-3.128	-5.2	-2.0	-3.121	-3.116	-3.116	-3.122
logK3	1.0	1.220	0.0	2.0	1.225	1.230	1.230	1.621
logSite	-2.0	-1.964	-3.0	-1.0	-1.970	-1.975	-1.975	-2.279

Table 2. Entropy values of prior and posterior parameter distributions based on the Gaussian and formal generalized likelihood function.

Parameter Name	Prior Entropy	Gaussian Likelihood			Formal Generalized Likelihood		
		Posterior Entropy	Entropy Reduction	Rank of Reduction	Posterior Entropy	Entropy Reduction	Rank of Reduction
logK1	1.9031	0.5238	1.3793	1	0.5940	1.3091	1
logK2	1.9031	1.0008	0.9023	3	1.6552	0.2479	4
logK3	1.9031	1.1675	0.7356	4	1.3936	0.5095	3
logSite	1.9031	0.9703	0.9328	2	1.1312	0.7719	2

993 Table 3. Parameter values in four cases to investigate parametric uncertainty. For Cases 1 and
 994 2, the simulated breakthrough curves of Experiment 2 are visually identical (Figure 13); for
 995 Cases 3 and 4, the simulated breakthrough curves of Experiment 8 are visually identical
 996 (Figure 14).

		logK1	logK2	logK3	logSite
Experiment 2	Case 1	-5.439	-3.910	1.621	-2.081
	Case 2	-4.475	-1.240	1.966	-3.462
Experiment 8	Case 3	-4.947	-4.800	1.600	-1.622
	Case 4	-4.947	-2.400	0.100	-1.622

997
 998
 999
 1000

LIST OF FIGURES

Figure 1. (a) Bimodal distribution of a single parameter θ , (b) model simulations (y) corresponding to parameter samples from the distribution, where y is an observation, and (c) SSR (sum of squared residuals) as objective function, where SSR_{min} and SSR_{accepted} are minimum and accepted SSR . This figure illustrates that multiple minima of the objective function are caused by multiple modes of the parameter density function.

Figure 2. Two-dimensional surface of $-\log_{10}SSWR$ when $\log K1$, $\log Site$, and $\log K2$ vary simultaneously (the least important parameter, $\log K3$, is fixed at its global optimum). The red arrow points to the global optima of $\log K1$, $\log Site$, and $\log K2$.

Figure 3. Results of model calibration using UCODE_2005: (a-1) – (a-3) comparison of observed and simulated breakthrough curves for experiments 1, 2, and 8, (b) graph of weighted simulations and weighted observations, (c) graph of residuals and simulations to demonstrate, (d) actual histogram of residuals and fit to a Gaussian density function, (e) partial autocorrelation coefficient of the residuals of Experiment 1 (blue lines represent the 95% confidence interval).

Figure 4. Convergence behaviors of the Gelman-Rubin Statistics for DREAM_{ZS} realizations using Gaussian likelihood function.

Figure 5. Evolution of three MCMC chains for the four parameters when using Gaussian likelihood function.

Figure 6. Convergence behaviors of the Gelman-Rubin Statistics for DREAMzs realizations using formal generalized likelihood function.

Figure 7. Evolution of three MCMC chains for the four parameters when using formal generalized likelihood function.

Figure 8. Histograms of four model parameters from DREAMzs samples based on Gaussian (black) and formal generalized (red) likelihood functions. The blue lines plot the normal distributions based on UCODE results, i.e., using the parameter estimates as the means and the estimation variances as the variances.

Figure 9. Scatter plots of parameter correlation using the MCMC samples based on the Gaussian likelihood.

Figure 10. Scatter plots of parameter correlation using the MCMC samples based on the formal generalized likelihood function.

Figure 11. (a) 95% linear and nonlinear confidence intervals evaluated using UCODE_2005, (b) 95% nonlinear credible intervals evaluated using the Gaussian and formal generalized likelihood functions for Experiment 5.

Figure 12. (a) 95% linear and nonlinear confidence intervals evaluated using UCODE_2005, (b) 95% nonlinear credible intervals evaluated using the Gaussian and formal generalized likelihood functions for Experiment 7.

Figure 13. (a) Local sensitivity estimated using UCODE_2005 for the global parameter optima and (b) global sensitivity estimated using the MOAT method. The order of parameter is consistent that the most and second most important parameters are logK1 and logSite, and the two least important parameters are logK2 and logK3.

Figure 14. (a) Comparison of calculated and observed breakthrough curves for Experiment 2 with two sets of different parameter values (Case 1 – 2 in Table 3). Simulated breakthrough curves of three U(VI) complexation species for (b) Case 1 and (c) Case 2.

Figure 15. (a) Comparison of calculated and observed breakthrough curves for Experiment 8 with two sets of different parameter values (Case 3 – 4 in Table 3). Simulated breakthrough curves of three U(VI) complexation species for (b) Case 3 and (c) Case 4.

Appendix B

This appendix contains the paper of Lu et al. (2012), which was published in Water Resources Research. This research helps answer the questions below explained in detail in the report:

Question 3: How to quantify parametric uncertainty of groundwater reactive transport modeling?

It in particular answers the question below regarding the regression and Bayesian methods developed for quantifying parametric uncertainty:

What are the relations between the confidence intervals of regression methods and credible intervals of Bayesian methods?

Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification

Dan Lu,¹ Ming Ye,¹ and Mary C. Hill²

Received 16 August 2011; revised 27 June 2012; accepted 4 July 2012; published 14 September 2012.

[1] Confidence intervals based on classical regression theories augmented to include prior information and credible intervals based on Bayesian theories are conceptually different ways to quantify parametric and predictive uncertainties. Because both confidence and credible intervals are used in environmental modeling, we seek to understand their differences and similarities. This is of interest in part because calculating confidence intervals typically requires tens to thousands of model runs, while Bayesian credible intervals typically require tens of thousands to millions of model runs. Given multi-Gaussian distributed observation errors, our theoretical analysis shows that, for linear or linearized-nonlinear models, confidence and credible intervals are always numerically identical when consistent prior information is used. For nonlinear models, nonlinear confidence and credible intervals can be numerically identical if parameter confidence regions defined using the approximate likelihood method and parameter credible regions estimated using Markov chain Monte Carlo realizations are numerically identical and predictions are a smooth, monotonic function of the parameters. Both occur if intrinsic model nonlinearity is small. While the conditions of Gaussian errors and small intrinsic model nonlinearity are violated by many environmental models, heuristic tests using analytical and numerical models suggest that linear and nonlinear confidence intervals can be useful approximations of uncertainty even under significantly nonideal conditions. In the context of epistemic model error for a complex synthetic nonlinear groundwater problem, the linear and nonlinear confidence and credible intervals for individual models performed similarly enough to indicate that the computationally frugal confidence intervals can be useful in many circumstances. Experiences with these groundwater models are expected to be broadly applicable to many environmental models. We suggest that for environmental problems with lengthy execution times that make credible intervals inconvenient or prohibitive, confidence intervals can provide important insight. During model development when frequent calculation of uncertainty intervals is important to understanding the consequences of various model construction alternatives and data collection strategies, strategic use of both confidence and credible intervals can be critical.

Citation: Lu, D., M. Ye, and M. C. Hill (2012), Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification, *Water Resour. Res.*, 48, W09521, doi:10.1029/2011WR011289.

1. Introduction

[2] Environmental modeling is often used to predict effects of future anthropogenic and/or natural occurrences. Predictions are always uncertain. Epistemic and aleatory prediction uncertainty is caused by data errors and scarcity, parameter uncertainty, model structure uncertainty, and scenario uncertainty [Morgan and Henrion, 1990; Neuman and Wierenga, 2003; Meyer *et al.*, 2007; Renard *et al.*, 2011; Clark *et al.*, 2011]. Here we mostly explore the propagation of parameter uncertainty into measures of prediction

uncertainty using two common measures: individual confidence intervals based on classical regression theories (Draper and Smith [1998] and Hill and Tiedeman [2007] show how prior information can be included in regression theories) and individual credible intervals (also known as probability intervals) based on Bayesian theories [Box and Tiao, 1992; Casella and Berger, 2002]. On the other hand, alternative models are also constructed in this study to allow consideration of these measures of parameter uncertainty in the context of model structure uncertainty.

[3] Confidence intervals are of interest because they generally can be calculated using tens to thousands of model runs versus the tens of thousands to millions of model runs needed to calculate credible intervals. While computational cost for evaluating Bayesian credible intervals can be reduced by using surrogate modeling such as response surface surrogates and low-fidelity physically based surrogates [Razavi *et al.*, 2012], surrogate modeling generally requires thousands of model runs to establish the response surface surrogate and the

¹Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA.

²U.S. Geological Survey, Boulder, Colorado, USA.

Corresponding author: M. C. Hill, U.S. Geological Survey, 3215 Marine St., Boulder, CO 80303, USA. (mchill@usgs.gov)

©2012. American Geophysical Union. All Rights Reserved.
0043-1397/12/2011WR011289

surrogate model. Considering the utility of frequently calculating uncertainty measures to evaluate the consequences of various model construction alternatives and data collection strategies [Tiedeman *et al.*, 2003, 2004; Dausman *et al.*, 2010; Neuman *et al.*, 2012; Lu *et al.*, 2012], less computationally demanding methods are appealing.

[4] While both confidence and credible intervals have been used in environmental modeling [Schoups and Vrugt, 2010; Krzysztofowicz, 2010; Aster *et al.*, 2012], it is not possible to provide a literature review and examples from such a large field. Here, we focus on groundwater modeling for the literature review below and for the complex synthetic test case considered. Since groundwater modeling and many other fields of environmental modeling share similar modeling protocols and mathematical and statistical methods for uncertainty quantification and interpretation, guidelines for using confidence and credible intervals drawn from the theoretical analyses and test cases considered in this work are expected to be generally applicable to other fields of environmental modeling.

[5] Among the pioneering groundwater works, Cooley [1977, 1979, 1983], Yeh and Yoon [1981], Carrera and Neuman [1986], and Yeh [1986] used linear confidence intervals based on observations and prior information; Cooley and Vecchia [1987] developed a method for estimating nonlinear confidence intervals; Kitanidis [1986] presented a Bayesian framework of quantifying parameter and predictive uncertainty. More recent development and applications of the intervals can be found in Sciortino *et al.* [2002], Tiedeman *et al.* [2003, 2004], Cooley [2004], Montanari and Brath [2004], Samanta *et al.* [2007], Hill and Tiedeman [2007], Gallagher and Doherty [2007], Montanari and Grossi [2008], Dausman *et al.* [2010], Li and Tsai [2009], Wang *et al.* [2009], Parker *et al.* [2010], Schoups and Vrugt [2010], and Fu *et al.* [2011], among others. The study of confidence and credible intervals in the context of multimodel analysis has been considered by Neuman [2003], Ye *et al.* [2004], Poeter and Anderson [2005], Poeter and Hill [2007], Neuman *et al.* [2012], Lu *et al.* [2012], and L. Foglia *et al.* (Analysis of model discrimination techniques, the case of the Maggia Valley, Southern Switzerland, submitted to *Water Resources Research*, 2012), among others.

[6] A related field is sensitivity analysis, which is commonly used to diagnose contributions to uncertainty measures [e.g., Saltelli *et al.*, 2008]. In sensitivity analysis, comparisons of linear methods related to the linear confidence intervals in this work and global methods with Monte Carlo sampling more similar to the nonlinear Bayesian methods have suggested both utility and lack of utility of the linear methods. Foglia *et al.* [2007] suggest a high level of utility; Tang *et al.* [2007] suggest little utility. Of note is that the linear measure used by Tang *et al.* [2007] (labeled CS) has dimensions of the reciprocal of the parameter values, with the predictable result that parameters with small values (such as LZPK) are rated as “most important.” This contradicts their results from global methods. Multiplying by the parameter values to obtain dimensionless linear measures results in LZPK being rated as “less important” which is consistent with their results from global methods.

[7] Confidence and credible intervals are conceptually different and understanding their fundamental differences and similarities can aid appropriate use. In this study, we consider linear confidence and credible intervals for linear and linearized-nonlinear models, and nonlinear confidence and

credible intervals for nonlinear models. These intervals are defined in section 2. For nonlinear models, nonlinear confidence and credible intervals account for model nonlinearity and are thus expected to be more accurate than linear intervals calculated from linearized-nonlinear models. Table 1 summarizes previous studies that compare confidence and credible intervals. The previous studies report linear confidence intervals and nonlinear credible intervals that differ by between 2% and 82%, while nonlinear confidence and credible intervals differ by between –24% and 7%. The previous studies mainly focused on comparing calculated confidence and credible intervals without a thorough discussion of underlying theories, and the large differences have not been fully understood in the environmental modeling community. In this work we seek a deeper understanding, and begin by considering the work of Jaynes [1976], Box and Tiao [1992] and Bates and Watts [1988].

[8] Jaynes [1976] presents a spirited discussion on estimation of confidence and credible intervals of distribution parameters (e.g., mean of truncated exponential distribution in his Example 5). A similar but more systematic discussion is found in Box and Tiao [1992]. The authors conclude that, given sufficient statistics, what we call confidence and credible intervals are “identical” [Jaynes, 1976, p. 199] or “numerically identical” [Box and Tiao, 1992, p. 86]. Sufficient statistics are completely specified by a defined set of statistics that can be calculated from the data. For example, random variable $\mathbf{X}$ follows a Gaussian distribution with unknown mean θ and variance σ^2 ; then the sample mean and sample variance are sufficient statistics for θ and σ^2 , respectively. In this case, the confidence and credible intervals for the two parameters θ and σ^2 are identical [Box and Tiao, 1992, p. 61]. Box and Tiao [1992, p. 113] extend the discussion to linear models with multivariate Gaussian observation errors, concluding that what are herein called the parameter credible regions are “numerically identical” to the parameter confidence regions used in linear regression [Box and Tiao, 1992, p. 118]. As a result, confidence and credible intervals on predictions are numerically identical. This conclusion can also be found in Bates and Watts [1988, p. 7]. Bates and Watts [1988] further extend the analysis of parameter confidence and credible regions to nonlinear models with multivariate Gaussian observation errors. They conclude that, when intrinsic model nonlinearity is small, the two regions are mathematically similar and so are confidence and credible intervals on parameters and predictions (confidence and credible regions, their relation to confidence and credible intervals, and intrinsic model nonlinearity are discussed in section 2.6).

[9] Issues not considered in the references cited include the extent to which the two kinds of intervals differ from each other when intrinsic model nonlinearity is moderate and large, as is common in models of environmental systems, and resulting relative utility of the confidence intervals in these nonideal circumstances. In addition, informative prior in the Bayesian methods and inclusion of consistent prior information in regression is not considered. Prior information is of considerable utility in environmental models. This paper seeks to provide some insight into these two issues.

[10] This paper first compares confidence and credible intervals theoretically, establishing the foundations developed by Box and Tiao [1992] and Bates and Watts [1988] and extending the theory for linear models to include to

Table 1. List of Selected Recent Studies Comparing Confidence and Credible Intervals^a

Reference and Quantity for Which the Intervals are Constructed	Model	Linear Confidence Interval (LCo)	Nonlinear Confidence Interval (NCo)	Nonlinear Credible Interval (NCr) ^b	Increase in Width Relative to NCr (%) ^c
<i>Vrugt and Bouten</i> [2002], Soil hydraulic parameters	Near linear	X	N/A	X	Similar
<i>Gallagher and Doherty</i> [2007], Watershed Parameters	Nonlinear	XX	N/A	X	L:82%
<i>Finsterle and Pruess</i> [1995], Two-phase flow Parameters	Nonlinear	XXX	X	XX	L:2%; N:–24%
<i>Christensen and Cooley</i> [1999], Groundwater Parameters	Nonlinear	XXX	X	XX	L:71%; N:–7%
<i>Predictions</i>	Nonlinear	X	XX	N/A	–
<i>Liu et al.</i> [2010], ^d DNAPL Parameters	Nonlinear	Not consistently larger or smaller	N/A	N/A	–
<i>This study, Predictions</i>	Nonlinear	Not consistently larger or smaller	N/A	N/A	–
	Linear simple	e	N/A	e	–
	Very nonlinear simple I	X	X	X	Same
	Mildly nonlinear simple II	XXX	XX	X	L:70%; N:13%
	Nonlinear complex ^f	X	X	X	Same
		XX ^g	XXX	X	L: 0.1 to 33% N: 0.2 to 69%
	See Figure 5	X ^h	XX	XXX	L: –18 to –11% N: –15 to –6%

^aThe relative size of the intervals is represented by the number of Xs, with more Xs indicating larger intervals. N/A: interval types not considered. For all, consistent prior is used for the confidence and credible intervals. Except as noted, there is noninformative prior for credible intervals, and no prior for confidence intervals.

^bCalculated using Markov-Chain Monte Carlo.

^cL: $100 \times (\text{LCo} - \text{NCr}) / \text{NCr}$; N: $100 \times (\text{NCo} - \text{NCr}) / \text{NCr}$. Negative values mean the LCo or NCo interval is smaller than the NCr. –, results are presented graphically, so values not determined.

^d*Liu et al.* [2010] calculated linear credible intervals using the maximum a posteriori (MAP) approach. The listing under linear confidence intervals is consistent with results of this work that show the equivalence of linear credible and confidence intervals. Same informative prior is used for linear and nonlinear credible interval.

^eResults of the comparison between LCo and NCr were not consistently larger or smaller for different parameters.

^fThe prior is informative and prior information is included for confidence intervals.

^gApplies for most simulated circumstances.

^hFor drawdown predicted by model INT.

include informative prior for credible intervals and prior information for confidence intervals. In particular, we prove the equivalence when the prior information used for confidence intervals is consistent with the informative prior used for credible intervals; consistency is limited to the informative prior being multi-Gaussian. For nonlinear models, we pursue a largely heuristic approach. We use simple analytical test cases to explore the conditions under which parameter confidence and credible regions and associated prediction confidence and credible intervals are numerically equivalent or nearly equivalent. We further explore computational difficulties that cause differences between the two kinds of intervals in practice. Finally, we use a complex synthetic test case to compare the differences between confidence and credible intervals in the context of differences between alternative models of a single system. The general guidelines drawn for the nonlinear models in this study are expected to be useful to practical environmental modeling.

2. Theoretical Consideration of the Uncertainty Intervals

[11] In this section, conceptual differences between confidence and credible intervals are discussed in section 2.1. In section 2.2, we discuss the situation when linear confidence and credible intervals for linear models are numerically identical. In section 2.3, we show that for linear models with multi-Gaussian observation errors, the two kinds of intervals are numerically identical without priors

(noninformative priors for credible intervals and no priors for confidence intervals), and in section 2.4 with consistent priors (informative priors for credible intervals are consistent with priors defined for confidence intervals). As shown in section 2.5, the same conclusions also apply to linearized-nonlinear models. The relations between confidence and credible intervals are significantly more complicated for nonlinear models. They depend on relations between parameter confidence and credible regions defined in section 2.6. Results identify the consequences of local minima on confidence intervals and inadequate exploration of parameter space for credible intervals.

2.1. Conceptual Differences Between Confidence and Credible Intervals

[12] For a variable X , both confidence and credible intervals can be defined symbolically as

$$\text{Prob}(l \leq X \leq u) = 1 - \alpha, \quad (1)$$

where l and u are the lower and upper interval limits, α is significance level, and $1 - \alpha$ is confidence level. However, the definition is interpreted in different ways when estimating confidence intervals and credible intervals, rendering the two kinds of intervals conceptually different.

[13] Take the intervals for model predictions as an example. From the frequentist point of view, a true value of the prediction exists and a confidence interval with a confidence level of, for example, 95% ($\alpha = 5\%$), is an interval

that is expected to include the true value “95% of the time” in repeated sampling of observations and prior information used in regression [Jaynes, 1976, p. 200; Box and Tiao, 1992, p. 86; McClave and Sincich, 2000, p. 282]. In other words, the interval, not the true value of the unknown prediction, is random. The interval varies with samples used to calculate the interval. If N sets of observations are sampled based on their error distribution and, for each set, the confidence interval of a prediction is evaluated, 95% of the N intervals are expected to contain the true value of the prediction. This concept was used in Hill [1989] and Cooley [1997] to test methods of estimating nonlinear confidence intervals.

[14] From the Bayesian point of view, the prediction is a random variable and a 95% credible interval is expected to include 95% of the probability distribution function (PDF) of the prediction [Box and Tiao, 1992, p. 86; Casella and Berger, 2002]. The posterior distribution summarizes the state of knowledge about the unknown prediction based on available data and prior information. The credible interval is determined via

$$\int_l^u p(g(\beta)|y)dg(\beta) = 1 - \alpha, \quad (2)$$

where β and $g(\beta)$ are model parameters and predictions, respectively, and $p(g(\beta)|y)$ is the posterior distribution of $g(\beta)$ conditioned on data y . In this study, the credible interval limits l and u are determined using the easily computed equal-tailed method as [Casella and Berger, 2002]

$$p(g(\beta) \leq l|y) = p(g(\beta) \geq u|y) = \alpha/2. \quad (3)$$

Other methods of estimating the credible intervals (e.g., highest posterior density (HPD) interval) are also available [Box and Tiao, 1992; Chen and Shao, 1999; Casella and Berger, 2002]. The HPD method is not used in this work partly because it tends to produce the smallest interval and partly simply for the convenience of reporting one set of results.

[15] In this study, we focus our discussion on relations between confidence and credible intervals within the framework of *Gaussian distributed observation errors*, which is the basis of using Gaussian likelihood functions below. The Gaussian distribution is commonly used in environmental problems and is also used in statistical books such as Draper and Smith [1998], Box and Tiao [1992], and Hill and Tiedeman [2007]. Box and Tiao [1992, p. 151] note that approximately Gaussian distributed observations are expected when the central limit theorem applies. This requires that the contributions to observation error be many independent sources, “none of which are dominant.” Considering that in environmental models there are clearly many sources of error (including different sources of measurement, epistemic, and aleatory errors) and that a major goal of model development is to resolve dominant errors, the assumption of Gaussian observation errors is expected to have wide applicability.

2.2. When Linear Confidence and Credible Intervals for Linear Models are Numerically Identical

[16] For linear models, classical regression confidence intervals of *fixed but unknown* true model parameters β are estimated based on the distributions $p(\hat{b})$ of *parameter estimates* $\hat{b}$, which are functions of the observations y . Bayesian

credible intervals of *random* model parameters β are evaluated based on the distributions $p(\beta|y)$ of *parameters* β , which are calculated according to the Bayes’ theorem,

$$p(\beta|y) = \frac{p(y|\beta)p(\beta)}{p(y)} = \frac{l(\beta|y)p(\beta)}{p(y)}. \quad (4)$$

If $p(\beta|y)$ is equivalent to $p(\hat{b})$, the credible and confidence intervals of β and the intervals of linear model predictions $g(\beta)$ are numerically equivalent. Jaynes [1976] and Box and Tiao [1992] established this equivalence using the concept of sufficient statistics without prior information (noninformative prior in (4) and no prior used for the confidence interval).

[17] According to Box and Tiao [1992, p. 62], for a vector of observations y whose distribution depends on the parameter vector β , the set of statistics T is said to be jointly *sufficient* for β if the likelihood function $l(\beta|y)$ can be expressed in the form

$$l(\beta|y) \propto l_T(\beta|T) \propto p(T|\beta), \quad (5)$$

where T is a set of statistics calculated from y that fully define the likelihood function and the second relation of equation (5) uses the basic relation between likelihood and probability. If y is multivariate Gaussian distributed with mean simulated by a linear model with model parameters β and known variance, the corresponding Gaussian likelihood function $l(\beta|y)$ can be fully defined by the set of sufficient statistics $T = \hat{b}$ so that $l(\beta|y) \propto p(\hat{b}|\beta)$ [Box and Tiao, 1992, p. 115]. If non-informative prior is used in (4), $p(\beta|y) \propto l(\beta|y)$ and thus $p(\beta|y) \propto p(\hat{b}|\beta)$ [Box and Tiao, 1992, p. 115]. This is the basis of the conclusion of Jaynes [1976, p. 199] that, “if the confidence interval is based on a sufficient statistic, ... it turns out to be so nearly equal to the Bayesian interval that it is difficult to produce any appreciable difference in the numerical results; in an astonishing number of cases, they are identical. ... Similarly, the shortest confidence interval for the mean of a normal distribution, whether the variance is known or unknown, ... turn out to be identical with the shortest Bayesian intervals at the same level (based on a uniform prior density for location parameters and the Jeffreys prior $d\sigma/\sigma$ for scale parameters).”

[18] Building on the theoretical foundation of Jaynes [1976] and Box and Tiao [1992], we next prove that linear confidence and credible intervals are numerically identical for the Gaussian likelihood function used widely in regression and Bayesian analysis in environmental modeling. We consider two cases, without and with informative prior information. To our knowledge, the derivation of the equivalence using informative prior information is new. Confidence intervals with prior information and credible intervals with informative prior were discussed by Hill [2010]. Doherty and Hunt [2010] and Fienen et al. [2010] suggested equivalence in limited circumstances. The proof of equivalence presented in this work is more theoretically rigorous and shows the equivalence to be more general than the previous works. As a corollary, we point out that linear confidence and credible intervals are also numerically identical when calculated for linearized-nonlinear models.

2.3. Equivalence of Regression and Bayesian Distributions and Intervals for Gaussian Linear Models

[19] In classical regression, a linear model without prior information on parameters is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (6)$$

with

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{C}_\varepsilon), \quad (7)$$

where transformations can be applied to achieve the Gaussian distribution of (7). The covariance matrix $\mathbf{C}_\varepsilon$ is often related as $\mathbf{C}_\varepsilon = \sigma^2 \boldsymbol{\omega}^{-1}$ to weighting $\boldsymbol{\omega}$ (used in regression) and a scalar σ^2 (generally unknown but can be estimated). When σ^2 is known, the distribution of estimates $\hat{\mathbf{b}}$ of the unknown true parameters $\boldsymbol{\beta}$ is multivariate Gaussian, i.e., [Draper and Smith, 1998, p. 94; Seber and Lee, 2003, p. 47],

$$\hat{\mathbf{b}} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X})^{-1}) \propto \exp \left[-\frac{1}{2} (\hat{\mathbf{b}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X} (\hat{\mathbf{b}} - \boldsymbol{\beta}) \right]. \quad (8)$$

Ideally $\mathbf{X}$ in (8) is the same as in (6), but in reality alternative models may have different $\mathbf{X}$ (more discussion see Appendix C of Hill and Tiedeman [2007]).

[20] For a linear prediction function $g(\boldsymbol{\beta})$ with $g(\boldsymbol{\beta}) = \mathbf{Z}\boldsymbol{\beta}$, it has

$$g(\hat{\mathbf{b}}) \sim N(g(\boldsymbol{\beta}), \mathbf{Z}^T (\mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X})^{-1} \mathbf{Z}) \propto \exp \left[-\frac{1}{2} (g(\hat{\mathbf{b}}) - g(\boldsymbol{\beta}))^T (\mathbf{Z}^T (\mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X})^{-1} \mathbf{Z})^{-1} \cdot (g(\hat{\mathbf{b}}) - g(\boldsymbol{\beta})) \right]. \quad (9)$$

Correspondingly, the $(1 - \alpha) \times 100\%$ confidence interval for $g(\boldsymbol{\beta})$ is given as

$$g(\hat{\mathbf{b}}) \pm z_{1-\alpha/2} \left[\mathbf{Z}^T (\mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X})^{-1} \mathbf{Z} \right]^{1/2}, \quad (10)$$

where $z_{1-\alpha/2}$ is a z statistic of standard normal distribution with significance level α . When σ^2 is unknown and estimated using $s^2 = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T \boldsymbol{\omega} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) / (n - p)$, the $(1 - \alpha) \times 100\%$ confidence interval of $g(\boldsymbol{\beta})$ is

$$g(\hat{\mathbf{b}}) \pm t_{1-\alpha/2, n-p} \left[s^2 \mathbf{Z}^T (\mathbf{X}^T \boldsymbol{\omega} \mathbf{X})^{-1} \mathbf{Z} \right]^{1/2}, \quad (11)$$

where $t_{1-\alpha/2, n-p}$ is a t statistic with significance level α and freedom $n-p$, where n is the number of observations and p is the number of parameters.

[21] In the Bayesian analysis, according to (4), with the same model defined in (6) and (7), considering Jeffreys' noninformative prior defined in Box and Tiao [1992, pp. 41–54], it is derived in Appendix A in the auxiliary material¹ that, when σ^2 is known, the posterior distribution of $\boldsymbol{\beta}$ is multivariate Gaussian, i.e.,

$$\boldsymbol{\beta} \sim N_p(\hat{\mathbf{b}}, (\mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X})^{-1}) \propto \exp \left[-\frac{1}{2} (\boldsymbol{\beta} - \hat{\mathbf{b}})^T \mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X} (\boldsymbol{\beta} - \hat{\mathbf{b}}) \right], \quad (12)$$

where $\hat{\mathbf{b}}$ is the maximum likelihood estimate (also the least squares estimate) of $\boldsymbol{\beta}$. Correspondingly, the posterior distribution of a linear prediction $g(\boldsymbol{\beta})$ (i.e., $g(\boldsymbol{\beta}) = \mathbf{Z}\boldsymbol{\beta}$) is multivariate Gaussian,

$$g(\boldsymbol{\beta}) \sim N_p(g(\hat{\mathbf{b}}), \mathbf{Z}^T (\mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X})^{-1} \mathbf{Z}) \propto \exp \left[-\frac{1}{2} (g(\boldsymbol{\beta}) - g(\hat{\mathbf{b}}))^T (\mathbf{Z}^T (\mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X})^{-1} \mathbf{Z})^{-1} \cdot (g(\boldsymbol{\beta}) - g(\hat{\mathbf{b}})) \right]. \quad (13)$$

[22] Comparing equations (12) and (13) with equations (8) and (9), respectively, shows that the distributions of $\hat{\mathbf{b}}$ and $g(\hat{\mathbf{b}})$ are equivalent to those of $\boldsymbol{\beta}$ and $g(\boldsymbol{\beta})$. Therefore, the $(1 - \alpha) \times 100\%$ credible interval of $g(\boldsymbol{\beta})$ is equivalent to (10) if σ^2 is known, and (11) if σ^2 is unknown. In other words, for linear models without prior information and Gaussian distributed observation errors, the linear confidence and credible intervals are numerically identical.

2.4. Equivalence When Including Informative Prior in Bayesian Equations and Prior Information in Regression for Gaussian Linear Models

[23] The equivalence between confidence and credible intervals can also be obtained with consideration of prior information, when it is available, in Bayesian and regression analysis. Regression comes from a frequentist background, and the idea of adding prior information can sometimes seem strange. For example, Stark and Tenorio [2011] and Kitaniadis [2010] suggest that only Bayesian methods include prior, while, for example, Schweppe [1973], Cooley [1983], Hill and Tiedeman [2007], and Poeter et al. [2005] have used prior within the context of regression methods for decades. The Maximum a Posteriori (MAP) method [Carrera and Neuman, 1986; Oliver et al., 2008; Liu et al., 2010] is essentially equivalent to the regression-with-prior method used in this work, but was developed from the Bayesian perspective.

[24] In the Bayesian analysis, for the model defined in (6) and (7), when prior parameter distributions are informative, a conventional form is the conjugate prior that is multivariate Gaussian with mean $\boldsymbol{\beta}_p$ and covariance matrix $\mathbf{C}_p$ [Kitaniadis, 1986, 1997]. For this prior, it is derived in Appendix B of the auxiliary material that the $(1 - \alpha) \times 100\%$ credible interval for linear model prediction, $g(\boldsymbol{\beta}) = \mathbf{Z}\boldsymbol{\beta}$, is

$$g(\boldsymbol{\beta}'_p) \pm z_{1-\alpha/2} \left[\mathbf{Z}^T (\mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X} + \mathbf{C}_p^{-1})^{-1} \mathbf{Z} \right]^{1/2}, \quad (14)$$

where $\boldsymbol{\beta}'_p = (\mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{X} + \mathbf{C}_p^{-1})^{-1} (\mathbf{X}^T \mathbf{C}_\varepsilon^{-1} \mathbf{y} + \mathbf{C}_p^{-1} \boldsymbol{\beta}_p)$ is mean of the Bayesian posterior distribution.

[25] In regression analysis, the prior information is defined as knowledge about model parameters estimated by the regression procedure [Schweppe, 1973, p. 104; Cooley, 1983; Hill and Tiedeman, 2007, p. 288], as might be derived from measurements or expert knowledge of transmissivity or hydraulic conductivity, recharge or discharge, specified boundary flows and heads. This knowledge is commonly included in the regression to complement the observations $\mathbf{y}$ of state variables. This is accomplished by treating the prior information as the measurement vector $\mathbf{y}\boldsymbol{\beta}$ of the estimated model parameters with error vector $\boldsymbol{\varepsilon}_\beta$ and by appending $\mathbf{y}_\beta$

¹Auxiliary materials are available in the HTML. doi:10.1029/2011WR011289.

to observation vector $\mathbf{y}$ and $\boldsymbol{\varepsilon}_\beta$ to the observation error vector $\boldsymbol{\varepsilon}$ in equation (6). For example, when the prior information has the form $\mathbf{y}_\beta = \boldsymbol{\beta} + \boldsymbol{\varepsilon}_\beta$ with the errors defined as $\boldsymbol{\varepsilon}_\beta \sim N_{n_{pri}}(\mathbf{0}, \mathbf{C}_\beta)$, it is derived in Appendix B of the auxiliary material that the $(1 - \alpha) \times 100\%$ confidence interval for linear model prediction, $g(\boldsymbol{\beta}) = \mathbf{Z}\boldsymbol{\beta}$, is equivalent to (14) when the prior information $\mathbf{y}_\beta$ is equal to the mean of Bayesian prior distribution $\boldsymbol{\beta}_p$ and the covariance of errors of the prior information $\mathbf{C}_\beta$ is equal to the covariance of the prior distribution $\mathbf{C}_p$. For a Gaussian distribution, this yields consistent priors in the two methods, even though conceptually the prior information in regression and Bayesian analysis are different. In classical regression context, the parameters $\boldsymbol{\beta}$ are taken as fixed and unknown and the randomness of the prior estimates in $\mathbf{y}_\beta$ is from the error $\boldsymbol{\varepsilon}_\beta$; in Bayesian analysis, $\boldsymbol{\beta}$ is conceptualized as being random [Kitanidis, 2010, among others]. For linear models with consistent priors, the linear confidence and credible intervals are numerically identical.

2.5. Equivalence of Linear Confidence and Credible Intervals for Linearized-Nonlinear Models

[26] In classical regression, the linear confidence interval is also available for a nonlinear model,

$$\mathbf{y} = f(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \quad (15)$$

with $\boldsymbol{\varepsilon}$ defined in equation (7) when prior information is not available. To estimate a linear confidence interval, the nonlinear function $f(\mathbf{b})$ of model parameters $\mathbf{b}$ (a general vector of model parameters) is approximated by its first-order Taylor series expansion in the neighborhood of $\hat{\mathbf{b}}$, the generalized least squares estimator of $\boldsymbol{\beta}$. That is, $f(\mathbf{b}) \approx f(\hat{\mathbf{b}}) + \mathbf{X}_{\hat{\mathbf{b}}}(\mathbf{b} - \hat{\mathbf{b}})$, where $\mathbf{X}_{\hat{\mathbf{b}}} = [\partial f / \partial \mathbf{b}]_{\mathbf{b}=\hat{\mathbf{b}}}$ is the sensitivity matrix [Seber and Wild, 2003, p. 23–24]. If the nonlinear prediction function $g(\mathbf{b})$ is also approximated to the first-order with the prediction sensitivity vector $\mathbf{Z}_{\hat{\mathbf{b}}} = [\partial g / \partial \mathbf{b}]_{\mathbf{b}=\hat{\mathbf{b}}}$ [Seber and Wild, 2003, p. 192], then for σ^2 known, the $(1 - \alpha) \times 100\%$ linear confidence interval of $g(\boldsymbol{\beta})$ is similar to (10) but with $\mathbf{X}$ replaced by $\mathbf{X}_{\hat{\mathbf{b}}}$ and $\mathbf{Z}$ replaced by $\mathbf{Z}_{\hat{\mathbf{b}}}$. For σ^2 unknown, the $(1 - \alpha) \times 100\%$ linear confidence interval of $g(\boldsymbol{\beta})$ is similar to (11) but with $\mathbf{X}$ replaced by $\mathbf{X}_{\hat{\mathbf{b}}}$ and $\mathbf{Z}$ replaced by $\mathbf{Z}_{\hat{\mathbf{b}}}$.

[27] In the Bayesian analysis, the linear credible interval for the nonlinear model is evaluated after the nonlinear model is linearized. As shown in Appendix C of the auxiliary material, the derived linear credible intervals are numerically identical to the linear confidence intervals for nonlinear models. Though for a linearized model confidence and credible intervals may be numerically identical, if the linearization method is not adequate for approximating confidence and credible intervals for the parameters and predictions of a nonlinear model, they will both be equivalently wrong. Linearization tends to become progressively less adequate as the model becomes more nonlinear. More discussion on how model nonlinearity affects the above approximations is discussed by Hill and Tiedeman [2007, p. 393–398].

2.6. Relations Between Nonlinear Confidence and Credible Intervals for Nonlinear Models

[28] The relations between nonlinear confidence and credible intervals of nonlinear models are complicated. In classical

regression, the confidence intervals of model parameters and predictions are not estimated based on the distributions of their estimates (e.g., equations (8) and (9)) as for linear models, because analytical expressions of the distributions are in general not available. Instead, the nonlinear confidence intervals are evaluated using parameter confidence regions as described below. To facilitate the discussion, we also use posterior parameter credible regions (defined below) to evaluate credible intervals. The step of determining posterior parameter credible regions is rare because credible intervals are generally calculated directly from posterior distributions of model parameters and predictions. However, a parameter credible region can be determined and is used here to relate prediction confidence and credible intervals.

[29] The nonlinear confidence and credible intervals can be numerically identical if (1) the interval limits are equivalent to the maximum and minimum values of prediction function $g(\mathbf{b})$ on the boundary of the parameter confidence and credible regions and (2) the two parameter regions are equivalent [Vecchia and Cooley, 1987; Cooley 1993a, 1993b, 1999]. The discussions below for the two conditions involve definitions of exact confidence and credible regions and approximate regions defined by the likelihood method. Based on Vecchia and Cooley [1987], the exact $(1 - \alpha) \times 100\%$ confidence region for $\boldsymbol{\beta}$ is a region in p dimensional Euclidean space, say, R_α , that depends on $\mathbf{y}$, where $\mathbf{y}$ includes observations and possible prior information, and for which $p(\boldsymbol{\beta} \in R_\alpha) = 1 - \alpha$ holds exactly in regression theory. The exact $(1 - \alpha) \times 100\%$ posterior credible region is bounded by a contour of the posterior density function within which the posterior probability of model parameters is $(1 - \alpha) \times 100\%$ exactly [Bates and Watts, 1988, p. 220]. However, estimating the exact region is computationally difficult, and the approximate likelihood method is commonly used.

[30] The likelihood confidence region is defined as the set of parameter values whose corresponding objective function values, $S(\mathbf{b})$, satisfy [Christensen and Cooley, 1999; Cooley, 2004; Hill and Tiedeman, 2007, p. 178]

$$S(\mathbf{b}) \leq S(\hat{\mathbf{b}}) \left[\frac{1}{n-p} t_{\alpha/2, n-p}^2 + 1 \right]. \quad (16)$$

The confidence region defined in (16) is for calculating individual confidence intervals. Scheffé-type simultaneous confidence intervals are calculated based on the region with $S(\mathbf{b})$ satisfying

$$S(\mathbf{b}) \leq S(\hat{\mathbf{b}}) \left[\frac{p}{n-p} F_{\alpha/2, p, n-p} + 1 \right], \quad (17)$$

which is equivalent to (16) if $p = 1$. The test cases in this work use individual confidence intervals calculated using (16). The parameter region defined in (16) contains the true model parameters $\boldsymbol{\beta}$ with approximate probability of $(1 - \alpha) \times 100\%$ for errors defined in (7) when prior information is not available and σ^2 is unknown. When intrinsic model nonlinearity is small, it contains the true model parameters with exact $(1 - \alpha) \times 100\%$ probability [Donaldson and Schnabel, 1987; Bates and Watts, 1988, p. 201].

[31] The likelihood credible region is also bounded by the contours of (16) and (17) as follows. When intrinsic model nonlinearity is small, consider independent Jeffrey's

noninformative priors for parameters β and σ , i.e., $p(\beta) \propto C$ and $p(\sigma) \propto 1/\sigma$ (or $p(\beta, \sigma) \propto 1/\sigma$). For errors defined in (7), the resulting likelihood function for parameters β and σ is

$$l(\beta, \sigma | \mathbf{y}) = p(\mathbf{y} | \beta, \sigma) \\ = \frac{1}{(2\pi\sigma^2)^{n/2}} |\omega|^{1/2} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{y} - f(\beta))^T \omega (\mathbf{y} - f(\beta)) \right] \quad (18)$$

where $S(\beta) = (\mathbf{y} - f(\beta))^T \omega (\mathbf{y} - f(\beta))$ is the objective function, same as $S(\mathbf{b})$ used in regression. Correspondingly, the joint posterior distribution of parameters β and σ is

$$p(\beta, \sigma | \mathbf{y}) \propto p(\mathbf{y} | \beta, \sigma) \sigma^{-1} \\ \propto \sigma^{-(n+1)} \exp \left(-\frac{1}{2\sigma^2} S(\beta) \right). \quad (19)$$

Integrating out parameter σ leads to the posterior distribution of model parameters β [Bates and Watts, 1988, p. 220],

$$p(\beta | \mathbf{y}) = \int_0^\infty p(\beta, \sigma | \mathbf{y}) d\sigma \propto [S(\beta)]^{-n/2}, \quad (20)$$

and a credible region is bounded by a contour of the posterior density function or equivalently by a contour of the objective function $S(\beta)$. Bates and Watts [1988, p. 220] proved that the approximate $(1 - \alpha) \times 100\%$ individual posterior credible region is bounded by the $S(\mathbf{b})$ contour defined in (16).

[32] Thus, when observation errors are multivariate Gaussian distributed and the intrinsic model nonlinearity is very small, without prior information, the confidence and credible regions are equivalent and so are the nonlinear confidence and credible intervals of predictions evaluated on the regions. When consistent prior information is used for regression and Bayesian analysis in the way described in section 2.4, it is straightforward to develop the equivalence between the parameter confidence and credible regions. As a result, the confidence and credible intervals are numerically identical.

[33] Regions defined by the above approximate likelihood methods and the calculated confidence and credible intervals are accurate given the following assumptions:

[34] 1. The model accurately represents the system;

[35] 2. Model prediction $g(\mathbf{b})$ is monotonic enough that any local extreme of $g(\mathbf{b})$ within the closed parameter regions lies between the maximum and minimum values of $g(\mathbf{b})$ that occur along the boundary of the regions [Cooley, 1993a];

[36] 3. There is a single minimum in the objective function;

[37] 4. The residuals (defined as the difference between observed and simulated values) are multivariate Gaussian distributed to obtain a valid critical value, and

[38] 5. Model intrinsic nonlinearity is small [Vecchia and Cooley, 1987; Christensen and Cooley, 1999; Cooley, 2004].

[39] The assumptions can be evaluated using existing techniques such as those described in Hill and Tiedeman [2007] and shown below. The model intrinsic nonlinearity of assumption 5, when combined with parameter effects nonlinearity, results in total model nonlinearity [Christensen and Cooley, 1999]. Parameter effects nonlinearity is what can be

removed after transforming model parameters in suitable ways, while intrinsic model nonlinearity results from model structure and cannot be removed by any parameter transformations. Cooley [2004] and Christensen and Cooley [2005] developed methods of measuring model total nonlinearity based on Beale [1960], Linszen [1975], and Bates and Watts [1980]. More discussion of model nonlinearity measures can be found in Hill and Tiedeman [2007, p. 142–145]. The residuals mentioned in assumption 4 are expected to have a variance-covariance matrix equal to $(\mathbf{I} - \mathbf{X}(\mathbf{X}^T \omega \mathbf{X})^{-1} \mathbf{X}^T \omega)$, so are not expected to be independent even when the elements of ϵ are independent. [Cook and Weisberg, 1982, p. 11; Cooley and Naff, 1990; Hill and Tiedeman, 2007, p. 111–113; Aster et al., 2012; Finsterle and Zhang, 2011]. This also applies when equations (6) and (7) are augmented to include prior information.

[40] With recently developed Markov chain Monte Carlo (MCMC) techniques, the nonlinear credible intervals can be evaluated directly without relying on the concept of parameter credible region. The reason is that posterior parameter distributions can be directly simulated and that the credible intervals of predictions can be evaluated by first sorting the prediction samples from the smallest to the largest and then identifying, for example, the 2.5% and 97.5% thresholds to form 95% individual credible intervals of the predictions. However, the estimated credible interval based on the MCMC results may be inaccurate if the MCMC sampling does not correctly simulate the posterior distribution, as discussed below. Therefore, when comparing the nonlinear confidence intervals determined from a confidence region and credible intervals from MCMC methods, the assumptions of calculating accurate confidence intervals need to be examined and the MCMC sampling process needs to be checked. This is further demonstrated in two numerical examples presented in section 3.

[41] Two MCMC codes are used in this study. One is MICA developed by Doherty [2003]. MICA uses the Metropolis-Hastings algorithm revised by introducing a simple adaptive algorithm of the covariance matrix of the proposal distribution to obtain a reasonable acceptance rate [Haario et al., 2001]. When the acceptance rate is high, all elements of the covariance matrix are simultaneously increased by a user-defined multiplier; and when the acceptance rate is low, all elements are simultaneously decreased. While MICA works well for estimating *unimodal* posterior parameter distributions, for problems with multimodal parameter distributions, it cannot sample the distribution efficiently with a single proposal distribution, because it does not have algorithms for chain jumps among multiple modes. As pointed out by Gallagher and Doherty [2007] and also found in this study, for problems with multiple minima the probability density functions obtained from MICA are not accurate in that the resulting density functions do not have multiple modes corresponding to the minima. To solve this problem, DREAM of Vrugt et al. [2008, 2009] is used. For the results presented here, the algorithm was implemented in FORTRAN and parallelized. DREAM automatically tunes the scale and orientation of the proposal distribution in the sampling process by using multiple different chains simultaneously for global exploration. It uses the current location of the chains to generate candidate points, allowing the jumps between modes of parameter distributions, and thus efficiently accommodates complex and multimodal

Table 2. Predictions and 95% Linear Confidence and Credible Intervals for Linear Simple Test Function^a

	Prediction	Lower Limit	Upper Limit
Noninformative Prior			
Unknown σ^2 , equation (11)			
Linear confidence interval	63.76	62.26	65.26
Linear credible interval (MCMC)	63.76	62.27	65.27
Known σ^2 , equation (10)			
Linear confidence interval	63.76	62.37	65.16
Linear credible interval (MCMC)	63.76	62.37	65.19
Informative Prior ^b			
Linear confidence interval	63.73	62.73	64.74
Linear credible interval (MCMC)	63.73	62.72	64.75

^aTrue value of the prediction is 63. MCMC results are obtained with MICA [Doherty, 2003] using 1,000,000 model runs. Linear confidence intervals require 10 model runs).

^bInformative Prior with $C_\beta = C_p = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$, equation (14).

parameter posterior distributions. The test cases below illustrate the consequences of this difference between MICA and DREAM.

3. Simple Test Cases

[42] Three simple test cases with one linear and two nonlinear analytic functions are designed to demonstrate the theoretical findings above and to help better understand the similarities and differences between confidence and credible intervals. These test cases are not derived from environmental problems, but are easily designed to produce linearity or desired kinds of relevant nonlinearities. In the next section, a complicated groundwater problem is explored.

3.1. Linear Test Function

[43] The linear test function is

$$y = ax + m + \epsilon, \quad (21)$$

where the true values are $a = 2$ and $m = 3$. Twenty samples of y are first generated with $x = \{1, 2, \dots, 20\}$, and subsequently corrupted using one realization of white noise ϵ , with mean zero and constant variance $\sigma^2 = 1$. Linear confidence and credible intervals, calculated for y at $x = 30$ with unknown and known σ^2 , and noninformative and informative priors, are listed in Table 2, where the linear credible intervals are estimated from 500,000 MCMC parameter samples. Except for negligible numerical discrepancy, the table confirms that the linear confidence and credible intervals are numerically identical for this linear problem in all three situations. Figures 1a–1c show probability density functions (PDFs) of the two parameters and prediction based on the classical regression theory (equations (8) and (9)) and Bayesian theory (equations (12) and (13)). This simple numerical test case demonstrates the theoretical analysis in section 2.3 that confidence and credible intervals for linear models are numerically identical. The conclusions would also be expected to apply to linearized-nonlinear models.

3.2. Very Nonlinear Test Function I

[44] The nonlinear test function is

$$y = x/a + \sin(amx) + \epsilon \quad (22)$$

where $a = 2$, $m = 0.1$. Twenty samples of y are first generated with $x = \{1, 2, \dots, 20\}$, and subsequently corrupted using one realization of white noise ϵ , with mean zero and constant variance $\sigma^2 = 1$. This nonlinear function is designed to be nonlinear and to have local minima. For this nonlinear function, linear and nonlinear confidence intervals with no prior information and credible intervals with noninformative priors are calculated for y at $x = 30$ with known σ^2 . Table 3 shows that the linear confidence and credible intervals are numerically the same but differ from the nonlinear intervals. Compared to the previous studies listed in Table 1, our results are similar to those of the nonlinear test case of *Vrugt and Bouten* [2002] in that the linear confidence intervals are larger than the nonlinear credible intervals. Our results differ from those of *Gallagher and Doherty* [2007] in that the nonlinear credible interval is smaller than the nonlinear confidence interval. Although sufficient numbers of MCMC simulations were performed, the nonlinear credible intervals of MICA and DREAM differ (Table 3).

[45] To understand better the linear intervals for nonlinear models, Figures 1d–1f were constructed to show the linear confidence and credible intervals. Similar to Figures 1a–1c, Figures 1d–1f plot the PDFs of parameters and prediction based on classical regression theory and Bayesian theory for the linearized model obtained from 500,000 MCMC samples. Again, the figures show that the classical regression (red curve) and Bayesian (blue curve) distributions are equivalent and the linear confidence (red bar) and credible (blue bar) intervals are coincident. This numerically confirms the theoretical analysis in section 2.5 for linear intervals calculated for a nonlinear problem.

[46] Figures 1d–1f also show nonlinear credible intervals. The figures show that PDFs of the linearized model (blue curves) are significantly different from those of the nonlinear model (green curves) obtained from 500,000 MCMC samples. While the PDFs of model parameters are Gaussian for the linearized model and almost Gaussian for the nonlinear model, the PDF of the model prediction for the nonlinear model (Figure 1f) is narrower and taller. This may be attributed to model nonlinearity: total and intrinsic nonlinearities of this model are 0.54 and 0.20, respectively, so that for both this model is rated as nonlinear on the four-tiered scale as shown in footnote of Table 6.

[47] Of the intervals considered, the MCMC credible intervals should provide the most accurate assessment of uncertainty. Model nonlinearity is accounted for to some degree by nonlinear confidence intervals, and Table 3 shows that for this problem the nonlinear confidence intervals are closer to the nonlinear credible intervals than are the linear intervals. The nonlinear confidence intervals are not included in Figure 1, because the underlying theory does not include definition of a PDF. Here, the nonlinear confidence intervals are evaluated on the confidence region. The plotted 95% parameter confidence region is defined using (16) and shown as the black ellipse with objective function value of 18.3 in Figure 2a. Figure 2a also shows prediction contours, and the intersection of two red prediction contours with the black ellipse define the limits (the two red dots) of the nonlinear confidence interval on the prediction with values shown in Table 3. For this nonlinear function defined in (22), the prediction is not monotonic with respect to the parameter values. While, for some orientations of the objective function contours, the shape of the prediction contours could cause the

Figure 1. Probability density functions (PDFs) of parameters (a and m) and prediction (y) based on regression and Bayesian theories for the (a–c) linear and (d–f) very nonlinear simple test functions. Figures 1d–1f show results for linearized and nonlinear models. In each plot, the bars correspond to the 95% interval limits. The limits of the linear confidence (red) and credible (blue) intervals overlap for the linear model (Figures 1a–1c) and the linearized model (Figures 1d–1f). The PDFs and intervals are centered on estimated values.

calculation of nonlinear confidence intervals to produce nonunique results, Figure 2a clearly shows that the intersection of the objective function contour (the black ellipse) and the relevant prediction contours (the two red lines) produce unique maximum and minimum predictions (the two red dots). This analysis is consistent with the good performance of UCODE_2005 [Poeter *et al.*, 2005] in finding the nonlinear confidence interval for this problem.

[48] To better illustrate the relation between the confidence and credible intervals, the 500,000 MCMC parameter samples obtained from MICA are plotted in Figure 2a. MICA results in Figure 2a use a multivariate Gaussian proposal distribution with the covariance matrix calculated for the optimal parameters. The plotting procedure is consistent with how the credible interval is calculated, as follows. First, the MCMC parameter samples are ordered using the prediction

Table 3. Predictions and 95% Linear and Nonlinear Confidence and Credible Intervals for Very Nonlinear Simple Test Function I^a

Type of Interval	Prediction	Lower Limit	Upper Limit	Interval Width	Change in Width Relative to NCrDr(%)
Linear confidence	14.83	10.61	19.05	8.44	65% larger
Linear credible	14.83	10.61	19.05	8.44	65% larger
Nonlinear confidence	14.83	12.80	18.42	5.62	10% larger
Nonlinear credible (MICA)	14.82	13.12	18.09	4.97	3% smaller
Nonlinear credible (DREAM) (NCrDr)	14.84	13.04	18.15	5.11	–

^aTotal and intrinsic model nonlinearities are 0.54 and 0.20, respectively, indicating this is a very nonlinear model. True value of the prediction is 17.2. Intervals are for known σ^2 . Results with unknown σ^2 would be similar except that all intervals would be larger.

Figure 2. The 95% individual nonlinear confidence and credible intervals for (a and b) nonlinear test function I (dots in Figure 1a are samples from MICA; dots in Figure 1b are from DREAM) and (c) test function II (dots are from DREAM and similar to those from MICA). The black ellipses for an objective function value of 18.3 in (a) and (b) and of 34.5 in (c) are 95% confidence regions defined in equation (16). MCMC samples are plotted as dots: the upper 2.5% and lower 2.5% of predictions are green; the middle 95% of predictions are yellow.

values they produced, starting with the parameter set that produces the smallest value of the prediction and progressing to the parameter set that produced the largest value of the prediction. The green dots at the bottom right corner correspond to the parameter sets that produce the smallest 2.5% of the predictions, and those at the top left corner to the parameter sets that produce the largest 2.5% of the predictions. The yellow dots correspond to the parameter sets that produce the middle 95% of the predictions. The nonlinear credible interval limits equal the predictions that fall exactly at the 2.5% levels. They are identified in Figure 2a by blue dots and the values are listed in Table 3. The two blue lines are contours for the prediction values represented by the two blue dots (i.e., the nonlinear credible interval limits).

[49] Figure 2a shows that the nonlinear confidence interval is larger than the credible interval (the distance between the two red lines is larger than that between the two blue lines), but numerically they should be identical if all the assumptions are satisfied to calculate the confidence interval and the posterior distribution is correctly sampled by MCMC in the calculation of the credible interval. To better understand the discrepancy, Figure 2b was constructed similarly to Figure 2a but using results from DREAM. Figure 2b indicates that there are two local minima (black stars) of the objective function in addition to the one detected by nonlinear regression and MICA (black plus) and the parameter values corresponding to the minima differ from each other substantially, as shown in Table 4. The nonlinear confidence intervals defined by (16) are evaluated on the objective function contours centered on a minimum. For the three different minima, the objective function contours defined by (16) are different and so are the nonlinear confidence intervals, as shown in Table 4.

[50] Most likely, local minima affect estimation of confidence and credible intervals centered at the global minimum more as the values of the local minima approach the value of the global minimum. In Figure 2b the objective function contours defined by (16) centered at the local minima are larger than that of 18.3 centered at the global minimum, and the effect on calculated confidence intervals centered at the global minimum is modest (10% larger as shown in Table 3 and 4). MICA only samples the region surrounding the global objective function minimum (black plus in Figure 2a). DREAM successfully identifies the multimodal target parameter posterior distribution as shown in Figure 2b. Figure 2b indicates that, because of the local minima, the percent of samples covered by the 95% parameter confidence region as shown in Figure 2a is actually less than 95%. In other words, because (16) uses the statistics corresponding to a 95% confidence level to define the confidence region, the calculated nonlinear confidence interval on the prediction is larger than what it should be. It appears that this would be typical in the presence of multimodal posterior PDFs for confidence intervals constructed using the global minimum. Table 3 indicates that the more accurate nonlinear credible interval from DREAM is a little smaller than the nonlinear confidence interval but larger than the incorrect nonlinear credible interval from MICA produced using a single proposal distribution.

3.3. Mildly Nonlinear Test Function II

[51] Nonlinear test function I gives different nonlinear confidence and credible intervals because of unsatisfied

Table 4. Optimal Parameter Values, Minimum Objective Function Values, and Corresponding Prediction Uncertainty Intervals for Very Nonlinear Simple Test Function I^a

Type of minima	Parameter Value		Objective Function	Prediction	Lower Limit	Upper Limit	Interval Width	Change in Width Relative to NCrDr (%) ^c
	<i>a</i>	<i>m</i>						
Local 1	1.86	0.00046	23.1	16.16	13.89 ^b	18.40 ^b	4.51 ^b	12% smaller
Local 2	1.83	0.305	24.9	15.50	14.07 ^b	17.55 ^b	3.48 ^b	32% smaller
Global	1.94	0.096	14.7	14.83	12.80 ^b	18.42 ^b	5.62 ^b	10% larger
Global ^a	1.94	0.098	14.7	14.84	13.04 ^c	18.15 ^c	5.11 ^c	–

^aThe first two rows correspond to two local minima marked as asterisks in Figure 2b; the third row corresponds to global minimum marked as “plus” in Figure 2a; the last row corresponds to the local minimum marked as “plus” in Figure 2b.

^bNonlinear confidence intervals.

^cNonlinear credible interval (DREAM) (NCrDr).

assumptions for the calculation of the confidence interval: there are important multiple minima in the objective function and the intrinsic model nonlinearity is high. Here, we used another nonlinear test function to evaluate whether the calculated nonlinear confidence and credible intervals are closer if the assumptions are satisfied to a greater degree: there are no competing local minima and the intrinsic nonlinearity is smaller. This nonlinear test function, adopted from *Draper and Smith* [1998, p. 475], is

$$y = a + (0.49 - a)e^{-m(x-8)} + \epsilon \quad (23)$$

where $a = 0.4$, $m = 0.1$. Forty samples of y are first generated with $x = \{8, 9, \dots, 47\}$, and subsequently corrupted using one realization of white noise ϵ , with mean zero and constant variance $\sigma^2 = 0.01$. For this nonlinear function, nonlinear confidence intervals and credible intervals are calculated for y at $x = 50$.

[52] As expected, the nonlinear confidence and credible intervals are almost the same (Table 5). Figure 2c explores their relations by plotting the confidence interval calculated based on equation (16) and the credible interval determined from DREAM parameter samples based on the equal-tailed method. The coincidence of the red and blue curves confirms the similarity between nonlinear confidence and credible intervals. The reason is that the parameter confidence region is identical to its credible region. As shown in Figure 2c, the 95% individual confidence region is delineated by the black objective function contour with value of 34.5 calculated using equation (16). This contour corresponds to the 95% individual parameter credible region because the probability of each individual parameter within or on the boundary of the region is 95%. This conclusion is drawn in the following procedure for the two parameters a and m . Based on the 95% individual confidence region (the black ellipse), the minimum and maximum values of parameter a are located and represented by the two vertical dotted dash blue lines; the parameter samples between the two lines is 95.02% of all the MCMC samples. Following the same procedure, the minimum and maximum values of parameter m are represented by the two horizontal dotted dash red lines and 95.01% of all the MCMC samples are enclosed by the two lines. We can conclude that the 95% confidence and credible regions are numerically identical. Based on the 95% individual parameter confidence region (the black contour in Figure 2c), the 95% individual prediction confidence interval can be evaluated. According to the discussion in section 2.6, these confidence interval limits are the minimum and maximum prediction values on the

confidence region, which is equivalent to the credible region. This equivalence means that these two predictions define both 95% individual confidence and credible intervals. The credible intervals produced in this manner include about 95.02% of all the MCMC samples, which again verifies the equivalence of the 95% confidence and credible intervals. This suggests that for this mildly nonlinear problem with small intrinsic nonlinearity, the 95% parameter credible region is identical to its 95% confidence region and the 95% credible interval of prediction evaluated on the credible region is identical to its 95% confidence interval evaluated on the confidence region.

[53] The correspondence of the regions and intervals between classical regression and Bayesian analysis is not an accident for the following reasons. (1) The problem is well conditioned: the ratio of the largest and smallest singular values of $(X^T \omega^{1/2})$ is about 6.4, which corresponds to the composite scaled sensitivities values being substantial and similar in value and parameter correlation being small [Hill, 2010]. In Figure 2c, this is clear because the contour is not elongated parallel to either axis or at an angle to the axes. (2) There is one minimum. (3) The single mode parameter posterior probability distribution can be easily sampled by MCMC techniques and MICA and DREAM produce similar results. (4) The prediction function is monotonic and the maximum and minimum values are on the boundary of the region. (5) The intrinsic model nonlinearity is very small, 0.002, suggesting that the approximate likelihood region is almost correct. Consequently, in this case, equation (16) can be used to calculate an accurate individual nonlinear confidence interval and MCMC samples can accurately represent the posterior distribution of parameters and an accurate individual nonlinear credible interval.

4. Complex Test Case of Synthetic Groundwater Models

[54] The complex test case of synthetic groundwater models is more directly relevant to environmental problems than the simple test cases and is designed to extend our

Table 5. Predictions and 95% Nonlinear Confidence and Credible Intervals for Mildly Nonlinear Simple Test Function II^a

Type of Interval	Prediction	Lower Limit	Upper Limit
Nonlinear confidence	0.406	0.3998	0.4109
Nonlinear credible	0.406	0.3997	0.4109

^aIntrinsic model nonlinearity is 0.002. True value of the prediction is 0.401.

Figure 3. (a) Modeling domain with the confining unit, blue Lake, and Straight River, (b) the true horizontal hydraulic conductivity field (K field) with values from 6 to 150 m/d, and (c–d) the calibrated K fields for the three-zone model (3Z) with K2 up to 317 m/d and the interpolation model (INT) with highest value 180 m/d. (Figure 3c) Configuration of the three zones of K field for 3Z. (Figure 3d) Dots are nodes of linear triangular basis functions constructed to interpolate K field for INT. The hydraulic conductivities are changed in the regression at the sixteen numbered inner dots and at the three squares labeled A, B and C; the hydraulic conductivities of the other seven squares without labels are set. In Figure 3c dots represent 27 wells where hydraulic heads are observed at the top and bottom of the system. Black dots (same as those in Figure 3d) also have hydraulic conductivity measurements and yellow dots only have heads observations; pumping is taken at wells P1 and P3 (blue dots), and drawdown is predicted at well P3.

understanding of the confidence and credible intervals to more practical issues in the context of model uncertainty. We consider several alternative models with different levels of model error, and evaluate the measures of uncertainty in the context of true prediction values and the assumptions required to obtain accurate confidence and credible intervals.

4.1. Model Description, Sensitivity Analysis, and Calibration

[55] This test case includes simulations using a synthetic true model and three alternative calibrated models developed based on data from the true model; all the models are steady state. The true model is described in *Hill et al.* [1998], where it was used to study nonlinear regression methods. The three-dimensional modeling domain is an undeveloped alluvial valley with top water table boundary (subject to areal recharge) and impermeable bottom and lateral boundaries. As shown in Figure 3a, the system has a clay confining

unit, a lake (Blue Lake), and a river (Straight River). In the true and all calibrated modes, the river is modeled as a head-dependent boundary. All models use the true distribution of the confining layer shown in Figure 3a because *Hill et al.* [1998] showed that results were insensitive to reasonable variations. Each model layer is characterized by heterogeneous horizontal hydraulic conductivity with values ranging from 6 to 150 m/d in the true model (Figure 3b) and 27 to 317 m/d in the calibrated models (Figures 3c and 3d). No vertical variation in aquifer hydraulic conductivity is considered in any model.

[56] Development scenario A of *Hill et al.* [1998] considered in this study includes pumpage at wells P1 and P3 (locations shown in Figure 3c). Model predictions of interest are drawdown of the water table at well P3 and percent decrease of streamflow at the gauge site G2 (Figure 3a).

[57] The three alternative models represent three different ways of parameterizing the heterogeneous horizontal

hydraulic conductivity. The first model (HO) is the simplest, treating the domain as homogeneous. The second model (3Z) has three zones shown in Figure 3c, which produced the best overall model calibration among many other zone configurations considered by Hill *et al.* [1998]. In the third model (INT), the field of horizontal hydraulic conductivity is parameterized by interpolation using linear triangular basis functions constructed on the points in Figure 3d. Model HO is developed in this study; models 3Z and INT are modified from those of Hill *et al.* [1998].

[58] The alternative models differ from the true model in several ways. The most important difference is that, instead of modeling the lake as a head-dependent boundary as in the true model and Hill *et al.* [1998], the volume of the lake is simulated as high hydraulic conductivity cells to avoid the modeler intervention described by Hill *et al.* [1998]. The true model has five layers; the calibrated models all have three as follows. In the north, the bottom of layer 1 coincides with the bottom of the lake; in the south, the bottom of layer 1 is deep enough to ensure that no cell goes dry during the simulation. The bottom of layer 2 coincides with the top of the confining unit which is simulated as vertical leakage between model layers 2 and 3.

[59] The defined parameters common to all the models are net lake recharge (LAKERCH), areal recharge rate (RCH), leakage of the confining unit (KV), and vertical anisotropy (VANI). The calibrated parameters specific to the individual models are hydraulic conductivities and the use of one (for models HO and 3Z) or three (for model INT) parameters for the conductance of the riverbed.

[60] Calibration data include 54 observations of hydraulic head from 27 wells (two heads from each well in layers 1 and 3, Figure 3c) and one observation of lake stage. Observations of streamflow gain (groundwater discharge to the stream) are also used in the calibration. To investigate the effects of more streamflow data on prediction accuracy and uncertainty, two streamflow gain data sets are used: two observations from gauges G1 and G2 (Case I) or eighteen observations from each cell of the river (Case II). In Case I, the observation of G2 and the difference between G1 and G2 are used in the calibration. The level of detail provided by Case II is generally not realistic in practice but here the value of such data is evaluated. Prior information on LAKERCH is used in all three models. For model INT, prior information also includes sixteen measurements of hydraulic conductivity from the wells shown in Figure 3d.

[61] Errors added to the observed heads and the lake stage are based on typical values for the data type; the weights used in model calibration are calculated solely based on the added measurement errors and do not account for model error. The observations of hydraulic head and lake stage have white noise with variance of 0.01 m^2 , the inverse of which is the weights. For measurements of net lake recharge and hydraulic conductivity used as prior information, coefficients of variation of 50% and 20%, respectively, are used to define the added white noise and the weights. No measurement errors are added to the created streamflow data observations. Realistic coefficients of variation are used to determine the weighting: 10% and 20% in cases I and II, respectively. A larger value is used in Case II because it is expected that such detailed flows are likely to be measured with larger percent errors. Because no errors are added, any discrepancies are due only to model error.

[62] Estimated parameters are selected based on the composite scaled sensitivity (CSS) and parameter correlation coefficient (PCC) obtained from a sensitivity study [Hill and Tiedeman, 2007, p. 50–54]. CSS values identify parameters that possibly can be reasonably estimated using the calibration data. Sensitive parameters that are highly correlated with one or more other parameters, as indicated using PCC, possibly cannot be estimated uniquely with the available data. If present, selected highly correlated parameters generally need to be removed from the set of estimated parameters. Take the 3Z model as an example: Figures D1a and D1c in Appendix D of the auxiliary material show that parameter RCH has the highest CSS values and VANI and KV have the smallest. This is also true for the HO and INT models. In this work, using heads alone would result in all parameters except VANI to be completely correlated (absolute value of PCC equal to 1.00) so that the parameters would be interdependent and nonunique. This is because all hydraulic conductivities and boundary flows would have been estimated with no flow specified or observed. Adding the measured streamflow gains and losses provide sufficient information that all the parameters can be uniquely estimated: the largest absolute value of PCC was 0.93, which is too much smaller than 1.00 to cause nonunique parameter estimates.

[63] Model calibration is conducted using the modified Gauss-Newton method in UCODE_2005 to minimize the following objective function

$$S(\mathbf{b}) = \sum_{i=1}^{NH} \omega_{h_i} (y_{h_i} - \hat{y}_{h_i}(\mathbf{b}))^2 + \sum_{j=1}^{NQ} \omega_{q_j} (y_{q_j} - \hat{y}_{q_j}(\mathbf{b}))^2 + \sum_{k=1}^{NPR} \omega_{p_k} (y_{p_k} - \hat{y}_{p_k}(\mathbf{b}))^2 \quad (24)$$

where $\mathbf{b}$ is a vector of NP parameters; NH and NQ equal the number of hydraulic-head and streamflow observations, respectively; NPR is the number of prior information values; y_{h_i} and $\hat{y}_{h_i}(\mathbf{b})$ represent the observed and simulated hydraulic head, respectively; ω_{h_i} represents the weight for the head observation; y_{q_j} and $\hat{y}_{q_j}(\mathbf{b})$ represent the observed and simulated streamflow, respectively; ω_{q_j} represents the weight for the streamflow; y_{p_k} and $\hat{y}_{p_k}(\mathbf{b})$ represent the prior parameter value and corresponding estimate of the parameter in the regression, respectively; and ω_{p_k} represents the weight for the prior estimate.

4.2. Simulating the Confidence and Credible Intervals

[64] The MCMC simulation of credible intervals is conducted using DREAM. The prior distributions of parameters used in the Bayesian analysis are consistent with the prior information used in the classical regression. For parameters without prior information in the regression, uniform prior distributions with large bounds are used in MCMC to create noninformative priors; for parameters with prior information in the regression, Gaussian prior distributions are applied in MCMC with the mean and variance set to equal y_{p_k} and $\omega_{p_k}^{-1}$ of (24). Details of the parameter prior distribution can be found in Appendix E of the auxiliary material.

[65] Ten chains are used in the MCMC simulation, each of which generated 20,000 samples after the chains converge. The convergence is monitored by the factor R , as

Figure 4. Linear and nonlinear confidence intervals and nonlinear credible intervals (≤ 106 , ≤ 1594 , and $420,000$ model runs, respectively) for two predictions: (a and c) streamflow change at gauge site G2 and (b and d) drawdown of the water table at well P3. Here 2 or 18 observations of streamflow gain are used. The horizontal lines represent true values of the predictions. The nonlinear credible intervals are calculated using DREAM.

defined and suggested by Gelman *et al.* [1995]. The largest R for the first 2000 iterations (total 20,000 parameter samples of 10 chains) is 1.05, which suggests convergence to the posterior probability distribution because it is below the critical value 1.2). Model predictions are calculated using the 200,000 samples, and for each prediction the 95% equal-tailed credible interval is estimated subsequently.

[66] Linear measures of parameter uncertainty required $2 \times NP + 1$ forward model runs without pumping, where $NP = 6, 8$, or 26 for models HO, 3Z, and INT, respectively. The same number of prediction model runs with pumping is required to propagate the parameter uncertainty into prediction uncertainty (i.e., based on equation (11)). Thus, to calculate linear confidence intervals on predictions, 26, 34, and 106 model runs are needed for the HO, 3Z, and INT models, respectively.

[67] Nonlinear confidence intervals on predictions required an optimization process for each interval limit. In this work, the two nonlinear intervals for each of the HO, 3Z, and INT models required about 303, 412, and 1594 model runs, respectively.

[68] Nonlinear credible intervals require about 420,000 model runs without and with pumping for each model. Any number of predictions can be simulated with the MCMC results, so that if many predictions are considered MCMC can become competitive with nonlinear confidence intervals [Shi *et al.*, 2012]. However, the number of predictions is rarely that large.

4.3. Linear and Nonlinear Confidence and Credible Intervals

[69] The 95% linear and nonlinear confidence and credible intervals calculated for two predictions and either two or eighteen streamflow gain observations are shown in Figure 4. We have already shown and demonstrated (Table 1 and Figures 1a–1c) that linear confidence and credible intervals are numerically identical (as discussed in sections 2.3 and 2.4), so only linear confidence intervals are plotted.

[70] Figure 4 shows that, for models 3Z and INT, the three intervals for both predictions in both cases are very close to each other relative to the distance from the prediction to the true value. For HO, the credible intervals are smaller than both the linear and nonlinear confidence intervals, especially for predicted drawdown (Figure 5). The reasons are multiple. First, HO is the least accurate model. Second, the intrinsic nonlinearity of HO is the largest as shown in Table 6 and Figure 5. Third, the prediction function of drawdown is not monotonic with respect to parameters LAKERCH and RCH, though the insensitivity of these parameters to drawdown and the slight nonmonotonicity suggest the last effect may be small.

[71] The results presented in this work cannot be used to test the significance level of the intervals because only a few samples are considered. However, true values located far outside the intervals can be used to suggest that the intervals are too small. The intervals for HO in Figures 4a, 4b, and 4d exclude the true values by enough that the accuracy of the prediction significance level is drawn into question. For model 3Z, the true value of drawdown is excluded from both the confidence and credible intervals after including the more detailed observations, suggesting that model 3Z may not produce reliable measures of prediction uncertainty. These examples are consistent with the idea that simple models tend to underestimate prediction uncertainty.

[72] Using the more detailed observations of streamflow gain reduces interval width of predictions for all models due to reduced parameter uncertainty. In model 3Z, for example, the 95% linear confidence interval for parameter K2 is reduced from 134–741 m/d to 136–557 m/d by adding the more detailed observations. Using the more detailed observations of streamflow gain also improves predictions accuracy slightly, as evidenced by the simulated values in Figures 4c and 4d being slightly closer to the true values than those in Figures 4a and 4b.

4.4. Evaluation

[73] In this section, the reasons that the true values are excluded from the confidence and credible intervals of models HO and 3Z are investigated by examining

Figure 5. Relative increase in width of linear and nonlinear confidence intervals to nonlinear credible intervals for streamflow change and drawdown with (a–b) 2 and (c–d) 18 observations of streamflow gain. The measures of intrinsic nonlinearity from Table 6 are inserted in the shaded boxes for easy comparison.

assumptions required for deriving accurate intervals. Derivation of accurate linear and nonlinear confidence intervals assumes that true errors are Gaussian with zero mean and that model errors are either insignificant relative to measurement errors (i.e., the models are nearly error-free representations of the system) or the model errors are zero-mean and random (i.e., the models have no distinct biases), and en masse produce a Gaussian contribution to errors. The Gaussian assumptions are needed to construct confidence intervals (though not for the regression to optimize parameter values) but in theory are not needed for credible intervals because non-Gaussian likelihood functions and prior distributions are allowed in Bayesian analysis. For consistent comparison, this work uses a Gaussian likelihood function and, when applicable, a Gaussian prior distribution. Thus, this work does not investigate situations in which non-Gaussian errors are accounted for by credible intervals, but only considers the ability of credible intervals to correctly account for model nonlinearity, including resulting local minima. Using the Gaussian structure for non-Gaussian errors may lead to less accurate model simulations [Sun *et al.*, 2009] and overestimation of predictive uncertainty [Schoups and Vrugt, 2010]. The assumptions about Gaussian distributed errors and the size and characteristics of model errors are examined for both the confidence and credible intervals.

[74] However, the focus on Gaussian residual distributions is expected to be of wide interest because they are apparently often applicable and commonly used [Renard *et al.*, 2011]. For non-Gaussian residuals, as models of a given system evolve, bias in residuals may diminish and residual structures may become Gaussian after certain transforms (e.g., the Box-Cox transformation). In this sense, the Gaussian assumption of residuals may become increasingly realistic. On the other hand, our complex example suggests that model errors are similarly debilitating for both confidence and credible intervals, and dominate over the differences between confidence and credible intervals. Therefore, faith that somehow credible intervals are more reliable in the face of model error seems somewhat questionable based on the results shown here.

[75] To investigate whether true errors are Gaussian or equivalently weighted true errors are standard normal, the weighted-residuals are evaluated. Weighted-residuals tend to be correlated even when true errors are independent, but

Table 6. Total Nonlinearity, Intrinsic Nonlinearity, Correlation Coefficient and Regression Statistics and Their Critical Values, in Parentheses, of the Three Alternative Groundwater Models

Statistic	Case I: 2 Streamflow Gains			Case II: 18 Streamflow Gains		
	HO	3Z	INT	HO	3Z	INT
Total nonlinearity ^a	2.44	0.81	13.85	2.79	0.91	2.27
Intrinsic nonlinearity ^a	0.54	0.04	0.18	0.57	0.07	0.11
R_N^2 ^b	0.989 (0.96)	0.986 (0.96)	0.989 (0.97)	0.987 (0.97)	0.988 (0.97)	0.982 (0.97)
s ^c	1.49 (1.25–1.84)	1.27 (1.06–1.57)	1.05 (0.88–1.30)	1.33 (1.14–1.60)	1.17 (1.0–1.4)	0.95 (0.81–1.14)

^aCalculated using equation (7.15) and (7.16) of Hill and Tiedeman [2007] by UCODE_2005. Critical values for both measures: >1.0 highly nonlinear; 0.09–1.0 nonlinear; 0.01–0.09 moderately nonlinear; <0.01 effectively linear.

^bCalculated using equation (6.18) of Hill and Tiedeman [2007] by UCODE 2005. R_N^2 normality test for weighted-residuals evaluated for the observations and the prior information. Critical values are in parentheses. Larger R_N^2 values indicate normally distributed weighted-residuals.

^cConfidence intervals for s are shown in parentheses and are calculated using equation (6.2) of Hill and Tiedeman [2007] by UCODE 2005. Confidence intervals that are entirely above 1.0 indicate model fit that is significantly worse than would be consistent with the weighting. If the weighting is error-based, as in this study, such large values of s indicate important model error and possibly model bias.

normality of the weighted-residuals suggests normality of the true errors. Also, it is the normality of the weighted-residuals that is important for (16) used to calculate nonlinear confidence intervals. The statistical variable, R_N^2 is the correlation coefficient between the weighted-residuals (ordered from smallest to largest) and the normal order statistics [Brockwell and Davis, 1987, p. 304; Hill and Tiedeman, 2007, p. 110], and can be used to test if the weighted-residuals are normally distributed and also satisfy the more demanding condition of independence. R_N^2 provides a more powerful test than the common Kolmogorov-Smirnov statistic. Values of R_N^2 close to 1.0 indicate that the weighted-residuals are independent and normally distributed. Its values and corresponding critical values at significance level of 0.05 are listed in Table 6 for the three models. Because the R_N^2 values are all larger than the critical values, it is concluded that the assumption of error normality is satisfied for all models in the synthetic groundwater test case.

[76] To investigate whether model errors are significant relative to measurement errors, Hill and Tiedeman [2007, p. 303], Aster et al. [2012], and others suggest using the standard error of regression, $s = [S(\hat{\mathbf{b}})/(NH + NQ + NPR - NP)]^{1/2}$, where $S(\hat{\mathbf{b}})$ is from minimizing equation (24). If the model fit is consistent with assigned weighting then the calculated standard error of regression s is close to 1.0. Larger values of s indicate either additional errors, commonly epistemic and aleatory errors, not accounted for in the weighting or that the weights do not correctly reflect the intended errors [Hill and Tiedeman, 2007]. In this synthetic study the measurement error is known and incorporated properly in the weights for all observations except flows, but no model error is included. For flows, the lack of added error and use of coefficients of variation of 10% and 20% mean that even values of 1.0 suggest some model error. Thus, s values larger than 1.0 indicate model errors. The s values and their corresponding 95% confidence intervals listed in Table 6 suggest significant model errors for models HO and 3Z. The model errors of HO and 3Z mainly result from oversimplified conceptualizations of the horizontal hydraulic conductivity (Figure 3).

[77] Other investigations of model error consider plots of weighted-residuals and optimized parameter values. Figures D1b and D1d in Appendix D of the auxiliary material plot weighted-residuals against weighted simulated values from model 3Z for the case of using two and eighteen observations of streamflow gain for calibration, respectively. While the calibration results are acceptable, the weighted-residuals of model 3Z are not fully random in that they are mostly positive for weighted simulated values between 26 and 32. This may indicate systematic model error in model 3Z. Though results for HO and INT are not shown, model HO has worse overall fit than 3Z and model INT has the best fit with weighted-residuals randomly distributed. Both models HO and 3Z suffer from biased parameter estimates. For example, for model 3Z the estimated value of K2 equals 317 m/d, more than two times the largest hydraulic conductivity of the true model (Figure 3); the estimated values of riverbed conductance for models HO and 3Z, which equal 312.94 and 303.29, respectively, are far from the true value of 244 m²/d per meter of stream, however this is not clear evidence of model bias because streambed conductances depend on grid size [Mehl and Hill, 2010]. In this synthetic

problem, we know that the bias is mainly caused by oversimplification of horizontal hydraulic conductivity. No such problems occur for the INT model, for which the hydraulic conductivity parameterization is most closely coordinated with the pattern of the true field (Figures 3b and 3d).

[78] The groundwater examples suggest that both intrinsic model nonlinearity and model error affect the differences between nonlinear confidence and credible intervals, but their relation is not simple. Table 6 and Figure 5 show that the largest differences between the confidence and credible intervals occur for model HO, which has the largest intrinsic model nonlinearity and worst model fit. However, results for models 3Z and INT do not support this relation clearly. For example, the smaller standard error of regression s in Table 6, the distribution of weighted-residuals, and more reasonable parameter values indicate that INT has less model error than 3Z, but some of the confidence intervals are closer to the credible intervals for 3Z than for INT (Figure 5d). This is probably because the intrinsic model nonlinearity of INT (0.18 and 0.11 for case I and II) is larger than that of 3Z (0.04 and 0.07) (Figure 5). For both, difficulties associated with model error remain.

5. Discussion

[79] This work is focused on understanding differences and similarities between the confidence and credible intervals from theoretical and heuristic perspectives. We limit our discussions and conclusions to the problems that the observation errors are multivariate Gaussian distributed, which is commonly useful in environmental modeling. In this study, linear and nonlinear confidence and credible intervals are considered for both linear and nonlinear models; results are summarized and compared to other published work in Table 1.

[80] For *linear* confidence and credible intervals of *linear* models, analytical expressions were used to show that, when no prior information is used in the regression and the non-informative prior parameter distribution is used in the Bayesian calculation, the two kinds of intervals are numerically identical. If prior information used in the regression is consistent with the prior distribution used in the Bayesian calculation, the confidence intervals are equivalent to the credible intervals; otherwise, the two kinds of intervals differ.

[81] For *linear* confidence and credible intervals of *linearized-nonlinear* models, we show that the two kinds of intervals are again numerically identical. However, both confidence and credible intervals are approximate, and their accuracy depends on model nonlinearity and model error.

[82] For *nonlinear* confidence and credible intervals of *nonlinear* models, analytical expressions are not available and heuristic investigations were pursued. Confidence intervals were calculated using the approximate likelihood methods of Christensen and Cooley [1999] and Cooley [2004], while credible intervals were calculated using the MCMC techniques of Gallagher and Doherty [2007] and Vrugt et al. [2008, 2009].

[83] Three numerical experiments are conducted to explore the differences between linear and nonlinear confidence and credible intervals for nonlinear models. The experiments with simple nonlinear functions indicate that the intervals are similar for consistent priors, as long as the

assumptions used to calculate the confidence intervals are satisfied and the credible intervals are correctly simulated by MCMC techniques. Therefore, in practice, the expected advantage produced by calculating computationally demanding nonlinear credible intervals can be evaluated by considering the underlying assumptions. In the very nonlinear test function I, the nonlinear credible interval is smaller than the linear and nonlinear confidence intervals because the intrinsic model nonlinearity is large and multiple minima exist in the objective function. But in the mildly nonlinear test function II, when all the assumptions are satisfied, the confidence and credible parameters regions are numerically coincident, and correspondingly the nonlinear confidence and credible intervals of predictions are numerically identical. Even in the very nonlinear test case I, the linear and nonlinear confidence intervals constructed using the global minima were close enough to the credible intervals to suggest potential utility given the small number of model runs required by the confidence intervals (10 and 232 for the linear and nonlinear confidence intervals, respectively; 1,000,000 for the credible intervals).

[84] The numerical experiment of groundwater modeling introduces the complication of model error and resulting prediction bias on the performance of confidence and credible intervals. The confidence and credible intervals are generally closer to each other when model error is small and the intrinsic model nonlinearity is not very large, despite some exception for the 3Z and INT models (Figure 5). For model HO, which has greater model error and nonlinearity, the difference between confidence and credible intervals is the largest. However, the more important result is that for model HO all of the intervals perform similarly and not very well in that they excluded the true value by a wide margin. While more experience is needed to determine common performance, this example and theoretical considerations suggest poor performance of all intervals is likely when model bias dominates. This suggests that it may be useful to calculate the less computationally demanding confidence intervals early in model development, and calculate the computationally demanding credible intervals as the model becomes a better representation of the system. An advantage of this approach is that it allows more routine calculation of uncertainty intervals and associated measures of, for example, the importance of observations and parameters to predictions [Tonkin *et al.*, 2007; Dausman *et al.*, 2010]. This results in greater understanding of simulated dynamics and increased model transparency.

[85] We suggest that model nonlinearity can be measured using the intrinsic model nonlinearity statistic [Bates and Watts, 1980; Cooley, 2004]. A difficulty with this statistic is that it considers only the nonlinearity of observations with respect to the parameters, a difficulty discussed by Hill and Tiedeman [2007, p. 189–193] and Cooley [2004]. In this work, the predictions considered are similar to the observations used and intrinsic model nonlinearity is expected to provide information about the linearity of the predictions as well. When predictions differ substantially from observations due to differences between calibration and prediction conditions or because different quantities are considered, it may not be as useful a measure of nonlinearity.

[86] The results of very nonlinear simple test case I suggests that the match between nonlinear confidence and credible intervals is affected strongly by the existence of

local minima. Of concern are how common local minima are in practice, what types of models are likely to have local minima, and how local minima can be detected, as also noted, for example, by Kavetski *et al.* [2006] in the context of surface hydrologic modeling. When models include many kinds of observations and are characterized by few parameters or many parameters with substantial prior information or regularization, parameter sensitivities are likely to be larger, absolute values of parameter correlation coefficients are likely to be further from 1.00, the inverse problem is more likely to be well posed, and local minima are likely to be less common. Thus, decisions made regarding model data and construction are critical to how similar confidence and credible intervals are likely to be in a given problem. Kavetski and Clark [2010] also discuss the importance of model numerical methods to the presence of local minima.

6. Conclusions

[87] The results presented in this work suggest that for linear models, confidence and credible intervals are mathematically and numerically identical when either (1) parameter prior information is absent for the confidence intervals and noninformative for the credible intervals or (2) any prior information defined for the confidence intervals is consistent with informative prior defined for the credible intervals. Because prior information for the confidence intervals is defined using only means, variances, and covariances, consistency means that the prior information in the credible intervals needs to be Gaussian.

[88] For nonlinear models, confidence and credible intervals can be numerically identical when all the assumptions are satisfied in the calculation of confidence intervals and credible intervals are calculated precisely. In practice, problem complexity leads to difficulties for both types of intervals. This work shows how competing local minima degrade the accuracy of confidence intervals. MCMC methods used to calculate credible intervals can suffer from numerical imprecision caused by inadequate sampling of the parameter space; long execution times and large parameter dimensions of many practical models can make it difficult or impossible to conduct enough model runs for sufficiently sampling. The accuracies of both confidence and credible intervals suffer from model errors. In practice nonlinear confidence intervals often differ from nonlinear credible intervals and the differences do not in themselves indicate which uncertainty interval should be more trusted. It is often assumed that the difficulties cited here for confidence intervals are more debilitating than those cited for credible intervals. This work suggests that the situation is likely to be more nuanced, and that both confidence and credible intervals can be important to uncertainty evaluation of environmental models.

[89] The computational expense of credible intervals (for the complex problem considered in this work, about 500,000 model runs versus about 100 and 1500 model runs for linear and nonlinear confidence intervals, respectively) is likely to influence the choice of the uncertainty analysis method used. Whether one uses confidence intervals or credible intervals, the results of this work indicate that the difference between alternative models can be more critical than the differences between confidence and credible intervals in many practical circumstances.

[90] Suggestions for users include the following. (1) Report all uncertainty intervals with an evaluation of model error. Model error can be evaluated using the sum of squared weighted-residuals when the weighting is based on observation error (including epistemic error), as well as plots of weighted-residuals and evaluation of estimated parameter values. For nonlinear models, linear intervals and nonlinear confidence intervals need to be reported with an evaluation of model nonlinearity. (2) Nonlinear confidence intervals can be estimated accurately using the approximate likelihood method, if the assumptions are approximately satisfied. The intervals should be reported with an examination of the assumptions. (3) Nonlinear credible intervals calculated from MCMC techniques can be estimated accurately if the parameter space is adequately explored. MCMC techniques face their own challenges for high-dimension problems with multimodal distributions because it is difficult to adequately search high-dimensional parameter spaces. (4) It appears likely that less computationally demanding confidence intervals are adequate early in model development when model error is likely to adversely affect all types of intervals. This provides more opportunity to take advantage of the rich set of related methods for identifying observations and parameters important to predictions. Use of the computationally demanding credible intervals can be reserved for when the model becomes a better representation of the system.

[91] **Acknowledgments.** The authors thank David Draper and Bruno Mendes for inspiring discussions, John Doherty and Jasper Vrugt for providing the MICA and DREAM codes, respectively, used for the MCMC simulation, Dmitri Kavetski for helping to verify the MCMC results, and Alberto Montanari and three anonymous reviewers for comments on the manuscript. This work was supported in part by NSF-EAR grant 0911074, DOE-SBR grant DE-SC0002687 and DOE Early Career Award DE-SC0008272.

References

- Aster, R. C., B. Borchers, and C. H. Thurber (2012), *Parameter Estimation and Inverse Problems*, 2nd ed., 360 pp., Elsevier, Amsterdam.
- Bates, D. M., and D. G. Watts (1980), Relative curvature measures of nonlinearity, *J. R. Stat. Soc. B*, 42(1), 1–25.
- Bates, D. M., and D. G. Watts (1988), *Nonlinear Regression Analysis and Its Applications*, 365 pp., John Wiley and Sons, New York.
- Beale, E. M. L. (1960), Confidence regions in non-linear estimation, *J. R. Stat. Soc. B*, 22(1), 41–76.
- Box, E. P., and G. C. Tiao (1992), *Bayesian Inference in Statistical Analysis*, 588 pp., Wiley, New York.
- Brockwell, P. J., and R. A. Davis (1987), *Time Series, Theory and Methods*, 600 pp., Springer, New York.
- Carrera, J., and S. P. Neuman (1986), Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, *Water Resour. Res.*, 22(2), 199–210, doi:10.1029/WR022i002p00199.
- Casella, G., and R. L. Berger (2002), *Statistical Inference*, 2nd ed., 660 pp., Duxbury, Belmont, Calif.
- Chen, M., and Q. Shao (1999), Monte Carlo estimation of Bayesian credible and HPD intervals, *J. Comput. Graph. Statist.*, 8(1), 69–92.
- Christensen, S., and R. L. Cooley (1999), Evaluation of confidence intervals for a steady-state leaky aquifer model, *Adv. Water Resour.*, 22(8), 807–817, doi:10.1016/S0309-1708(98)00055-4.
- Christensen, S., and R. L. Cooley (2005), User guide to the UNC process and three utility programs for computation of nonlinear confidence and prediction intervals using MODFLOW-2000, *U.S. Geol. Surv. Tech. Methods Rep.*, 2004-1349.
- Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, 47, W09301, doi:10.1029/2010WR009827.
- Cook, R. D., and S. Weisberg (1982), *Residuals and Influence in Regression*, 230 pp., Chapman and Hall, New York.
- Cooley, R. L. (1977), A method of estimating parameters and assessing reliability for models of steady-state groundwater flow: 1. Theory and numerical properties, *Water Resour. Res.*, 13(2), 318–324, doi:10.1029/WR013i002p00318.
- Cooley, R. L. (1979), A method of estimating parameters and assessing reliability for models of steady state groundwater flow: 2. Application of statistical analysis, *Water Resour. Res.*, 15(3), 603–617, doi:10.1029/WR015i003p00603.
- Cooley, R. L. (1983), Incorporation of prior information on parameters into nonlinear regression groundwater flow models: 2. Applications, *Water Resour. Res.*, 19(3), 662–676, doi:10.1029/WR019i003p00662.
- Cooley, R. L. (1993a), Exact Scheffé-type confidence intervals for output from groundwater flow models: 1. Use of hydrogeologic information, *Water Resour. Res.*, 29(1), 17–33, doi:10.1029/92WR01863.
- Cooley, R. L. (1993b), Exact Scheffé-type confidence intervals for output from groundwater flow models 2. Combined use of hydrogeologic information and calibration data, *Water Resour. Res.*, 29(1), 35–50, doi:10.1029/92WR01864.
- Cooley, R. L. (1997), Confidence intervals for ground water models using linearization, likelihood, and bootstrap method, *Ground Water*, 35(5), 869–880, doi:10.1111/j.1745-6584.1997.tb00155.x.
- Cooley, R. L. (1999), Practical Scheffé-type credibility intervals for variables of a groundwater model, *Water Resour. Res.*, 35(1), 113–126, doi:10.1029/98WR02819.
- Cooley, R. L. (2004), A theory for modeling groundwater flow in heterogeneous media, *U.S. Geol. Surv. Prof. Pap.*, 1679.
- Cooley, R. L., and R. L. Naff (1990), Regression modeling of ground-water flow, in *U.S. Geological Survey Techniques of Water-Resources Investigations*, Book 3, Chap. B4, 232 pp., U.S. Geol. Surv., Washington, D.C.
- Cooley, R. L., and A. V. Vecchia (1987), Calculation of nonlinear confidence and prediction intervals for groundwater flow models, *Water Resour. Bull.*, 23(4), 581–599, doi:10.1111/j.1752-1688.1987.tb00834.x.
- Dausman, A. M., J. Doherty, C. D. Langevin, and M. C. Sukop (2010), Quantifying data worth toward reducing predictive uncertainty, *Ground Water*, 48(5), 729–740, doi:10.1111/j.1745-6584.2010.00679.x.
- Doherty, J. (2003), *MICA: Model-independent Markov Chain Monte Carlo Analysis*, Watermark Numer. Comput., Brisbane, Australia.
- Doherty, J., and R. J. Hunt (2010), Response to comment on “Two statistics for evaluating parameter identifiability and error reduction,” *J. Hydrol.*, 380, 489–496, doi:10.1016/j.jhydrol.2009.10.012.
- Donaldson, J. R., and R. B. Schnabel (1987), Computational experience with confidence regions and confidence intervals for nonlinear least squares, *Technometrics*, 29, 67–82.
- Draper, N. R., and H. Smith (1998), *Applied Regression Analysis*, 3rd ed., John Wiley, New York.
- Fienen, M. N., J. E. Doherty, R. J. Hunt, and H. W. Reeves (2010), Using prediction uncertainty analysis to design hydrologic monitoring networks: Example applications from the great lakes water availability pilot project, *U.S. Geol. Surv. Sci. Invest. Rep.*, 2010-5159.
- Finsterle, S., and K. Pruess (1995), Solving the estimation-identification problem in two-phase flow modeling, *Water Resour. Res.*, 31(4), 913–924, doi:10.1029/94WR03038.
- Finsterle, S., and Y. Zhang (2011), Error handling strategies in multiphase inverse modeling, *Comput. Geosci.*, 37, 724–730, doi:10.1016/j.cageo.2010.11.009.
- Foglia, L., S. W. Mehl, M. C. Hill, P. Perona, and P. Burlando (2007), Testing alternative ground water models using cross-validation and other methods, *Ground Water*, 45(5), 627–641, doi:10.1111/j.1745-6584.2007.00341.x.
- Fu, G., D. Butler, S.-T. Khu, and S. Sun (2011), Imprecise probabilistic evaluation of sewer flooding in urban drainage systems using random set theory, *Water Resour. Res.*, 47, W02534, doi:10.1029/2009WR008944.
- Gallagher, M., and J. Doherty (2007), Parameter estimation and uncertainty analysis for a watershed model, *Environ. Model. Softw.*, 22, 1000–1020, doi:10.1016/j.envsoft.2006.06.007.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, 696 pp., Chapman and Hall, London.
- Haario, H., E. Saksman, and J. Tamminen (2001), An adaptive Metropolis algorithm, *Bernoulli*, 7(2), 223–242, doi:10.2307/3318737.
- Hill, M. C. (1989), Analysis of accuracy of approximate, simultaneous, nonlinear confidence intervals on hydraulic heads in analytical and numerical test cases, *Water Resour. Res.*, 25(2), 177–190, doi:10.1029/WR025i002p00177.
- Hill, M. C. (2010), Comment on “Two statistics for evaluating parameter identifiability and error reduction” by John Doherty and Randall J. Hunt, *J. Hydrol.*, 380, 481–488, doi:10.1016/j.jhydrol.2009.10.011.

- Hill, M. C., and C. R. Tiedeman (2007), *Effective Calibration of Ground Water Models, With Analysis of Data, Sensitivities, Predictions, and Uncertainty*, 480 pp., John Wiley and Sons, New York.
- Hill, M. C., R. L. Cooley, and D. W. Pollock (1998), A controlled experiment in ground water flow model calibration, *Ground Water*, 36(3), 520–535, doi:10.1111/j.1745-6584.1998.tb02824.x.
- Jaynes, E. T. (1976), Confidence intervals vs Bayesian intervals, in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, edited by W. L. Harper and C. A. Hooker, pp. 175–257, D. Reidel Publ., Dordrecht, Netherlands.
- Kavetski, D., and M. P. Clark (2010), Ancient numerical daemons of conceptual hydrological models revisited: 2. Impact of time stepping schemes on model analysis and prediction, *Water Resour. Res.*, 46, W10511, doi:10.1029/2009WR008896.
- Kavetski, D., G. Kuczera, and S. W. Franks (2006), Calibration of conceptual hydrological models revisited: 1. Improving optimization and analysis, *J. Hydrol.*, 320, 187–201, doi:10.1016/j.jhydrol.2005.07.013.
- Kitanidis, P. K. (1986), Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *Water Resour. Res.*, 22(4), 499–507, doi:10.1029/WR022i004p00499.
- Kitanidis, P. K. (1997), *Introduction to Geostatistics, Applications in Hydrology*, 272 pp., Cambridge Univ. Press, New York, doi:10.1017/CBO9780511626166.
- Kitanidis, P. K. (2010), Bayesian and geostatistical approaches to inverse problems, in *Large-scale inverse problems and quantification of uncertainty*, edited by L. Biegler et al., pp. 71–86, Wiley, New York, doi:10.1002/9780470685853.ch4.
- Krzysztofowicz, R. (2010), Decision criteria, data fusion and prediction calibration: A Bayesian approach, *Hydrol. Sci. J.*, 55(6), 1033–1050, doi:10.1080/02626667.2010.505894.
- Li, X., and F. T.-C. Tsai (2009), Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod, *Water Resour. Res.*, 45, W09403, doi:10.1029/2008WR007488.
- Linssen, H. N. (1975), Nonlinearity measures: A case study, *Stat. Neerl.*, 29, 93–99, doi:10.1111/j.1467-9574.1975.tb00253.x.
- Liu, X., M. A. Cardiff, and P. K. Kitanidis (2010), Parameter estimation in nonlinear environmental problems, *Stochastic Environ. Res. Risk Assess.*, 24, 1003–1022, doi:10.1007/s00477-010-0395-y.
- Lu, D., M. Ye, S. P. Neuman, and L. Xue (2012), Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs, *Adv. Water Resour.*, 35, 69–82, doi:10.1016/j.advwatres.2011.10.007.
- McClave, J. T., and T. Sincich (2000), *Statistics*, 8th ed., Prentice Hall, Upper Saddle River, N.J.
- Mehl, S. W., and M. C. Hill (2010), Grid-size dependence of Cauchy boundary conditions used to simulate stream-aquifer interactions, *Adv. Water Resour.*, 33, 430–442, doi:10.1016/j.advwatres.2010.01.008.
- Meyer, P. D., M. Ye, M. L. Rockhold, S. P. Neuman, and K. J. Cantrell (2007), *Combined Estimation of Hydrogeologic Conceptual Model, Parameter, and Scenario Uncertainty*, NUREG/CR-6940, PNNL-16396, U.S. Nucl. Regul. Comm., Washington, D.C.
- Montanari, A., and A. Brath (2004), A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, 40, W01106, doi:10.1029/2003WR002540.
- Montanari, A., and G. Grossi (2008), Estimating the uncertainty of hydrological forecasts: A statistical approach, *Water Resour. Res.*, 44, W00B08, doi:10.1029/2008WR006897.
- Morgan, M. G., and M. Henrion (1990), *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge Univ. Press, Cambridge, U.K., doi:10.1017/CBO9780511840609.
- Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, *Stochastic Environ. Res. Risk Assess.*, 17(5), 291–305, doi:10.1007/s00477-003-0151-7.
- Neuman, S. P., and P. J. Wierenga (2003), *A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites*, NUREG/CR-6805, U.S. Nucl. Regul. Comm., Washington, D.C.
- Neuman, S. P., L. Xue, M. Ye, and D. Lu (2012), Bayesian analysis of data-worth considering model and parameter uncertainties, *Adv. Water Resour.*, doi:10.1016/j.advwatres.2011.02.007, in press.
- Oliver, D. S., A. C. Reynolds, and N. Liu (2008), *Inverse Theory for Petroleum Reservoir Characterization and History Matching*, 392 pp., Cambridge Univ. Press, New York.
- Parker, J. C., U. Kim, M. Widdowson, P. Kitanidis, and R. Gentry (2010), Effects of model formulation and calibration data on uncertainty in dense nonaqueous phase liquids source dissolution predictions, *Water Resour. Res.*, 46, W12517, doi:10.1029/2010WR009361.
- Poeter, E. P., and D. A. Anderson (2005), Multimodel ranking and inference in ground water modeling, *Ground Water*, 43(4), 597–605, doi:10.1111/j.1745-6584.2005.0061.x.
- Poeter, E. P., and M. C. Hill (2007), MMA, A computer code for multimodel analysis, *U.S. Geol. Surv. Tech. Methods*, 6–E3.
- Poeter, E. P., M. C. Hill, E. R. Banta, S. W. Mehl, and S. Christensen (2005), UCODE 2005 and six other computer codes for universal sensitivity analysis, inverse modeling, and uncertainty evaluation, *U.S. Geol. Surv. Tech. Methods*, 6–A11.
- Razavi, S., B. A. Tolson, and D. H. Burn (2012), Review of surrogate modeling in water resources, *Water Resour. Res.*, 48, W07401, doi:10.1029/2011WR011527.
- Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, *Water Resour. Res.*, 47, W11516, doi:10.1029/2011WR010643.
- Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2008), *Global Sensitivity Analysis, The Primer*, John Wiley and Sons, New York.
- Samanta, S., D. S. Mackay, M. K. Clayton, E. L. Kruger, and B. E. Ewers (2007), Bayesian analysis for uncertainty estimation of a canopy transpiration model, *Water Resour. Res.*, 43, W04424, doi:10.1029/2006WR005028.
- Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, 46, W10531, doi:10.1029/2009WR008933.
- Schweppe, F. C. (1973), *Uncertainty Dynamic Systems*, Prentice Hall, Englewood Cliffs, N.J.
- Sciortino, A., T. C. Harmon, and W. W.-G. Yeh (2002), Experimental design and model parameter estimation for locating a dissolving dense nonaqueous phase liquid pool in groundwater, *Water Resour. Res.*, 38(5), 1057, doi:10.1029/2000WR000134.
- Seber, G. A. F., and A. J. Lee (2003), *Linear Regression Analysis*, 549 pp., Wiley, New York, doi:10.1002/9780471722199.
- Seber, G. A. F., and C. J. Wild (2003), *Nonlinear Regression*, 768 pp., Wiley, New York.
- Shi, X., M. Ye, S. Finsterle, and J. Wu (2012), Comparing nonlinear regression and Markov chain Monte Carlo methods for assessment of predictive uncertainty in vadose zone modeling, *Vadose Zone J.*, doi:10.2136/vzj2011.0147, in press.
- Stark, P. B., and L. Tenorio (2011), A primer of frequentists and Bayesian inference in inverse problems, in *Large-scale Inverse Problems and Quantification of Uncertainty*, edited by L. Biegler et al., pp. 9–32, Wiley, New York.
- Sun, A. Y., A. P. Morris, and S. Mohanty (2009), Sequential updating of multimodal hydrogeologic parameter fields using localization and clustering techniques, *Water Resour. Res.*, 45, W07424, doi:10.1029/2008WR007443.
- Tang, Y., P. Reed, T. Wagener, and K. van Werkhoven (2007), Comparing sensitivity analysis methods to advance lumped watershed model identification and calibration, *Hydrol. Earth Syst. Sci.*, 11, 793–817, doi:10.5194/hess-11-793-2007.
- Tiedeman, C. R., M. C. Hill, F. A. D'Agnese, and C. C. Faunt (2003), 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.
- Tiedeman, C. R., D. M. Ely, M. C. Hill, and G. M. O'Brien (2004), A method for evaluating the importance of system state observations to model predictions, with application to the Death Valley regional groundwater flow system, *Water Resour. Res.*, 40, W12411, doi:10.1029/2004WR003313.
- Tonkin, M., C. R. Tiedeman, D. M. Ely, and M. C. Hill (2007), OPR-PPR, a computer program for assessing data importance to model predictions using linear statistics, *U.S. Geol. Surv. Tech. Methods*, 6–E2, 115 pp. [Available at <http://water.usgs.gov/software/OPR-PPR.html>.]
- Vecchia, A. V., and R. L. Cooley (1987), Simultaneous confidence and prediction intervals for nonlinear regression models with application to a groundwater flow model, *Water Resour. Res.*, 23(7), 1237–1250, doi:10.1029/WR023i007p01237.
- Vrugt, J. A., and W. Bouten (2002), Validity of first-order approximations to describe parameter uncertainty in soil hydraulic models, *Soil Sci. Soc. Am. J.*, 66, 1740–1751, doi:10.2136/sssaj2002.1740.
- Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling:

- Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.
- Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon (2009), Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*(3), 271–288.
- Wang, Q. J., D. E. Robertson, and F. H. S. Chiew (2009), A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, *Water Resour. Res.*, *45*, W05407, doi:10.1029/2008WR007355.
- Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*, W05113, doi:10.1029/2003WR002557.
- Yeh, W. W.-G. (1986), Review of parameter identification procedures in groundwater hydrology: The inverse problem, *Water Resour. Res.*, *22*(2), 95–108, doi:10.1029/WR022i002p00095.
- Yeh, W. W.-G., and Y. S. Yoon (1981), Aquifer parameter identification with optimum dimension in parameterization, *Water Resour. Res.*, *17*(3), 664–672, doi:10.1029/WR017i003p00664.

Appendix C

This appendix contains the paper of Lu et al. (2013), which was published in Water Resources Research. This research helps answer the questions below explained in detail in the report:

Question 4: How to quantify model uncertainty of groundwater reactive transport modeling?

It in particular answers the question below regarding the evaluation of model averaging weights for groundwater reactive transport models:

What is the impact of disregarding the residual correlation and how to incorporate the correlation in the evaluation of model averaging weights?

Effects of error covariance structure on estimation of model averaging weights and predictive performance

Dan Lu,¹ Ming Ye,¹ Philip D. Meyer,² Gary P. Curtis,³ Xiaoqing Shi,^{1,4} Xu-Feng Niu,⁵ and Steve B. Yabusaki²

Received 14 January 2013; revised 25 June 2013; accepted 23 July 2013.

[1] When conducting model averaging for assessing groundwater conceptual model uncertainty, the averaging weights are often evaluated using model selection criteria such as *AIC*, *AICc*, *BIC*, and *KIC* (Akaike Information Criterion, Corrected Akaike Information Criterion, Bayesian Information Criterion, and Kashyap Information Criterion, respectively). However, this method often leads to an unrealistic situation in which the best model receives overwhelmingly large averaging weight (close to 100%), which cannot be justified by available data and knowledge. It was found in this study that this problem was caused by using the covariance matrix, C_e , of measurement errors for estimating the negative log likelihood function common to all the model selection criteria. This problem can be resolved by using the covariance matrix, C_{ek} , of total errors (including model errors and measurement errors) to account for the correlation between the total errors. An iterative two-stage method was developed in the context of maximum likelihood inverse modeling to iteratively infer the unknown C_{ek} from the residuals during model calibration. The inferred C_{ek} was then used in the evaluation of model selection criteria and model averaging weights. While this method was limited to serial data using time series techniques in this study, it can be extended to spatial data using geostatistical techniques. The method was first evaluated in a synthetic study and then applied to an experimental study, in which alternative surface complexation models were developed to simulate column experiments of uranium reactive transport. It was found that the total errors of the alternative models were temporally correlated due to the model errors. The iterative two-stage method using C_{ek} resolved the problem that the best model receives 100% model averaging weight, and the resulting model averaging weights were supported by the calibration results and physical understanding of the alternative models. Using C_{ek} obtained from the iterative two-stage method also improved predictive performance of the individual models and model averaging in both synthetic and experimental studies.

Citation: Lu, D., M. Ye, P. D. Meyer, G. P. Curtis, X. Shi, X.-F. Niu, and S. B. Yabusaki (2013), Effects of error covariance structure on estimation of model averaging weights and predictive performance, *Water Resour. Res.*, 49, doi:10.1002/wrcr.20441.

1. Introduction

[2] Considerable progress has been made in the past three decades on uncertainty quantification in environmental modeling [Liu and Gupta, 2007; Matott et al., 2009; Tartakovsky, 2013; and references therein]. Initially, the emphasis has been on uncertainty in model parameters. A more recent trend has been to consider uncertainties in both model structures and pa-

rameters [Ye et al., 2010a; Gupta et al., 2012]. This has been motivated by a growing recognition that environmental systems are open and complex, rendering them prone to multiple conceptualizations and mathematical descriptions, regardless of the quantity and quality of available data and knowledge [Beven, 2002; Bredehoeft, 2003, 2005; Neuman, 2003]. Multi-model analysis has become popular for quantification of model uncertainty [Burnham and Anderson, 2002; Ye et al., 2004, 2005, 2008a, 2008b, 2010b, 2010c; Poeter and Anderson, 2005; Marshall et al., 2005; Beven, 2006; Foglia et al., 2007; Ajami et al., 2007; Vrugt and Robinson, 2007; Tsai and Li, 2008a, 2008b; Wohling and Vrugt, 2008; Rojas et al., 2008, 2009; Rubin et al., 2010; Winter and Nychka, 2010; Riva et al., 2011; Neuman et al., 2012; Lu et al., 2011, 2012; Nowak et al., 2012; Seifert et al., 2012; Rings et al., 2012; Parrish et al., 2012; Dai et al., 2012]. In multimodel analysis, rather than choosing a single model, modeling predictions and associated uncertainty from multiple competing models are aggregated, typically in a model averaging process. Consider a set of models, $\mathbf{M} = (M_1, \dots, M_K)$, and denote $\hat{y}_k$ as a prediction (e.g., mean prediction or probability distribution) of model M_k for a

¹Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA.

²Pacific Northwest National Laboratory, Richland, Washington, USA.

³U.S. Geological Survey, Menlo Park, California, USA.

⁴School of Earth Sciences and Engineering, Nanjing University, Nanjing, Jiangsu, China.

⁵Department of Statistics, Florida State University, Tallahassee, Florida, USA.

Corresponding author: M. Ye, Department of Scientific Computing, Florida State University, Tallahassee, FL 32306-4120, USA. (mye@fsu.edu)

quantity of interest. The weighted average estimate, $\hat{y}$, of the prediction is

$$\hat{y} = \sum_{k=1}^K w_k \hat{y}_k, \quad (1)$$

where w_k is the averaging weight associated with model M_k , the most critical variable to be estimated in the process of model averaging. It is still an open question how to estimate the averaging weights with mathematical and statistical rigor and computational efficiency.

[3] This study is focused on evaluating model averaging weights using

$$w_k = \frac{\exp(-\Delta IC_k/2)}{\sum_{l=1}^K \exp(-\Delta IC_l/2)}, \quad (2)$$

where the IC (Information Criteria) are various model selection criteria, and $\Delta IC_k = IC_k - IC_{\min}$ is the difference between the IC of model M_k and the minimum IC , $IC_{\min}$. Four model selection criteria are considered in this study: AIC [Akaike, 1974], $AICc$ [Hurvich and Tsai, 1989], BIC [Schwarz, 1978], and KIC [Kashyap, 1982]. They are defined for model M_k as [Ye et al., 2008a]

$$AIC_k = -2\ln [L(\hat{\theta}_k|\mathbf{D})] + 2N_k \quad (3)$$

$$AICc_k = -2\ln [L(\hat{\theta}_k|\mathbf{D})] + 2N_k + \frac{2N_k(N_k + 1)}{N - N_k - 1} \quad (4)$$

$$BIC_k = -2\ln [L(\hat{\theta}_k|\mathbf{D})] + N_k \ln N \quad (5)$$

$$KIC_k = -2\ln [L(\hat{\theta}_k|\mathbf{D})] - 2\ln p(\hat{\theta}_k) - N_k \ln(2\pi) + \ln |\mathbf{F}_k| \quad (6)$$

where $\hat{\theta}_k$ is the maximum likelihood (ML) estimate of a vector θ_k of N_k adjustable parameters (which may include statistical parameters of the calibration data) associated with model M_k ; $\mathbf{D}$ is a vector of N observations collected in space time; $-\ln [L(\hat{\theta}_k|\mathbf{D})]$ is the minimum of the negative log likelihood (NLL) function, $-\ln [L(\theta_k|\mathbf{D})]$, occurring, by definition, at $\hat{\theta}_k$; $p(\hat{\theta}_k)$ is the prior probability of θ_k evaluated at $\hat{\theta}_k$; and $\mathbf{F}_k$ is the observed (implicitly conditioned on the observations $\mathbf{D}$ and evaluated at the maximum likelihood parameter estimates $\hat{\theta}_k$) Fisher information matrix having elements [Kashyap, 1982]

$$F_{k,ij} = -\frac{\partial^2 \ln [L(\theta_k|\mathbf{D})]}{\partial \theta_{ki} \partial \theta_{kj}} \Big|_{\theta_k = \hat{\theta}_k}. \quad (7)$$

[4] Models associated with smaller values of a given criterion are ranked higher than those associated with larger values and correspondingly assigned larger model averaging weights; the absolute value of the criterion being irrelevant. As shown in Neuman [2003] and Ye et al. [2008a], model averaging weight calculated using KIC is a maximum likelihood (ML) approximation to posterior model probability of Bayesian model averaging (BMA) described in Hoeting et al. [1999]. Therefore, BMA based on the model selection criteria is referred to as MLBMA hereinafter.

[5] The model selection criteria have been widely used in groundwater modeling for both model selection and model averaging, and they are default outputs of popular software of groundwater inverse modeling such as PEST [Doherty, 2005], UCODE [Poeter et al., 2005], iTOUGH2 [Finsterle, 2007], and MMA [Poeter and Hill, 2007; Ye, 2010]. Their popularity in model selection is due to their quantitative representation of the principle of parsimony. The first term of each criterion, $-2\ln [L(\hat{\theta}_k|\mathbf{D})]$, measures goodness-of-fit between predicted and observed data, $\mathbf{D}$; the smaller this term, the better the fit. The terms containing N_k represent measures of model complexity. The criteria thus embody (to various degrees) the principle of parsimony by penalizing models for having a relatively large number of parameters if this does not bring about a corresponding improvement in model fit. Their popularity in model averaging is due to their relative ease of computation and computational efficiency, particularly, in comparison with other methods that use Monte Carlo (MC) methods to calculate model averaging weights.

[6] However, the model selection criteria in equations (3)–(6) aggressively exclude inferior models with relatively large ΔIC values. For example, models receive less than 5% probability if their ΔIC values are larger than 6. Application of the criteria to hydrologic modeling has sometimes led to the assignment of close to 100% of the averaging weight to one model when available data and knowledge suggest that exclusion of other competing models is unjustifiable. For example, Meyer et al. [2007] developed four models simulating uranium transport at the Hanford Site 300 Area of the U.S. Department of Energy (DOE). All the model selection criteria assigned almost 100% averaging weight to a single model, whereas this model was not significantly superior to the other three models for matching calibration data. Singh et al. [2010] encountered a similar situation, when working with nine models developed for one of the corrective action units of the DOE Nevada National Security Site (NNSS), USA. For another NNSS corrective action unit, Pohlmann et al. [2007] and Ye et al. [2010a] considered 25 groundwater models, each of which has different recharge components and hydrostratigraphic frameworks. Based on the four model selection criteria, only two models received significant weights, and the weights of the other 23 models were negligible. However, evaluating the models based on expert judgment [Ye et al., 2008b] and examining calibration results of the models did not support discarding all 23 models (though it was reasonable to discard some). Similar situations occurred in Morales-Casique et al. [2010] when studying a number of geostatistical and air flow models, in Diks and Vrugt [2010] for two cases that involved eight watershed models and seven soil hydraulic models, respectively, and in Seifert et al. [2012] for six hydrological models with different conceptual geological configurations.

[7] Tsai and Li [2008a] proposed to address the problem of unjustifiable assignment of model averaging weight to a single model by calculating weights via

$$w_k = \frac{\exp(-\alpha \Delta IC_k/2)}{\sum_{l=1}^K \exp(-\alpha \Delta IC_l/2)}. \quad (8)$$

where α is a subjective factor. *Tsai and Li* [2008a] gave several examples in which the averaging weight of a single model was reduced from 100% to as little as 60% using reasonable values of α . As shown below, however, use of (8) with a similar value of α did not solve the problem that one model unreasonably receives 100% weight in our numerical experiments.

[8] *Diks and Vrugt* [2010] evaluated a number of methods for estimating model averaging weights that did not assign 100% weight to a single model in either of their two applications. Several of these methods allow negative model weights. As observed by *Raftery et al.* [2005], negative weights can be difficult to interpret since they imply a negative correlation between a model's simulated value and the predicted (model average) value. In addition, only positive weights can be used when calculating a model average probability density (to avoid negative densities). According to *Diks and Vrugt* [2010], the model averaging weights proposed by *Bates and Granger* [1969] had predictive performance significantly worse than the use of *AIC* or *BIC*. The other methods evaluated in *Diks and Vrugt* [2010] allow only positive model weights and had better predictive performance than the model selection criteria of *AIC* and *BIC*. These methods included Bayesian Model Averaging (BMA) with a likelihood based on a finite mixture model [*Raftery et al.*, 2005], BMA with a likelihood based on a linear regression model (with weights constrained to be positive) [*Raftery et al.*, 1997], and Mallows model averaging [*Hjort and Claeskens*, 2003; *Hansen*, 2007] (with weights constrained to be positive). With these methods, model weights are determined by fitting the model average result to the calibration data. This is in contrast to the use of equation (2) in which model selection criteria (and therefore the weights themselves) are determined on the basis of each individual model's fit to the calibration data and on complexity of the individual models. Unlike the model selection criteria, the BMA methods evaluated in *Diks and Vrugt* [2010] do not include a term representing model complexity. *Ye et al.* [2008a, 2010c] showed that model averaging weights determined from equation (6) have a rigorous mathematical basis in the context of the BMA method of *Hoeting et al.* [1999]. A comparative study with the BMA method of *Raftery et al.* [1997, 2005] is needed to better understand the theoretical and numerical similarities and differences. Similarly, the investigation of error correlation in this study is expected to be of general use to other model averaging methods. For example, it could be included in the log likelihood function of the BMA method of *Raftery et al.* [2005], in which independence of forecast errors in space and time is explicitly assumed. These additional studies, however, are beyond the scope of this paper.

[9] As described in the remainder of this paper, the problem of assigning unreasonably large model averaging weight to a single model when using the model selection criteria is caused by disregarding the correlation between total errors (including model errors and measurement errors) for calculation of the *NLL* term ($-\ln[L(\hat{\theta}_k|\mathbf{D})]$) common to all the model selection criteria. As discussed in section 2 below, the error correlation, reflected in the covariance structure used for maximum likelihood model calibration, affects the calculation of *NLL*, and correspondingly the evaluation of the

model averaging weights. To resolve this problem for temporal data, an iterative two-stage parameter estimation method is developed and introduced in section 3 to incorporate total error correlation into the covariance matrix of model calibration and *NLL* evaluation. The method is evaluated using synthetic data in section 4 and then applied to an experimental study in section 5 for estimating model averaging weights of several surface complexation models developed to simulate column experiments of hexavalent uranium [U(VI)]. The effects of disregarding the correlation of total errors on model averaging weights and predictive performance of individual models and model averages are evaluated for the synthetic and experimental studies.

2. Effect of Error Covariance Structure on Model Averaging Weights

[10] In this section, the negative log likelihood (*NLL*) function is first defined, followed by a discussion of the effect of error covariance structure on estimation of *NLL* and model averaging weights. The effect is illustrated at the end of this section using a simple example.

2.1. Error Covariance Structure in Evaluation of *NLL*

[11] Let $g(\boldsymbol{\theta})$ be a model that closely approximates the true system with negligible model errors, and use $\boldsymbol{\varepsilon}$ to denote measurement errors associated with observations $\mathbf{D}$. One can write the observation, $\mathbf{D}$, as

$$\mathbf{D} = g(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}. \quad (9)$$

[12] Assume that $\boldsymbol{\varepsilon}$ has a multivariate Gaussian distribution with zero mean and covariance matrix $\mathbf{C}_\varepsilon$ (non-Gaussian distribution can be transformed to a Gaussian form, though challenging sometimes), then the *NLL* term in equations (3)–(6) corresponding to the maximum likelihood parameter estimate $\hat{\boldsymbol{\theta}}$ is

$$NLL = -2\ln \left[L(\hat{\boldsymbol{\theta}}|\mathbf{D}) \right] = N\ln(2\pi) + \ln|\mathbf{C}_\varepsilon| + \mathbf{r}^T \mathbf{C}_\varepsilon^{-1} \mathbf{r}, \quad (10)$$

where $\mathbf{r}$ is the residual vector, i.e., the difference between observations and simulations, $|\mathbf{C}_\varepsilon|$ is determinant of $\mathbf{C}_\varepsilon$, and $\mathbf{C}_\varepsilon^{-1}$ is weight matrix used for calculating the sum of square weighted residuals (*SSWR*), i.e., $SSWR = \mathbf{r}^T \mathbf{C}_\varepsilon^{-1} \mathbf{r}$. The formulation of equation (10) invokes a number of assumptions as discussed in *Finsterle and Zhang* [2011]. When the model is well calibrated, the probability structure of $\mathbf{r}$, should be similar to that of $\boldsymbol{\varepsilon}$. If the measurement errors are correlated, the off-diagonal elements in the covariance matrix $\mathbf{C}_\varepsilon$ are not zeros. In practice, $\mathbf{C}_\varepsilon$ is generally diagonal, since measurement errors are most commonly taken to be uncorrelated [*Carrera and Neuman*, 1986]. *Tiedeman and Green* [2013] presented a special case, in which calibration data were calculated from multiple direct measurements and correlation between the estimated calibration data needs to be incorporated into model calibration and uncertainty quantification.

[13] When a model cannot accurately simulate the true system and relatively large model errors exist, observations $\mathbf{D}$ cannot be sufficiently explained by the model and measurement errors as in equation (9). Given a number of such

Figure 1. Illustration of effect of error covariance structure on the evaluation of $SSWR$, sum of squared weighted residuals. Point O represents the two observations D_1 and D_2 ; Points A and B are simulations of the observations by models A and B . The ellipses in solid red and blue lines are the $SSWR$ contours based on matrices $\mathbf{C}_{e_A}$ and $\mathbf{C}_{e_B}$ of the total error of models A and B , respectively. The ellipses in the dashed red and blue lines are the $SSWR$ contours based on the matrix $\mathbf{C}_e$ of the measurement errors.

models considered in multimodel analysis, use $f_k(\boldsymbol{\beta}_k)$ to denote model M_k with parameters $\boldsymbol{\beta}_k$. The difference, $\boldsymbol{\eta}_k = g(\boldsymbol{\theta}) - f_k(\boldsymbol{\beta}_k)$, is defined as the model error, the imperfections associated with $f_k(\boldsymbol{\beta}_k)$. Combining the model errors, $\boldsymbol{\eta}_k$, with the measurement errors, $\boldsymbol{\varepsilon}$, gives the total errors $\mathbf{e}_k$. The observation vector, $\mathbf{D}$, is thus written as

$$\mathbf{D} = f_k(\boldsymbol{\beta}_k) + \mathbf{e}_k = f_k(\boldsymbol{\beta}_k) + \boldsymbol{\eta}_k + \boldsymbol{\varepsilon}. \quad (11)$$

[14] Assume that the joint probability distribution function of the total errors, $\mathbf{e}_k$, is multivariate Gaussian with zero mean and covariance matrix $\mathbf{C}_{e_k}$. In practice, this assumption can be verified ad hoc by analyzing the residuals after model calibration, as shown in section 4. Further assume that the covariance matrix can be characterized by parameters, $\mathbf{a}_k$, then the NLL term in equations (3)–(6) of alternative model M_k evaluated at ML estimates is

$$NLL = -2\ln [L(\hat{\boldsymbol{\beta}}_k, \hat{\mathbf{a}}_k | \mathbf{D})] = N\ln(2\pi) + \ln |\mathbf{C}_{e_k}| + \mathbf{r}_k^T \mathbf{C}_{e_k}^{-1} \mathbf{r}_k. \quad (12)$$

[15] While equation (12) falls into the general framework of error-based weighting in model calibration [Hill and Tiedeman, 2007; Foglia et al., 2009] and correlation between total errors was extensively studied in Cooley and Christensen [2006] and Christensen and Doherty [2008], it appears to be the first time that the covariance matrix of total errors is estimated through the two-stage procedure described in the next section. Since model errors, and thus

total errors, are different for different alternative models, covariance matrix $\mathbf{C}_{e_k}$ has different correlation structures for different models.

[16] While the probability structure of the total errors, $\mathbf{e}_k$, is unknown, it can be inferred from the probability structure of the residuals, $\mathbf{r}_k$, after model M_k is calibrated [Finsterle and Zhang, 2011]. The residuals and total errors are distinguished in this study, because the residuals are caused not only by the model errors and measurement errors (i.e., the total errors) that cannot diminish during model calibration but also by parameter estimation errors that may gradually diminish during model calibration when calibrated parameter values approach to their optimum values. In other words, after model calibration, the misfit between observations and model simulations is caused by the total errors and cannot be further reduced by adjusting model parameters. Since the total errors and parameter estimation errors cannot be explicitly separated, inferring the probability structure of total errors from that of the residuals can only be performed in an iterative manner. This is the basis of the iterative two-stage method described in section 3 for serial data.

[17] Due to the model errors, the covariance matrix $\mathbf{C}_{e_k}$ is likely to be a full matrix with off-diagonal terms representing the correlations between the total errors, which was shown in a numerical study of Xu et al. [2012] using the Republican River Compact Administration model and the Spokane Valley-Rathdrum Prairie model of groundwater flow modeling. Doherty and Welter [2010] demonstrated that the level of model errors can be similar to or larger than that of measurement errors, and that total errors are expected to have a high degree of spatial and/or temporal correlation. When the total errors are significantly larger than the measurement errors and the total errors are correlated, disregarding the total errors (random or systematic) may cause convergence problems, give biased parameter estimates, and/or lead to poor predictive capabilities and misleading predictive uncertainty measures [Finsterle and Zhang, 2011]. It is demonstrated below that using equation (10), instead of equation (12), may yield inaccurate NLL and, consequently, wrong model averaging weights. It is worth mentioning that the way of handling error correlation is not limited to the Gaussian likelihood function. A more comprehensive study of handling error structure is referred to Schoups and Vrugt [2010], in which error correlation, heteroscedasticity, and non-Gaussianity were all considered and characterized using the skew exponential power density function.

2.2. Illustration Using a Simple Example

[18] Figure 1 illustrates the effect of using inappropriate error covariance structure (i.e., replacing the full covariance matrix of the total errors with the diagonal matrix of the measurement errors) on calculation of model averaging weights. The illustration considers a simple case with only two observations, $D_1 = 10$ and $D_2 = 30$ (plotted as point O in Figure 1) with the covariance matrix of measurement errors, $\mathbf{C}_e$, as

$$\mathbf{C}_e = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}. \quad (13)$$

[19] Consider two models, A and B . After the maximum likelihood parameter estimation, the model simulations of

D_1 and D_2 are ($\hat{D}_{A,1} = 9.11$, $\hat{D}_{A,2} = 30.89$) and ($\hat{D}_{B,1} = 11.96$, $\hat{D}_{B,2} = 30.62$), respectively, which are plotted as points A and B in Figure 1. For the purpose of demonstration, assume the following covariance matrix of total errors:

$$\mathbf{C}_{e_A} = \begin{bmatrix} 3.2 & 1.6 \\ 1.6 & 3.2 \end{bmatrix}, \text{ and } \mathbf{C}_{e_B} = \begin{bmatrix} 4.2 & 2.4 \\ 2.4 & 4.2 \end{bmatrix}. \quad (14)$$

[20] Although all the values above are chosen for the convenience of plotting Figure 1, the assumed covariance matrix in equation (14) can be calculated from the residuals based on the procedure discussed in section 3.

[21] It is demonstrated first the effect of using $\mathbf{C}_e$ instead of $\mathbf{C}_{e_k}$ on evaluating the goodness-of-fit of the two models as measured by $SSWR = \mathbf{r}_k^T \mathbf{C}_{e_k}^{-1} \mathbf{r}_k$. When the full covariance matrix of total errors, $\mathbf{C}_{e_A}$ is used, $SSWR$ is

$$MD_{M_k}^2 = \mathbf{r}_k^T \mathbf{C}_{e_k}^{-1} \mathbf{r}_k, \quad (15)$$

i.e., the Mahalanobis distance (MD) [Mahalanobis, 1936], and the quadratic form defines an ellipse in the data space. The residuals and covariance values above lead to

$$\begin{aligned} MD_A^2 &= \mathbf{r}_A^T \mathbf{C}_{e_A}^{-1} \mathbf{r}_A = 1.00 \\ MD_B^2 &= \mathbf{r}_B^T \mathbf{C}_{e_B}^{-1} \mathbf{r}_B = 1.00, \end{aligned} \quad (16)$$

which corresponds to the two $SSWR$ contours of solid lines (red for model A and blue for model B) plotted in Figure 1. Although the ellipses are different, the Mahalanobis distance between points A and O is the same as that between points B and O , suggesting that the two models perform equally well in terms of fitting to the observations. However, when the covariance matrix of measurement errors, $\mathbf{C}_e$, is used, $SSWR$ is

$$ED_{M_k}^2 = \mathbf{r}_k^T \mathbf{C}_e^{-1} \mathbf{r}_k, \quad (17)$$

i.e., the normalized Euclidian distance (ED), and they are

$$\begin{aligned} ED_A^2 &= \mathbf{r}_A^T \mathbf{C}_e^{-1} \mathbf{r}_A = 1.07 \\ ED_B^2 &= \mathbf{r}_B^T \mathbf{C}_e^{-1} \mathbf{r}_B = 3.96 \end{aligned} \quad (18)$$

based on the residual and covariance values above. This quadratic form defines another type of ellipses whose major and minor axes are aligned with the x and y axis, because the covariance matrix is diagonal. The ellipses of models A and B are plotted in dashed red and blue lines, respectively, in Figure 1. The figure shows that the normalized Euclidean distances of the two models are dramatically different and give wrong measures of the goodness-of-fit.

[22] Take BIC as an example for calculation of the model averaging weights. Without loss of generality, further assume that the number of calibrated parameters, N_k , is the same for the two models. When $\mathbf{C}_{e_k}$ is used, $BIC_A - BIC_B = \ln |\mathbf{C}_{e_A}| - \ln |\mathbf{C}_{e_B}| + \mathbf{r}_A^T \mathbf{C}_{e_A}^{-1} \mathbf{r}_A - \mathbf{r}_B^T \mathbf{C}_{e_B}^{-1} \mathbf{r}_B$, and the averaging weights of models A and B are 55.4 and 44.6%, respectively. The difference in model averaging weights is determined by the difference (-0.43) in $\ln |\mathbf{C}_{e_k}|$. However, when $\mathbf{C}_e$ is used, $BIC_A - BIC_B = \mathbf{r}_A^T \mathbf{C}_e^{-1} \mathbf{r}_A -$

$\mathbf{r}_B^T \mathbf{C}_e^{-1} \mathbf{r}_B$, and the averaging weights of models A and B become 80.9 and 19.1%, because the normalized Euclidian distance exaggerates the misfit of model B and distorts relative plausibility of the two models. The distortion increases as the covariance of total errors deviates more from that of measurement errors. For KIC , the $\ln |\mathbf{C}_{e_k}|$ term contributes not only to the calculation of NLL but also to the evaluation of the Fisher information matrix, considering that the Fisher information matrix is the inverse of the covariance matrix of parameter estimation uncertainty and that the covariance matrix is often estimated via $(\mathbf{X}_k^T \mathbf{C}_{e_k}^{-1} \mathbf{X}_k)^{-1}$, where $\mathbf{X}_k$ is sensitivity matrix of observations. Doherty and Welter [2010] cautioned that “if computation of postcalibration statistics such as AIC , $AICc$, BIC , and KIC ignores the unavoidable presence of structural noise of unknown covariance matrix that accompanies the use of any model (even a perfect model), then recommendations made on the basis of these statistics that favor one model over another, or one parameterization scheme over another, should be taken as suggestive rather than definitive.” The key issues to resolve this problem involve identification of the covariance matrix structure (full or diagonal), estimation of its characteristic parameters, and incorporation of the covariance matrix into model calibration and averaging weights evaluation. An iterative two-stage method is developed for serial data and introduced in the next section. The developed method calculates the Mahalanobis distance and uses it to evaluate model averaging weights.

3. Iterative Two-Stage Method for Serial Data

[23] For a time series of observations, $D_t = \{D_1, D_2, \dots, D_N\}$, measured at a sequence of discrete times $t = 1, 2, \dots, N$ (e.g., observations of flows, heads, and/or concentrations collected over time), analogous to equation (11), we have

$$D_t = f_k(\boldsymbol{\beta}_k) + e_{k,t}, \quad t = 1, 2, \dots, N. \quad (19)$$

[24] Several discrete stochastic time series models are commonly used to simulate temporal correlation of the total errors $e_{k,t}$, including the p th-order autoregressive model, $AR(p)$, the q th-order moving average model, $MA(q)$, and/or the mixed autoregressive moving averaging model, $ARMA(p,q)$ [Chatfield, 1989]. The $AR(p)$ model is used in this study, as it is shown to be appropriate for the groundwater reactive transport problems discussed below. Once the time series model is determined, the covariance matrix, $\mathbf{C}_{e_k}$, of the total errors can be constructed. For example, if an $AR(1)$ model is used, then $e_{k,t}$ can be quantified as $e_{k,t} = ae_{k,t-1} + \xi_t$, where a is the parameter and $\{\xi_t\}$ is a vector of white noise with mean zero and constant variance. If the series $\{e_{k,t}\}$ is stationary with constant variance σ^2 (the stationarity assumption can be verified by examining the parameter coefficients of the time series models as shown in section 4), the covariance matrix, $\mathbf{C}_{e_k}$, can be expressed as

$$\mathbf{C}_{e_k} = \sigma^2 \mathbf{V}, \quad (20)$$

where $\mathbf{V}$ is the correlation matrix in the form of

$$\mathbf{V} = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{N-1} \\ \rho_1 & & & \vdots \\ \vdots & & & \rho_1 \\ \rho_{N-1} & \cdots & \rho_1 & 1 \end{pmatrix} \quad (21)$$

with ρ_l (where $l = 1, 2, \dots, N-1$) representing the correlation coefficient at lag l . For the AR(1) model, since $\rho_1 = a$ and $\rho_l = \rho_1^l$ [Seber and Wild, 2003, p. 275], matrix $\mathbf{V}$ can be expressed as

$$\mathbf{V} = \begin{pmatrix} 1 & a & \cdots & a^{N-1} \\ a & & & \vdots \\ \vdots & & & a \\ a^{N-1} & \cdots & a & 1 \end{pmatrix}. \quad (22)$$

[25] The time series parameters, $\mathbf{a}_k = [a, \sigma^2]$, can be estimated simultaneously with parameters $\boldsymbol{\beta}_k$ of the deterministic model $f_k(\boldsymbol{\beta}_k)$ by minimizing the likelihood function,

$$\begin{aligned} NLL &= -2\ln[L(\boldsymbol{\beta}_k, \mathbf{a}_k | \mathbf{D})] \\ &= N \ln(2\pi) + \ln |\mathbf{C}_{e_k}(\mathbf{a}_k)| + (\mathbf{D} - f_k(\boldsymbol{\beta}_k))^T \mathbf{C}_{e_k}^{-1}(\mathbf{a}_k) (\mathbf{D} - f_k(\boldsymbol{\beta}_k)), \end{aligned} \quad (23)$$

which is similar to that of Kuczera [1983] and assumes that the joint probability distribution of the total errors is multivariate Gaussian with zero mean and the covariance matrix $\mathbf{C}_{e_k}$. This, however, requires determining the time series model *a priori* [Sorooshian and Dracup, 1980]. In a recent study, Schoups and Vrugt [2010] found that a fixed time series model may not be adequate to account for the error correlation and that adaptation of the time series model is necessary, for example, by changing the order of time series models. It is particularly true for this study, since the error correlation is unknown before model calibration starts and varies during the calibration. Therefore, model parameters, $\boldsymbol{\beta}_k$, and time series parameters, $\mathbf{a}_k$, are estimated separately in this study using the iterative two-stage method described below.

3.1. Iterative Two-Stage Parameter Estimation Method

[26] The iterative two-stage method is built on the basis that temporal residual analysis can be used as a means to infer the covariance matrix, $\mathbf{C}_{e_k}$, of the total errors. Hereinafter, all the analyses are for residuals, with the ultimate goal of estimating $\mathbf{C}_{e_k}$. The iterative two-stage method is implemented for each alternative model as follows:

[27] 1. Obtain the maximum likelihood parameter estimates, $\hat{\boldsymbol{\beta}}_{k,\varepsilon}$, by using the inverse of $\mathbf{C}_\varepsilon$ for weighting;

[28] 2. **Stage 1:** Compute the residual time series $\{r_{k,t}\} = D_t - f_k(\hat{\boldsymbol{\beta}}_{k,\varepsilon})$, $t = 1, 2, \dots, N$, estimate the error

variance $\hat{\sigma}^2$ by calculating the sample variance, $\hat{\sigma}^2 =$

$$\frac{1}{N-1} \sum_{t=1}^N (r_{k,t} - \bar{r}_k)^2 \quad (\bar{r}_k = \sum_{t=1}^N r_{k,t} / N \text{ is the sample mean}),$$

analyze the residuals serial correlation to determine the AR(p) model to simulate $\{r_{k,t}\}$, estimate $\hat{\mathbf{a}}_k$, and construct $\mathbf{C}_{e_k}(\hat{\mathbf{a}}_k)$.

[29] 3. **Stage 2:** Update the maximum likelihood parameter estimates, $\hat{\boldsymbol{\beta}}_{e_k}$, using the inverse of $\mathbf{C}_{e_k}(\hat{\mathbf{a}}_k)$ as weighting.

[30] 4. Replace $\hat{\boldsymbol{\beta}}_{k,\varepsilon}$ by $\hat{\boldsymbol{\beta}}_{e_k}$ and repeat Stages 1 and 2 until convergence (e.g., changes of the parameter estimates between two iterations are smaller than a user-specified tolerance). The AR(p) model determined in Stage 1 may vary between the iterations.

[31] After completing the above procedure, the final results of residuals, parameter estimates, and covariance matrix $\mathbf{C}_{e_k}(\hat{\mathbf{a}}_k)$ are used to calculate *NLL* using equation (23), which is subsequently used for evaluating the model selection criteria and model averaging weights. While the method is developed for serial data with temporal correlation, it can be adapted for data with spatial correlation using geostatistical theories. When both spatial and temporal correlations exist, one may first characterize them separately and then aggregate them in the manner of Carrera and Neuman [1986] and Riva *et al.* [2011]. Different from the two-stage method of Seber and Wild [2003, p. 279] in statistics and Sadeghipour and Yeh [1984] in groundwater modeling, the iterative process does not need to invoke any assumption on the order of the AR(p) model. More importantly, it is the first time that the two-stage approach is used for evaluation of model uncertainty, i.e., the calculation of model averaging weights.

3.2. Needed Techniques of Time Series Analysis

[32] Implementing Stage 1 above requires a number of techniques of time series analysis. Only the techniques needed for the numerical studies in the next two sections are briefly described here to make this paper self contained. Sample autocorrelation function (*ACF*) is a widely used technique to examine the serial correlation. For a stationary sequence of residuals, $r_1, r_2, \dots, r_N$, with constant intervals, the sample *ACF*, λ_l at lag l , is defined as

$$\lambda_l = \frac{\sum_{t=1}^{N-l} (r_t - \bar{r})(r_{t+l} - \bar{r})}{\sum_{t=1}^N (r_t - \bar{r})^2} \quad \text{for } l = 1, 2, \dots, N-1, \quad (24)$$

where $\bar{r} = \sum_{t=1}^N r_t / N$ is the overall mean. Under the null hypothesis that the data are not autocorrelated (i.e., $\rho_l = 0$ for $l \neq 0$), the sample *ACF* λ_l is normally distributed with zero mean and standard deviation of $1/\sqrt{N}$, according to Cryer and Chan [2008]. If the calculated sample *ACFs* based on (24) are all within the range of $\pm 2/\sqrt{N}$ (i.e., its 95% confidence interval), the time series is generally considered to be uncorrelated. A plot of sample *ACFs* is always used to examine the serial correlation. A variant of sample *ACF* is the sample partial *ACF*, i.e., *PACF*, defined as the correlation between r_t and r_{t-l} after removing the effect of the intervening variables $r_{t-1}, r_{t-2}, r_{t-3}, \dots, r_{t-l+1}$, where l is the number of lags. The *PACF* can be used to determine the order, p , of an AR(p) model. For example, if the calculated *PACF* of a residual series is nonzero for lag 1 but zero for all lags greater than 1, then the correlation of the residuals can be determined by the AR(1) model. In practice, based on the theory that the sample *PACF* at lags greater than p is approximately normally distributed with zero means and variances $1/N$, if the

calculated sample *PACFs* are within the range of $\pm 2/\sqrt{N}$ (i.e., its 95% confidence interval), then the correlation of the series can be determined by the $AR(p)$ model [Cryer and Chan, 2008; Chatfield, 1989]. This can be done by the sample *PACFs* plot, as shown in the next two sections.

[33] After the order of $AR(p)$ is determined, the next step is to estimate the parameters, $\mathbf{a}$, of the model that is defined as

$$r_t = a_1 r_{t-1} + a_2 r_{t-2} + \dots + a_p r_{t-p} + \xi_t, \quad (25)$$

[34] The method of moments described in Cryer and Chan [2008] is used in this study. Take the $AR(2)$ model

$$r_t = a_1 r_{t-1} + a_2 r_{t-2} + \xi_t, \quad (26)$$

as an example. If $\{r_t\}$ is stationary, then according to Seber and Wild [2003, p. 287] the autocorrelation function, ρ , at lag l is

$$\rho_l = a_1 \rho_{l-1} + a_2 \rho_{l-2}. \quad (27)$$

[35] For $l = 1$,

$$\rho_1 = a_1 \rho_0 + a_2 \rho_{-1} = a_1 + a_2 \rho_1, \quad (28)$$

where $\rho_0 = 1$ and $\rho_{-1} = \rho_1$. For $l = 2$,

$$\rho_2 = a_1 \rho_1 + a_2. \quad (29)$$

[36] The method of moments replaces the theoretical *ACFs* ρ_l and ρ_2 by the sample *ACFs* λ_1 and λ_2 calculated based on equation (24) to obtain

$$\lambda_1 = a_1 + a_2 \lambda_1 \quad \text{and} \quad \lambda_2 = a_1 \lambda_1 + a_2, \quad (30)$$

[37] Solving equation (29) gives the estimates of the $AR(2)$ model parameters as follows,

$$\hat{a}_1 = \frac{\lambda_1(1 - \lambda_2)}{1 - \lambda_1^2} \quad \text{and} \quad \hat{a}_2 = \frac{\lambda_2 - \lambda_1^2}{1 - \lambda_1^2}. \quad (31)$$

[38] This method can be applied to any $AR(p)$ models and has been implemented as built-in functions of popular software such as R, SAS, and MATLAB. An ad hoc model diagnostics is needed to examine the goodness-of-fit of the $AR(p)$ to the residual series. If the model is correctly specified and the parameter estimates are accurate, then the remaining terms of the residuals after subtracting the fitted $AR(p)$ model should be uncorrelated, which can be investigated using the sample *ACFs* plot. Otherwise, the specified $AR(p)$ model does not adequately capture the correlation information in the residuals and a more appropriate model should be considered.

[39] With the estimated $AR(p)$ model, the correlation and covariance matrices of the residuals can be constructed using equations (21) and (20). Take again the $AR(2)$ model as an example. Equations (28) and (29) give $\rho_1 = \hat{a}_1/(1 - \hat{a}_2)$ and $\rho_2 = \hat{a}_1 \rho_1 + \hat{a}_2$ for $l = 1$ and $l = 2$, respectively. With ρ_1 and ρ_2 known, ρ_l can be estimated using equation (25). Subsequently, the correlation matrix, $\mathbf{V}$, defined in equation (21) can be specified. By virtue of equation (20), the covariance matrix, $\mathbf{C}_{e_k}$, can be constructed after estimating the var-

Table 1. Surface Complexation Reactions and Parameters of the True Model in the Synthetic Study^a

U(VI) Surface Reaction	Site	log K	Site Fraction (f)
$S_1OH + UO_2^{2+} + H_2O = S_1OUO_2OH + 2H^+$	Weak site	-4.9748	0.967979
$S_2OH + UO_2^{2+} + H_2O = S_2OUO_2OH + 2H^+$	Strong site	-3.4547	0.031819
$S_2OH + UO_2^{2+} = S_2OUO_2^+ + H^+$	Strong site	0.6113	
$S_3OH + UO_2^{2+} + H_2O = S_3OUO_2OH + 2H^+$	Stronger site	-1.1926	0.0002
$S_4OH + UO_2^{2+} = S_4OUO_2^+ + H^+$	Strongest site	2.8388	0.000002

^aTotal site density used in this model is 1.3E-03 M/L. Summation of site fraction is one.

iance term via $\hat{\sigma}^2 = \frac{1}{N-1} \sum_{t=1}^N (r_t - \bar{r})^2$, i.e., the sample variance of the residuals. While the above techniques are for a stationary time series, which is the case for the numerical studies below as verified in section 4.6, the iterative two-stage method is general and can be applied to both stationary and nonstationary time series.

4. Method Evaluation Using Synthetic Data

[40] In this section, the iterative two-stage parameter estimation method is evaluated using a synthetic study based on the laboratory experiments of Kohler *et al.* [1996]. In order to study uranium reactive transport, Kohler *et al.* [1996] conducted eight column experiments in a well-characterized U(VI)-quartz-fluoride column system. These experiments were conducted at pH values less than 5 in order to minimize complexation of uranium by carbonate but provided an excellent test of using the surface complexation modeling approach to simulate U(VI) transport with variable geochemical conditions. The breakthrough curves of U(VI) exiting the column over the course of several pore volumes of water showed retardation effect due to uranium adsorption on the quartz surface. The uranium adsorption was simulated in Kohler *et al.* [1996] using seven alternative surface complexation models (SCMs) (C1–C7) with different numbers of surface functional groups and different reaction stoichiometries. The models were calibrated against three column experiments (Experiments 1, 2, and 8) conducted under different experimental conditions, and the calibrated models were used to predict the remaining four experiments (Experiments 3, 4, 5, and 7) for cross-validation study. The synthetic study is conducted in a similar manner.

4.1. Synthetic Data and Models

[41] In the synthetic study, the true model is designed based on models C5 and C6 of Kohler *et al.* [1996]. As shown in Table 1, the true model has four functional groups: weak site (S_1OH), strong site (S_2OH), stronger site (S_3OH), and the strongest site (S_4OH). Each site is associated with one reaction, except that the strong site is associated with two reactions. Each reaction is associated with two parameters. One is the formation constant, K , that measures adsorption affinity of uranium on the individual function groups. Taking reaction $S_jOH + UO_2^{2+} + H_2O = S_jOUO_2OH + 2H^+$ as an example, its formation constant,

Table 2. Alternative Surface Complexation Models and Estimated Parameters in the Both Synthetic and Experimental Studies

Model	Reactions	Estimated Parameter
C3	$S_1OH + UO_2^{2+} + H_2O = S_1OUO_2OH + 2H^+$	logK1
	$S_2OH + UO_2^{2+} = S_2OUO_2^+ + H^+$	logK2
C4	$S_1OH + UO_2^{2+} + H_2O = S_1OUO_2OH + 2H^+$	logK1
	$S_2OH + UO_2^{2+} + H_2O = S_2OUO_2OH + 2H^+$	logK2
C5	$S_1OH + UO_2^{2+} + H_2O = S_1OUO_2OH + 2H^+$	logK1
	$S_2OH + UO_2^{2+} + H_2O = S_2OUO_2OH + 2H^+$	logK2
C6	$S_1OH + UO_2^{2+} + H_2O = S_1OUO_2OH + 2H^+$	logK1
	$S_2OH + UO_2^{2+} + H_2O = S_2OUO_2OH + 2H^+$	logK2
	$S_2OH + UO_2^{2+} = S_2OUO_2^+ + H^+$	logK3
	$S_3OH + UO_2^{2+} = S_3OUO_2^+ + H^+$	logK4

K , is defined as $K_j = (S_jOUO_2OH)(H^+)^2 / ((S_jOH)(UO_2^{2+}))$, where the quantities in parenthesis denote the activity of each species (the activity coefficients of the surface species are assumed to be equal to one). The other parameters are the fractions, f , of each functional group. The fractions of the functional groups sum up to one and the total site concentration was calculated from the measured specific surface area [Kohler *et al.*, 1996]. In reality, the parameters are in general unknown and need to be estimated by calibrating reactive transport models against species concentrations. Based on the true model and the true parameter values listed in Table 1, the computer code RATEQ [Curtis, 2005] was used to generate synthetic concentration data (RATEQ was also used for the forward model during the model calibration below). Three data sets were generated under the chemical conditions of Experiments 1, 2, and 8 of Kohler *et al.* [1996]; for each experiment, the time interval between two data was the same. This process yielded a total of $N = 120$ true values of uranium concentrations; they were corrupted with measurement errors that followed multivariate Gaussian distributions with zero means and diagonal covariance matrix, i.e., the variance matrix. The standard deviation of measurement errors was estimated from the real data of Kohler *et al.* [1996], and its order of magnitude was around 10^{-3} . When the corrupted data were negative (at the tails of the breakthrough curves), their absolute values were used. The 120 noisy data were used for the parameter estimation and multimodel analysis. UCODE_2005 [Poeter *et al.*, 2005] was used for the maximum likelihood model calibration. In this study, assuming that the experiments are independent, the covariance matrix is constructed for the individual experiments, and these covariance matrices are then assembled to form the final covariance matrix. It is worth pointing out that the three experimental data sets can be symbolized as different sets of observations, such as multiple kinds of observations from multiple experiments and/or from multiple observation wells. The iterative two-stage method described in section 3 is applicable in these situations.

[42] The four alternative models considered in this study were models C3–C6 of Kohler *et al.* [1996] listed in Table 2. Models C3–C5 are nested in that C4 has one more reaction than C3 and C5 has one more reaction than C4. Models

C5 and C6 have the same number of reactions, but the reactions associated with the stronger site (S_3OH) are different. Following Kohler *et al.* [1996], the calibrated parameters included the formation constant of each reaction and the fraction of the strong site (Table 2); the fraction of the stronger site was fixed at the value used by Kohler *et al.* [1996], and the fraction of the weak site was calculated as 1 minus the fractions of the strong and stronger sites. All four alternative models are simpler than the true model. Model C5 is the closest to the true model, as the first four reactions of the true model are identical to those of C5. Model C6 is the second closest to the true model, because of the incorrect reaction associated with the stronger site. In line with this, the order of plausibility of the four models, from the most to the least plausible, is C5, C6, C4, and C3, which is the basis to interpret statistical results below.

4.2. Temporal Residual Correlation of the True Model

[43] Using the diagonal covariance matrix of measurement errors, a conventional maximum likelihood model calibration was conducted for the true model to estimate the formation constants (logK) of the five reactions listed in Table 1 (the site fractions and densities were fixed at their true values). This is to investigate whether the model calibration process yields correlated residuals (even when the measurement errors are independent), as indicated in literature [Cook and Weisberg, 1982, p. 11; Cooley and Naff, 1990; Hill and Tiedeman, 2007, p. 111–113; Aster *et al.*, 2012]. The residuals corresponding to the three data sets are plotted in Figures 2(a1)–2(a3), which shows that the residuals are randomly distributed around the zero line. The sample ACF plots in Figures 2(b1)–2(b3) suggest that the residuals are serially uncorrelated, because the calculated sample ACFs are all within the 95% confidence interval. It indicates that residual correlation due to calibration process has negligible effect on the use of the residual plots to examine the residuals serial correlation in this study, because $(N - N_k)/N$ is close to 1.0 (where $N = 120$ is the number of data and $N_k = 5$ is the number of calibrated parameters) [Draper and Smith, 1981, p. 152]. Thus, for the true model, it is appropriate to use the diagonal covariance matrix of the measurement errors in model calibration. This is confirmed by the standard error, s , defined as [Hill and Tiedeman, 2007, p. 95]

$$s = \left(\frac{SSWR}{N - N_k} \right)^{1/2}. \quad (32)$$

[44] According to Hill and Tiedeman [2007, p. 96], if s is significantly different from a value of 1.0 and its $(1 - \alpha)\%$ confidence interval does not include the value of 1.0, the residuals are inconsistent with the weighting used for model calibration and evaluation of SSWR; the confidence interval is calculated as $\left(\sqrt{(N - N_k)s^2/\chi_U^2}, \sqrt{(N - N_k)s^2/\chi_L^2} \right)$, where χ_U^2 and χ_L^2 are, respectively, the upper-tail and lower-tail value of a chi-square with $N - N_k$ degrees of freedom and significance level of α [Ott, 1993, p. 332]. As shown in Table 3, for the true model, the 95% confidence interval of s includes 1.0, and the s value of 1.1 is close to 1.0. This, however, is not true for the alternative models, even for model C5 that is closest to the true model.

Figure 2. Plots of residuals of true model for (a1) Experiment 1, (a2) Experiment 2, and (a3) Experiment 8; (b1–b3) are sample *ACFs* plots of corresponding residuals.

4.3. Temporal Residual Correlation of Alternative Models

[45] Table 2 lists the calibrated parameters of the alternative models. Taking model C3 as an example, its calibrated parameters are the 10-base logarithms of the two formation constants ($\log K1$ and $\log K2$) and the fraction of the strong site ($\log Site$). The fraction of the weak site is not considered explicitly, because the summation of all the site fractions is one. For models C5 and C6, the fraction of the stronger site (S_3OH) is not influential to simulated concentrations and fixed at the true value of 0.0002 (Table 1). For the alternative models, two calibration cases are considered with different weight matrices when minimizing the objective function

$$(\mathbf{D} - f_k(\boldsymbol{\beta}_k))^T \mathbf{Q}_k (\mathbf{D} - f_k(\boldsymbol{\beta}_k)). \quad (33)$$

[46] Case I is the conventional maximum likelihood model calibration using $\mathbf{Q}_k = \mathbf{C}_e^{-1}$ as the weight matrix. Case II is the iterative two-stage model calibration using $\mathbf{Q}_k = \mathbf{C}_{e_k}^{-1}$ as the weight matrix.

[47] In Case I, the residuals of the four alternative models are temporally correlated. Taking model C5 (the best model) as an example, Figures 3(a1)–3(a3) plot the residuals of C5 with pore volumes (equivalent to time) for the three experimental data sets. The residuals are of the order

Figure 3. Plots of residuals (left) and residuals-AR(p) (right) of model C5 for (a1 and b1) Experiment 1, (a2 and b2) Experiment 2, and (a3 and b3) Experiment 8 in the synthetic study.

of 10^{-2} , about one order of magnitude larger than that of measurement error, suggesting existence of model errors. For the residuals series, positive departures from the zero line tend to be followed by positive departures, and so do the negative departures. The temporal correlation is quantitatively shown by the sample *ACF* plots in Figures 4(a1)–4(a3) in that the calculated *ACFs* at most lags are above the 95% confidence interval of the sample *ACFs*. Because the *ACFs* do not become zero after a certain number of lags, AR(p) models (rather than MA(q) models) are appropriate to describe the correlation structure of the residuals [Cryer and Chan, 2008, p. 123]. The sample *PACF* plots shown in Figures 4(b1) and 4(b2) indicate that AR(1) models are proper to simulate the residual correlations for Experiment 1 and 2, because the calculated *PACF* of the residual series is nonzero for lag 1 but zero for all lags greater than 1. Similarly, Figure 4(b3) suggests AR(2) model for Experiment 8. For the other three alternative models (results are not shown), the residuals are larger and the temporal correlation is stronger. For example, the residuals of model C3 (the worst model) are of the order of 10^{-1} ; AR(2) and AR(3) models are needed for all three experimental data sets. Comparing the residual analysis

Table 3. Standard Error, s , and Its 95% Confidence Intervals of the True Model and Four Alternative Models for Case I and Case II in the Synthetic Study

Models	True	C3	C4	C5	C6
Case I: Using the Covariance matrix, $\mathbf{C}_e$, of the Measurement Errors					
s	1.10	48.08	15.77	3.30	4.70
95% confidence interval of s	0.98–1.27	42.63–55.13	13.98–18.10	2.92–3.79	4.16–5.40
R_N^2 ^a	0.99 (0.97)	0.91 (0.97)	0.97 (0.97)	0.98 (0.97)	0.97 (0.97)
Case II: Using the Covariance Matrix, $\mathbf{C}_{e_k}$, of the Total Errors					
s		0.95	0.88	0.93	0.98
95% confidence interval of s		0.84–1.09	0.78–1.01	0.83–1.07	0.87–1.13
R_N^2 ^a		0.92 (0.97)	0.98 (0.97)	0.97 (0.97)	0.97 (0.97)

^a R_N^2 normality test for weighted residuals evaluated for the observations. Critical values are in parentheses. Larger R_N^2 values indicate normally distributed weighted residuals.

Figure 4. Sample *ACFs* plots of residuals of model C5 for (a1) Experiment 1, (a2) Experiment 2, and (a3) Experiment 8; (b1–b3) are sample *PACFs* plots of the corresponding residuals; (c1–c3) are sample *ACFs* plots of residuals- $AR(p)$ in the synthetic study. The blue lines represent the 95% confidence intervals of the sample *ACFs* and *PACFs*.

results of the alternative models with those of the true model leads to the conclusion that the residual temporal correlation detected in Figures 3 and 4 for the alternative models are mainly caused by model errors.

[48] Using the above $AR(p)$ models as the starting point, the iterative two-stage model calibration of Case II is completed in four iterations with maximum parameter difference of the last two iterations less than 1%. Based on the calibration results of the last iteration, the covariance matrix, C_{e_k} , is first constructed for each individual experimental data set based on equation (20); they are then used to assemble the covariance matrix for the entire N residuals with the assumption that the residuals of each individual experimental data set are uncorrelated. Based on the final results of the model calibration, Figures 4(c1)–4(c3) plot the sample *ACF* of the remaining terms of the residuals after subtracting the fitted $AR(p)$ models (i.e., Residual- $AR(p)$ as the y axis label). The ad hoc diagnostic analysis confirms that the $AR(p)$ models are adequate to describe the residuals correlation structures of the three experimental data sets, because the calculated *ACFs* are all within the 95% confidence interval. This is seen physically clearer in Figures 3(b1)–3(b3) for the plots of the remaining terms along the pore volumes. The change of residual correlation before and after implementing the iterative two-stage method is more apparent for the other three alternative models (results not shown).

[49] Figure 5 plots the calibrated breakthrough curves of the three experiments for the four models in Case I (a1–a3) and Case II (b1–b3). The figure indicates that, among the four models, C3 has the worst fit for all three experiments, especially for Experiment 8. Models C5 and C6 have the best fit and their calibrated breakthrough curves are visually identical. The fit is similar for the two cases. The root mean square errors (RMSEs) for models C3–C6 are 1.484, 0.399, 0.099, and 0.103, respectively, in Case I and become 1.479, 0.310, 0.080, and 0.090 in Case II. While the iterative two-stage model calibration method only marginally improves the model fit, using the full matrix to incorporate serial correlation in model calibration of Case II dramatically affects the evaluation of *NLL* and subsequently the model averaging weights, as shown below.

4.4. Estimation of *NLL* and Model Averaging Weights

[50] Table 4 lists the values of *SSWR*, $\ln|C|$, *NLL*, $\ln|F|$, ΔNLL , ΔIC , and model averaging weights w_{IC} for the four alternative models in the two calibration cases. The *SSWR* of Cases I and II are the Euclidian and Mahalanobis distances, respectively. The results of *AIC* are not shown, as *AIC* is less accurate than *AICc* for evaluating the model averaging weights [Poeter and Anderson, 2005]. Table 4 shows that, when C_e^{-1} is used as the weighting in Case I, the *SSWR* values are dramatically different between different models. For example, the *SSWR* values of models C5 and

Figure 5. Comparison of observed and simulated breakthrough curves for (a1 and b1) Experiment 1, (a2 and b2) Experiment 2, and (a3 and b3) Experiment 8 for calibration Case I (left) and Case II (right) in the synthetic study.

C6 are 393.8 and 800.6, respectively. This, however, is unreasonable, because the fit of models C5 and C6 shown in Figures 5(a1)–5(a3) is visually identical. The reason is that the use of C_e^{-1} for weighting mistakenly exaggerates

the residual differences by about 10^6 in the calculation of $SSWR$, considering that the standard deviation of measurement errors is about 10^{-3} in magnitude and that the weighting is taken as the inverse of the variance of measurement

Table 4. Values of $SSWR$, $\ln|C|$, NLL , $\ln|F|$, ΔNLL , ΔIC , and Model Averaging Weights of the Four Alternative Models Calculated for Case I and Case II using $AICc$, BIC , KIC Averaging in the Synthetic Study

	Case I Using C_e				Case II Using C_{e_i}			
	C3	C4	C5	C6	C3	C4	C5	C6
$SSWR$	89007.9	8767.3	393.8	800.6	104.9	89.5	100.3	111.3
$\ln C $	-1391.9	-1391.9	-1391.9	-1391.9	-942.1	-1079.9	-1220.4	-1230.9
NLL	87616.0	7375.4	-998.1	-591.3	-837.2	-990.3	-1120.1	-1119.6
$\ln F $	24.3	40.6	61.9	58.3	24.0	40.1	57.8	57.7
ΔNLL	88614.1	8373.5	0.0	406.8	282.9	129.7	0.0	0.5
$\Delta AICc$	88609.8	8371.4	0.0	406.8	278.6	127.5	0.0	0.5
ΔBIC	88604.5	8368.7	0.0	406.8	273.3	124.9	0.0	0.5
ΔKIC	88593.8	8365.5	0.0	406.7	253.3	113.6	0.0	1.0
$\alpha \Delta KIC^a$	8572.7	809.5	0.0	39.4	24.5	11.0	0.0	0.09
w_{AICc} (%)	0.0	0.0	100.0	0.0	0.0	0.0	56.2	43.8
w_{BIC} (%)	0.0	0.0	100.0	0.0	0.0	0.0	56.2	43.8
w_{KIC} (%)	0.0	0.0	100.0	0.0	0.0	0.0	61.9	38.1
$w_{\alpha KIC}$ (%) ^a	0.0	0.0	100.0	0.0	0.0	0.2	51.1	48.7

^aResults based on equation (8) with scaling factor $\alpha = 1.06/\sqrt{N}$ from Table 1 of [Tsai and Li, 2008a], where $N = 120$ is the number of observations.

Figure 6. Predictions (solid blue lines) and their 95% linear confidence intervals (dashed blue lines) of the alternative models for Experiment 3 in Case I of the synthetic study. The KIC -based model averaging results are plotted in red lines in (c); the red lines are overlapped on the blue lines because model C5 received 100% model averaging weight. Black dots represent the observed breakthrough curve of Experiment 3.

errors. The standard errors, s , and the 95% confidence intervals of the four alternative models listed in Table 3 also indicate the misuse of $\mathbf{C}_e^{-1}$ for calculating $SSWR$ of the alternative models. The s values of the four models are all significantly larger than 1.0, and none of their confidence intervals includes 1.0. As ΔNLL_k between model NLL_k and NLL_{min} is determined solely by $\Delta SSWR_k$ in Case I, the large difference in $SSWR$ between models C3–C6 causes large differences in NLL , and correspondingly large differences in model selection criteria. As a result, model C5 receives an unreasonable 100% model averaging weight for all the model selection criteria (Table 4).

[51] When $\mathbf{C}_e^{-1}$ is used as the weighting in Case II, the difference of $SSWR$ between the alternative models becomes significantly smaller. For example, the $SSWR$ values of models C5 and C6 in Case II are 100.3 and 111.3, respectively, and the difference is only 11.0. The covariance matrix, $\mathbf{C}_e$, with consideration of temporal correlation reasonably reflects the covariance structure of the total errors. This is confirmed by the standard errors, s , and their confidence intervals for Case II listed in Table 3. The s values of the alternative models are close to 1.0, and all the confidence intervals include 1.0. As noted in Hill and Tiedeman [2007, p. 95], when model error is included in the weighted residuals, the s values cannot be used as a measure of overall fit to the observations. The $\ln|\mathbf{C}_e|$ term contributes not only to the calculation of $SSWR$ and NLL but also to the evaluation of NLL and model selection criteria of the individual models, as explain in section 2. The $\ln|\mathbf{C}_e|$ values are different for different models. Table 4 shows that in Case II model C3 has the largest $\ln|\mathbf{C}_e|$ value, because the model has the largest model error; the magnitude of the residuals variance of C3 is the largest, around 10^{-2} (the corresponding values are 10^{-3} , 10^{-4} , and 10^{-4} for C4–C6, respectively). While models C5 and C6 have the same magnitude of the residuals variance, the tem-

poral correlation of C6 is relatively larger than that of C5. The use of $\mathbf{C}_e$ in model calibration and calculation of NLL leads to a small difference in NLL between models C3–C6 (especially between models C5 and C6) and more reasonable model averaging weights. The KIC -based weights of models C5 and C6 are 61.9 and 38.1%, respectively; those based on $AICc$ and BIC are 56.2 and 43.8%, respectively. This agrees with the analysis in section 4.1 that model C5 is the most plausible model followed by model C6.

[52] Equation (8) from Tsai and Li [2008a] was used in this study to examine whether it can resolve the problem of unreasonable 100% model averaging weight. Table 4 lists the scaled KIC ($\alpha\Delta KIC$) values and corresponding model averaging weights. Among the six scaling factors (α values) given in Table 1 of Tsai and Li [2008a], $\alpha = 1.06/\sqrt{N}$ ($N=120$ being the number of calibration data in this study) was used in this study, because it provides the largest effect. In Case I, while the $\alpha\Delta KIC$ values are one order of magnitude smaller than ΔKIC values, they still result in 100% model averaging weight for C5. This is also the case for $\alpha\Delta AICc$ and $\alpha\Delta BIC$ -based model averaging weights (results not shown). This suggests that using (8) cannot correct the problem caused by using $\mathbf{C}_e^{-1}$ in the calculation of NLL when the model errors are significant. In the case of two models and 120 observations, for one model to receive at least 5% weight using the model selection criteria, its ΔIC_k value cannot be larger than 60; otherwise, equation (8) still assigns 100% weight to the best model. In Case II, the application of equation (8) results in more evenly distributed model averaging weights between the two best models. It is thus concluded that equation (8) can only change the model averaging weights when ΔIC_k values are moderate or residual correlations are not as significant as in this example.

4.5. Assessment of Predictive Performance

[53] The calibrated models in Cases I and II were used to predict the concentration data generated by the true model

Figure 7. Predictions (solid blue lines) and their 95% linear confidence intervals (dashed blue lines) of the alternative models for Experiment 3 in Case II of the synthetic study. The KIC -based model averaging results are plotted in red lines in (c). Black dots represent the observed breakthrough curve of Experiment 3.

Table 5. Predictive Logscore of Individual Models and KIC -Based and αKIC -Based MLBMA in the Both Synthetic and Experimental Studies

	Model	C3	C4	C5	C6	MLBMA	
						KIC -Based	αKIC -Based
Synthetic Study	Case I	394.2	8.2	13.7	3.2	13.7	13.7 ^a
	Case II	284.6	6.0	9.0	2.9	-3.4	-3.4 ^a
Experimental Study	Case I	465.7	223.3	85.3	120.8	85.3	85.3 ^b
	Case II	166.9	166.2	75.1	90.2	70.5	64.7 ^b

^aResults based on model averaging weights of $w_{\alpha KIC}$ (%) listed in Table 4.

^bResults based on model averaging weights of $w_{\alpha KIC}$ (%) listed in Table 6.

under the chemical condition of Experiment 3 of *Kohler et al.* [1996]. The predictive uncertainty of individual models was measured by the 95% linear confidence interval defined as

$$\hat{y}_k \pm 2\sqrt{Var_{\hat{y}_k}}, \text{ and } Var_{\hat{y}_k} = s^2 \mathbf{Z}_k^T (\mathbf{X}_k^T \mathbf{Q}_k \mathbf{X}_k)^{-1} \mathbf{Z}_k, \quad (34)$$

where $\hat{y}_k$ is a prediction of model M_k , s^2 is the estimated error variance where s is defined in equation (31), $\mathbf{Z}_k$ is the sensitivity vector of the prediction with respect to model parameters evaluated at their optimal values, $\mathbf{X}_k$ is the corresponding sensitivity matrix of the observations, and $\mathbf{Q}_k$ is the weighting¹ used in equation (32). The $s^2 \mathbf{Z}_k^T (\mathbf{X}_k^T \mathbf{Q}_k \mathbf{X}_k)^{-1} \mathbf{Z}_k$ term calculates prediction variance, a propagation of parameter estimate covariance $s^2 (\mathbf{X}_k^T \mathbf{Q}_k \mathbf{X}_k)^{-1}$ which in turn is a propagation of error covariance $\mathbf{Q}_k^{-1}$, i.e., $\mathbf{C}_\varepsilon$ in Case I and $\mathbf{C}_{\varepsilon_k}$ in Case II. The 95% confidence interval of model averaging is

$$\hat{\bar{y}} \pm 2\sqrt{Var_{\hat{\bar{y}}}}, \text{ and } Var_{\hat{\bar{y}}} = \sum_{k=1}^K w_k Var_{\hat{y}_k} + \sum_{k=1}^K w_k (\hat{y}_k - \hat{\bar{y}})^2, \quad (35)$$

where $\hat{\bar{y}}$ is the weighted average prediction calculated in equation (1), and $Var_{\hat{\bar{y}}}$ is the model averaging variance. The evaluation of model averaging mean and variance assumes that the models are independent, and does not account for potential model correlation.

[54] The model predictions and the 95% linear confidence intervals based on individual models (blue lines) and KIC -based model averaging results (red lines) are shown in Figure 6 for Case I and in Figure 7 for Case II. In Figure 6, the confidence intervals of the individual models C3 and C4 (dashed blue lines) are wider than those of models C5 and C6. This is attributed to the values of s^2 (Table 3) used in equation (33) as a scaling factor. If the scaling factor s^2 is not considered, the confidence intervals of all the individual models are extremely small, because $\mathbf{C}_\varepsilon$ with order of 10^{-6} is used as the error covariance matrix $\mathbf{Q}_k^{-1}$ and the small variance of measurement errors results in small prediction variance of the individual models. The 95% confidence intervals of the individual models in Figure 7 for Case II are similar in magnitude to those in Figure 6 for Case I. This, however, is caused not by the s^2 values (because they are close to 1, as shown in Table 3), but by the covariance matrix in that the order of magnitude of $\mathbf{C}_{\varepsilon_k}$ in Case II is $10^{-2}, 10^{-3}, 10^{-4}$, and 10^{-4} for models C3–C6, respectively.

[55] In Figure 6, the confidence interval of model averaging (dashed red lines) is the same as that of model C5, because of the model's 100% model averaging weight. This small predictive uncertainty may lead to overconfidence in model predictions and incorrect decision making. The confidence interval of model averaging for Case II in Figure 7 is wider than that of models C5 and C6, especially at the beginning of the climbing limb and at the end of the descending limb of the breakthrough curve. Model averaging also reduces biasness in model predictions, considering that model C5 under predicts and model C6 overpredicts the concentrations in the climbing limb (the pattern of under and overprediction is the opposite at the descending limb). Therefore, the model averaging gives not only a relatively large prediction confidence interval but also a less biased prediction, resulting in more measurements (represented by the black dots in Figure 7) included in the confidence interval of model averaging than in that of individual models. Following *Ye et al.* [2004], predictive logscore [Good, 1952; Volinsky et al., 1997; Hoeting et al., 1999] was used to quantitatively assess the predictive performance of individual models and model averaging for the two cases. Predictive logscore considers the predictive bias and predictive uncertainty jointly; smaller predictive bias and larger predictive uncertainty probably lead to smaller predictive logscore, indicating better predictive performance [Shi et al., 2012]. The predictive logscore of an individual model M_k is defined as $\ln p(\hat{\mathbf{y}}_k | M_k, \mathbf{D}) = - \sum_{\hat{\mathbf{y}}_k \in \hat{\mathbf{y}}_k} \ln p(\hat{\mathbf{y}}_k | M_k, \mathbf{D})$, where $\mathbf{y}$ is the pre-

diction data (i.e., data of Experiment 3 in this study) for analysis of model predictive performance and $\mathbf{D}$ is the data of model calibration (i.e., Experiments 1, 2, and 8 in this

Figure 8. Plots of the residuals (left panel) and residuals-AR(p) (right panel) of model C5 for (a1 and b1) Experiment 1, (a2 and b2) Experiment 2, and (a3 and b3) Experiment 8 in the experimental study. The scale of y axis is 10 times larger than that of Figure 3, the counterpart of this figure.

Figure 9. Sample ACFs plots of the residuals of model C5 for (a1) Experiment 1, (a2) Experiment 2, and (a3) Experiment 8; (b1)–(b3) are sample PACFs plots of the corresponding residuals; (c1)–(c3) are sample ACFs plots of residuals-AR(p) in the experimental study. The blue lines represent the 95% confidence intervals of the sample ACFs and PACFs.

study). The predictive logscore of model averaging is defined as

$$-\ln p(\hat{\mathbf{y}}_k | \mathbf{D}) = -\sum_{\hat{\mathbf{y}}_k \in \hat{\mathbf{Y}}_k} \ln \left[\sum_{k=1}^K w_k p(\hat{\mathbf{y}}_k | M_k, \mathbf{D}) \right]. \quad (36)$$

[56] The lower the predictive logscore of model M_k or model averaging based on observation $\mathbf{D}$, the higher the probability that M_k based on $\mathbf{D}$ would predict $\mathbf{y}$. The predictive logscore of the individual models and KIC-based MLBMA are listed in Table 5 for the two cases. The table shows that the individual models give better predictive performance in Case II than in Case I; the model averaging in Case II has the best predictive performance, because it has the smallest logscore calculated using the KIC and α KIC-based weights. These results demonstrate the importance of considering temporal correlation in the error covariance matrix.

4.6. Evaluation of Assumptions

[57] Two assumptions are involved in the above analysis, i.e., Gaussian likelihood function and stationary time series. These assumptions are verified in an ad hoc manner using statistical techniques. Following Hill and Tiedeman [2007, p. 110], Gaussianity of the weighted residuals is examined using the statistical variable, R_N^2 , the correlation coefficient

between the weighted residuals (ordered from smallest to largest) and the normal order statistics [Brockwell and Davis, 1987, p. 304]. The weighted residuals are considered to be Gaussian, if values of R_N^2 are close to 1.0. Table 3 lists the R_N^2 values of the four alternative models and the critical value at significance level of 0.05. For models C4–C6, because the R_N^2 values are all larger than or equal to the critical value, it is concluded that the Gaussianity assumption is valid for the three models and that using a Gaussian likelihood function is appropriate. This is consistent with the finding of Hill *et al.* [1998] for a number of alternative flow models. Although the Gaussianity assumption is not valid for model C3, it is unlikely to influence the model averaging weights, since this model is inferior to model C4 that receives zero model averaging weight.

[58] The assumption of stationary time series was verified by examining the parameter coefficients of the AR(p) models. According to Cryer and Chan [2008, p. 71], an AR(1) process is stationary if its coefficient satisfies $|a| < 1$; for an AR(2) process with parameters a_1 and a_2 , the stationarity conditions are that $a_1 + a_2 < 1$, $a_2 - a_1 < 1$, and $|a_2| < 1$. In this study, all the calibrated AR(p) models satisfied the stationarity conditions. Taking model C5 as an example, the AR(1) model used to simulate residual correlations of Experiments 1 and 2 has estimated parameters of $a = 0.79$ and $a = 0.71$, respectively; the parameters of the AR(2) model used for Experiment 8 has estimated

Figure 10. Comparison of observed and simulated breakthrough curves for (a1 and b1) Experiment 1, (a2 and b2) Experiment 2, and (a3 and b3) Experiment 8 for calibration Case I (left) and Case II (right) in the experimental study.

Figure 11. Predictions (solid blue lines) and their 95% linear confidence intervals (dashed blue lines) of the alternative models for Experiment 4 in Case II of the experimental study. The KIC-based model averaging results are plotted in red lines in (c). Black dots represent the observed breakthrough curve of Experiment 4.

parameters of $a_1 = 0.50$ and $a_2 = -0.38$. These estimated parameters indicate that the $AR(p)$ models are stationary. This is also true for the other alternative models (results not shown).

5. Application to Experimental Data

[59] The experimental application is similar to that of the synthetic study, except that the true model is unknown. The real concentration observations of Experiment 1, 2, and 8 from Kohler *et al.* [1996] were used to calibrate the same alternative models, C3, C4, C5, and C6. The calibrated models were then used to predict Experiment 4 that has significantly different geochemical conditions from those of Experiments 1, 2, and 8. As in the synthetic study, two cases of model calibration were conducted, using C_e^{-1} and C_k^{-1} as the weighting in Case I and Case II, respectively. The analysis of temporal correlation in residuals and effects of the error covariance structure on calculation of model averaging weights and predictive performance was conducted similarly to the synthetic study.

[60] The residual plots in Figures 8(a1)–8(a3) and sample ACF plots in Figures 9(a1)–9(a3) for model C5 show that the residuals are temporally correlated in Case I. The

Table 6. Values of $SSWR$, $\ln|C|$, NLL , $\ln|F|$, ΔNLL , ΔIC , and Model Averaging Weights of Four Alternative Models Calculated for Case I and Case II Using AIC_c , BIC , KIC Averaging in the Experimental Study

	Case I Using C_e				Case II Using C_{e_k}			
	C3	C4	C5	C6	C3	C4	C5	C6
$SSWR$	94032.8	30267.3	8395.7	10973.0	82.9	68.4	90.0	81.1
$\ln C $	-1391.9	-1391.9	-1391.9	-1391.9	-851.1	-942.7	-1018.0	-1004.6
NLL	92640.9	28875.4	7003.8	9581.2	-768.1	-874.3	-928.0	-923.5
$\ln F $	27.3	38.8	47.7	47.0	24.5	36.3	47.6	46.1
ΔNLL	85637.1	21871.6	0.0	2577.3	159.9	53.7	0.0	4.5
ΔAIC_c	85632.7	21869.4	0.0	2577.3	155.6	51.5	0.0	4.5
ΔBIC	85627.5	21866.8	0.0	2577.3	150.3	48.9	0.0	4.5
ΔKIC	85616.3	21863.2	0.0	2577.2	143.5	45.8	0.0	3.3
$\alpha \Delta KIC^a$	8284.6	2115.6	0.0	249.4	13.9	4.4	0.0	0.3
w_{AIC_c} (%)	0.0	0.0	100.0	0.0	0.0	0.0	90.6	9.4
w_{BIC} (%)	0.0	0.0	100.0	0.0	0.0	0.0	90.6	9.4
w_{KIC} (%)	0.0	0.0	100.0	0.0	0.0	0.0	83.9	16.1
$w_{\alpha KIC}$ (%) ^a	0.0	0.0	100.0	0.0	0.1	5.6	50.9	43.4

^aResults based on equation (8) with scaling factor $\alpha = 1.06/\sqrt{N}$ from Table 1 of [Tsai and Li, 2008a], where $N=120$ is the number of observations.

sample $PACF$ plots in Figures 9(b1)–9(b3) indicate that AR(2), AR(2), and AR(1) models are proper to and can adequately simulate the temporal correlation in the three experiments as indicated by Figures 8(b1)–8(b3) and Figures 9(c1)–9(c3). Comparing Figure 8 with Figure 3 shows that the model error is larger in the experimental application than in the synthetic study. Figure 10 plots the calibrated breakthrough curves of the four models for Case I (a1–a3) and Case II (b1–b3). Comparing the two panels in Figure 10 shows that the model fit is improved slightly in Case II, (with exception of model C3 for Experiment 2) especially that of C3 for Experiment 8. The RMSE of C5 and C6 changes from 0.51 and 0.56 in Case I to 0.47 and 0.50 in Case II. The model fit of C4 also improves with RMSE reducing from 0.74 in Case I to 0.62 in Case II. However, in both cases, none of the models can simulate well the peak of the observed breakthrough curve of Experiment 1, suggesting that all the models have significant model error. Better models are needed to adequately simulate the data; exploration of this topic was beyond the scope of this study.

[61] Table 6 lists the values of $SSWR$, $\ln|C|$, NLL , $\ln|F|$, ΔIC and the corresponding model averaging weights w_{IC} for the four alternative models in the two calibration cases. As in the synthetic study, model C5 receives 100% model averaging weight in Case I due to the use of C_e . While this problem could not be resolved by using equation (8), it was resolved in Case II by using C_{e_k} because the KIC -based model averaging weights of models C5 and C6 are 83.9 and 16.1%, respectively. Applying equation (8) to the results of Case II further reduced the difference in model averaging weights between the two models.

[62] Figure 11 plots the 95% confidence intervals of the individual models (blue lines) and KIC -based model averaging (red lines) for Case II. As in the synthetic study, the 95% linear confidence intervals of the individual models are similar in magnitude to those of Cases I (results of Case I are not shown). As in Case II of the synthetic study, the confidence interval of model averaging is larger than that of the individual models due to the reasonable model averaging weights. Table 5 shows that the predictive logscore of model averaging is smaller than that of the individual

models, indicating better predictive performance of model averaging, especially when equation (8) was used to calculate the weights. However, different from the synthetic study, the predictive performance of model averaging is not significantly better than that of the best model C5 in the experimental study. Since model averaging is a weighted average of the predictions of individual models, predictive performance of model averaging depends on that of individual models.

[63] Figure 11 shows that, although the models can satisfactorily reproduce the calibration data (Figure 10), all of them underestimate the retardation effect when predicting Experiment 4; the simulated retardation factor of ~ 1.9 is about 20% smaller than the observed value of ~ 2.4 . There are two major differences between Experiment 4 of validation and Experiments 1, 2, and 8 of calibration. First, Experiment 4 was conducted at pH 4.39 that exceeded the maximum pH 4.26 of Experiment 2, indicating a slight extrapolation outside the calibration conditions. The other major difference is that 100 μM fluoride was added to Experiment 4, which produced UO_2F_2 and UO_2F^+ that together account for more than 80% of the uranium in solution. The dominant influence of fluoride on uranium transport is captured in the prediction of Experiment 4, because the retardation factor would have been greater than 10 with the high pH 4.39 and in the absence of fluoride. The underestimation retardation may be attributed to model structure error that the reaction models are inadequate to simulate the joint effect of increased pH and added fluoride in Experiment 4. This model error cannot be included in the covariance matrix, C_{e_k} , since it is unknown during model calibration until the models are used to simulate Experiment 4. The solution to this problem is to collect more data, especially those that can simultaneously reduce errors of multiple models [Neuman et al., 2012; Lu et al., 2012]. The underestimation may also be due to parametric uncertainty, because Kohler et al. [1996] showed that the problem of underestimation did not occur, when the same models were used to simulate Experiment 4 but using a different set of parameter values. The major difference is the thermodynamic data for the formation of UO_2F_2 and UO_2F^+ . The thermodynamic formation constants used in

this study are more current; they are revised upward in *Guillaumont et al.* [2003], and this increase makes adsorption less favorable which may explain why the simulations show less retardation than was observed in the previous simulations. This sensitivity underscores the importance of having reliable thermodynamic data when conducting reactive transport simulations. In order to fully understand the reasons that retardation was underestimated in Experiment 4, one should tackle parametric uncertainty before evaluating model structure error, which, however, is beyond the scope of this study.

6. Conclusions and Discussion

[64] This work investigates the effects of error covariance structure on evaluation of model averaging weights. It is demonstrated in a simple example that using the covariance matrix, C_e , of measurement errors, instead of the covariance matrix, C_{e_k} , of total errors (including model errors and measurement errors), in the calculation of the sum of squared weighted residuals (*SSWR*) distorts the evaluation of goodness-of-fit between alternative models, because the Mahalanobis distance corresponding to C_{e_k} becomes the normalized Euclidian distance corresponding to C_e . This further affects the calculation of the model selection criteria and results in inaccurate estimates of model averaging weights.

[65] To resolve this problem, an iterative two-stage parameter estimation method was developed. The key to this method is to iteratively infer the covariance matrix, C_{e_k} , of the unknown total errors from residuals during the model calibration. The inferred covariance matrix is then used in the evaluation of model selection criteria and model averaging weights. Although the method presented in this study is for serial data and based on time series techniques, it can be adapted to spatial data by using geostatistical techniques to characterize spatial correlation.

[66] The method was first evaluated using a synthetic study with the true model and four alternative models. For the true model, it was appropriate to use C_e and the correlation caused by model calibration was negligible. However, for the alternative models, the model errors were significantly larger than the measurement errors, and the total errors were temporally correlated due to the model errors. When C_e was used, the best model received 100% model averaging weight regardless of model selection criteria, although the second best model had almost identical goodness-of-fit and the same number of calibrated model parameters. This problem was resolved by using the iterative two-stage method, because using C_{e_k} gave more reasonable and realistic model averaging weights. This was supported by the calibration results and physical understanding of the alternative models. Using C_{e_k} also improved predictive performance of the individual models by giving wider confidence intervals. Due to the reasonable model averaging weights obtained using C_{e_k} , predictive performance of model averaging was also improved by yielding less biased results and wider confidence intervals than the individual models. It is interesting to note that the calibrated breakthrough curves obtained using C_{e_k} were almost identical to those obtained using C_e .

[67] The same conclusions were drawn from the application of the iterative two-stage method to the experimental

problem. However, the improvement of predictive performance was not as significant as that in the synthetic study, because of inherent structure inadequacy of the alternative models used for the experimental problem. Predictive performance of model averaging depends on that of individual models. In addition, fully exploring model structural uncertainty requires postulating alternative models that reflect different aspects of the system of interest. This is necessary to avoid the problem of model dependence discussed in *Bishop and Abramowitz* [2012]. It can also potentially improve predictive performance of model averaging. As pointed out by *Winter and Nychka* [2010], the results of model averaging are better than those of individual models, only when the individual models produce very different forecasts. At last, it should be pointed out that, in the context of Bayesian model selection and averaging, parameter prior distributions may have significant impacts on evaluation of model averaging weights. The impacts, however, cannot be investigated using the likelihood-based model selection criteria considered in this study. Instead, a full Bayesian analysis using Markov chain Monte Carlo techniques is necessary, which is warranted in a future study.

[68] **Acknowledgments.** This work was supported in part by NSF-EAR grant 0911074 and DOE-SBR grant DE-SC0002687. We thank Matthias Kohler for providing the experimental data and concentration error estimates for the laboratory column study. We also thank Claire Tiedeman and the anonymous reviewers for their comments.

References

- Ajami, N. K., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resour. Res.*, **43**, W01403, doi:10.1029/2005WR004745.
- Akaike, H. (1974), A new look at statistical model identification, *IEEE Trans. Autom. Control*, **AC-19**, 716–722.
- Aster, R. C., B. Borchers, and C. H. Thurber (2012), *Parameter Estimation and Inverse Problems*, 2nd ed., 360 pp., Elsevier, Amsterdam.
- Bates, J. M., and C. W. J. Granger (1969), The combination of forecasts, *Oper. Res. Q.*, **20**, 451–468.
- Beven, K. (2002), Towards a coherent philosophy for modeling the environment, *Proc. R. Soc. London Ser. A*, **458**(2026), 2465–2484.
- Beven, K. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, **320**, 18–36.
- Bishop, C. H., and G. Abramowitz (2012), Climate model dependence and the replicate earth paradigm, *Clim. Dyn.*, **41**, 885–900, doi:10.1007/s00382-012-1610-y.
- Bredhoeft, J. D. (2003), From models to performance assessment: The conceptualization problem, *Ground Water*, **41**(5), 571–577.
- Bredhoeft, J. D. (2005), The conceptualization model problem-surprise, *Hydrogeol. J.*, **13**, 37–46.
- Brockwell, P. J., and R. A. Davis (1977), *Time Series: Theory and Methods*, Springer-Verlag, New York.
- Burnham, K. P., and D. R. Anderson (2002), *Model Selection and Multiple Model Inference: A Practical Information-Theoretical Approach*, 2nd ed., Springer, New York.
- Carrera, J., and S. P. Neuman (1986), Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, *Water Resour. Res.*, **22**, 199–210.
- Chatfield, C. (1989), *The Analysis of Time Series*, 4th ed., Chapman and Hall, Boca Raton, Fla.
- Christensen, S., and J. Doherty (2008), Predictive error dependencies when using pilot points and singular value decomposition in groundwater model calibration, *Adv. Water Resour.*, **31**, 674–700.
- Cook, R. D., and S. Weisberg (1982), *Residuals and Influence in Regression*, 230 pp., Chapman and Hall, New York.
- Cooley, R. L., and S. Christensen (2006), Bias and uncertainty in regression-calibrated models of groundwater flow in heterogeneous media, *Adv. Water Resour.*, **29**, 639–656.

- Cooley, R. L., and R. L. Naff (1990), Regression modeling of ground-water flow, in *U.S. Geological Survey Techniques of Water-Resources Investigations, Book 3*, chap. B4, 232 pp., U.S. Geol. Surv., Washington, D. C.
- Cryer, J. D., and K. Chan (2008), *Time Series Analysis With Applications in R*, 2nd ed., Springer, N. Y.
- Curtis, G. P. (2005), *Documentation and Applications of the Reactive Geochemical Transport Model RATEQ*, U.S. Geol. Surv., Menlo Park, Calif.
- Dai, Z., A. Wolfsberg, P. Reimus, H. Deng, E. Kwicks, M. Ding, D. Ware, and M. Ye (2012), Identification of sorption processes and parameters for radionuclide transport in fractured rock, *J. Hydrol.*, **414**, 220–230, doi:10.1016/j.jhydrol.2011.10.035.
- Diks, C. G. H., and J. A. Vrugt (2010), Comparison of point forecast accuracy of model averaging methods in hydrologic applications, *Stoch. Environ. Res. Risk Assess.*, **24**(6), 809–820, doi:10.1007/s00477-010-0378-z.
- Doherty, J. (2005), *PEST: Model-Independent Parameter Estimation, User Manual*, 5th ed., Watermark Numerical Computing, Brisbane, Australia.
- Doherty, J., and D. Welter (2010), A short exploration of structural noise, *Water Resour. Res.*, **4**, W05525, doi:10.1029/2009WR008377.
- Draper, N. R., and H. Smith (1981), *Applied Regression Analysis*, 709 pp., John Wiley, N. Y.
- Guillaumont, R., T. Fanghänel, J. Fuger, I. Grenthe, V. Neck, D. A. Palmer, M. H. Rand. (2003), *Update on the Chemical Thermodynamics of Uranium, Neptunium, Plutonium, Americium and Technetium*, 970 p., Elsevier, Amsterdam.
- Finsterle S. (2007), *iTOUGH2 User's Guide*, Earth Sci. Div., Lawrence Berkeley Natl. Lab., Berkeley, Calif.
- Finsterle, S., and Y. Zhang (2011), Error handling strategies in multiphase inverse modeling, *Comput. Geosci.*, **37**, 724–730.
- Foglia, L., S. W. Mehl, M. C. Hill, P. Perona, and P. Burlando (2007), Testing alternative ground water models using cross validation and other methods, *Ground Water*, **45**(5), 627–641.
- Foglia, L., M. C. Hill, S. W. Mehl, and P. Burlando (2009), Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function, *Water Resour. Res.*, **45**, W06427, doi:10.1029/2008WR007255.
- Good, I. J. (1952), Rational decisions, *J. R. Stat. Soc., Ser. B*, **57**(1), 107–114.
- Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resour. Res.*, **48**, W08301, doi:10.1029/2011WR011044.
- Hansen, B. E. (2007), Least squares model averaging, *Econometrica*, **75**, 1175–1189.
- Hill, M. C., and C. R. Tiedeman (2007), *Effective Calibration of Ground Water Models, With Analysis of Data, Sensitivities, Predictions, and Uncertainty*, 480 pp., John Wiley, New York.
- Hill, M. C., R. L. Cooley, and D. W. Pollock (1998), A controlled experiment in ground water flow model calibration, *Ground Water*, **36**(3), 520–535, doi:10.1111/j.1745-6584.1998.tb02824.x.
- Hjort, N. L., and G. Claeskens (2003), Frequentist model average estimators, *J. Am. Stat. Assoc.*, **98**, 879–899.
- Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, **14**(4), 382–417.
- Hurvich, C. M., and C.-L. Tsai (1989), Regression and time series model selection in small sample, *Biometrika*, **76**(2), 99–104.
- Kashyap, R. L. (1982), Optimal choice of AR and MA parts in autoregressive moving average models, *IEEE Trans. Pattern Anal. Mach. Intel.*, **4**(2), 99–104.
- Kohler, M., G. P. Curtis, D. B. Kent, and J. A. Davis (1996), Experimental investigation and modeling of uranium(VI) transport under variable chemical conditions, *Water Resour. Res.*, **32**, 3539–3551.
- Kuczera, G. (1983), Improved parameter inference in catchment models, 1. Evaluating parameter uncertainty, *Water Resour. Res.*, **19**, 1151–1162, doi:10.1029/WR019i005p01151.
- Liu, Y., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, **43**, W07401, doi:10.1029/2006WR005756.
- Lu, D., M. Ye, and S. P. Neuman (2011), Dependence of Bayesian model selection criteria and Fisher information matrix on sample size, *Math. Geosci.*, **43**, 971–993, doi:10.1007/s11004-011-9359-0.
- Lu, D., M. Ye, S. P. Neuman, and L. Xue (2012), Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs, *Adv. Water Resour.*, **35**, 69–82, doi:10.1016/j.advwatres.2011.10.007.
- Mahalanobis, P. C. (1936), On the generalized distance in statistics, *Proc. Natl. Inst. Sci. India*, **2**(1), 49–55.
- Marshall, L., D. Nott, and A. Sharma (2005), Hydrological model selection: A Bayesian alternative, *Water Resour. Res.*, **41**, W10422, doi:10.1029/2004WR003719.
- Matott, L. S., J. E. Babendreier, and S. T. Purucker (2009), Evaluating uncertainty in integrated environmental models: A review of concepts and tools, *Water Resour. Res.*, **45**, W06421, doi:10.1029/2008WR007301.
- Meyer, P. D., M. Ye, M. L. Rockhold, S. P. Neuman, and K. J. Cantrell (2007), Combined Estimation of Hydrogeologic Conceptual Model, Parameter, and Scenario Uncertainty With Application to Uranium Transport at the Hanford Site 300 Area, NUREG/CR-6940, PNNL-16396, Pac. Northwest Natl. Lab., Richland, Wash.
- Morales-Casique, E., S. P. Neuman, and V. V. Vesselinov (2010), Maximum likelihood Bayesian averaging of airflow models in unsaturated fractured tuff using Occam and variance windows, *Stochastic Environ. Res. Risk Assess.*, **24**, 863–830.
- Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, *Stochastic Environ. Res. Risk Assess.*, **17**(5), 291–305, doi:10.1007/s00477-003-0151-7.
- Neuman, S. P., L. Xue, M. Ye, and D. Lu (2012), Bayesian analysis of data-worth considering model and parameter uncertainties, *Adv. Water Resour.*, **36**, 75–85, doi:10.1016/j.advwatres.2011.02.007.
- Nowak, W., Y. Rubin, and F. P. J. de Barros (2012), A hypothesis-driven approach to optimize field campaigns, *Water Resour. Res.*, **48**, W06509, doi:10.1029/2011WR011016.
- Ott, L. (1993), *An Introduction to Statistical Methods and Data Analysis*, 4th ed., Duxbury, Belmont, Calif.
- Parrish, M. A., H. Moradkhani, and C. M. DeChant (2012), Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation, *Water Resour. Res.*, **48**, W03519, doi:10.1029/2011WR011116.
- Poeter, E. P., and D. A. Anderson (2005), Multimodel ranking and inference in ground water modeling, *Ground Water*, **43**(4), 597–605.
- Poeter, E. P., and M. C. Hill (2007), MMA: A Computer Code for Multi-Model Analysis, U.S. Geol. Surv. Tech. Methods, TM6-E3, 113 pp.
- Poeter, E. P., M. C. Hill, E. R. Banta, S. W. Mehl, and S. Christensen (2005), UCODE_2005 and six other computer codes for universal sensitivity analysis, inverse modeling, and uncertainty evaluation, U.S. Geol. Surv. Tech. Methods, 6–A11, 283 pp.
- Pohlmann, K. F., M. Ye, and G. Pohl (2007), Use of numerical ground-water modeling to evaluate uncertainty in conceptual models of recharge and hydrostratigraphy, *IEEE Int. Symp. Technol. Soc.*, 165–169.
- Raftery, A. E., D. Madigan, and J. A. Hoeting (1997), Bayesian model averaging for linear regression models, *J. Am. Stat. Assoc.*, **92**, 179–191.
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, **133**, 1155–1174.
- Rings, J., J. A. Vrugt, G. Schoups, J. A. Huisman, and H. Vereecken (2012), Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments, *Water Resour. Res.*, **48**, W05520, doi:10.1029/2011WR011607.
- Riva, M., M. Panzeri, A. Guadagnini, and S. P. Neuman (2011), Role of model selection criteria in geostatistical inverse estimation of statistical data- and model-parameters, *Water Resour. Res.*, **47**, W07502, doi:10.1029/2011WR010480.
- Rojas, R., L. Feyen, and A. Dassargues (2008), Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resour. Res.*, **44**, W12418, doi:10.1029/2008WR006908.
- Rojas, R., L. Feyen, and A. Dassargues (2009), Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modeling, *Hydrol. Processes*, **23**(8), 1131–1146.
- Rubin, Y., X. Chen, H. Murakami, and M. Hahn (2010), A Bayesian approach for inverse modeling, data assimilation and conditional simulation of spatial random fields, *Water Resour. Res.*, **46**, W10523, doi:10.1029/2009WR008799.
- Sadeghipour, J., and W. W.-G. Yeh (1984), Parameter identification of groundwater aquifer models: A generalized least squares approach, *Water Resour. Res.*, **20**(7), 971–979, doi:10.1029/WR020i007p00971.
- Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, **46**, W10531, doi:10.1029/2009WR008933.

- Schwarz, G. (1978), Estimating the dimension of a model, *Annu. Stat.*, 6(2), 461–464.
- Seber, G. A. F., and C. J. Wild (2003), *Nonlinear Regression*, 768 pp., John Wiley, New York.
- Seifert, D., T. O. Sonnenborg, J. C. Refsgaard, A. L. Højberg, and L. Trolborg (2012), Assessment of hydrological model predictive ability given multiple conceptual geological models, *Water Resour. Res.*, 48, W06503, doi:10.1029/2011WR011149.
- Shi, X., M. Ye, S. Finsterle, and J. Wu (2012), Comparing nonlinear regression and Markov chain Monte Carlo methods for assessment of predictive uncertainty in vadose zone modeling, *Vadose Zone J.*, 11(4), doi:10.2136/vzj2011.0147.
- Singh, A., S. Mishra, and G. Ruskauff (2010), Model averaging techniques for quantifying conceptual model uncertainty, *Ground Water*, 48, 701–715, doi:10.1111/j.1745-6584.2009.00642.x.
- Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resour. Res.*, 16, 430–442.
- Tartakovsky, D. (2013), Assessment and management of risk in subsurface hydrology: A review and perspective, *Adv. Water Resour.*, 51, 247–260, doi:10.1016/j.advwatres.2012.04.007.
- Tiedeman, C. R., and C. T. Green (2011), Effect of correlated observation error on parameters, predictions, and uncertainty, *Water Resour. Res.*, doi:10.1002/wrcr.20499, in press.
- Tsai, F. T.-C., and X. Li (2008a), Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window, *Water Resour. Res.*, 44, W09434, doi:10.1029/2007WR006576.
- Tsai, F. T.-C., and X. Li (2008b), Multiple parameterization for hydraulic conductivity identification, *Ground Water*, 46(6), 851–864.
- Volinsky, C. T., D. Madigan, A. E. Raftery, and R. A. Kronmal (1997), Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke, *J. R. Stat. Soc., Ser. C*, 46, 433–448.
- Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, 43, W01411, doi:10.1029/2005WR004838.
- Winter, C. L., and D. Nychka (2010), Forecasting skill of model averaging, *Stochastic Environ. Res. Risk Assess.*, 24, 633–638, doi:10.1007/s00477-009-0350-y.
- Wohling, T., and J. A. Vrugt (2008), Combining multiobjective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models, *Water Resour. Res.*, 44, W12432, doi:10.1029/2008WR007154.
- Xu, T., A. J. Valocchi, J. Choi, and E. Amir (2012), Improving groundwater flow model prediction using complementary data-driven models, paper presented at XIX International Conference on Computational Methods in Water Resources, Univ. of Ill., Urbana-Champaign, Ill.
- Ye, M. (2010), MMA: A computer code for multimodel analysis, *Ground Water*, 48(1), 9–12, doi:10.1111/j.1745-6584.2009.00647.x.
- Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum Likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, 40, W05113, doi:10.1029/2003WR002557.
- Ye, M., S. P. Neuman, P. D. Meyer, and K. F. Pohlmann (2005), Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff, *Water Resour. Res.*, 41, W12429, doi:10.1029/2005WR004260.
- Ye, M., P. D. Meyer, and S. P. Neuman (2008a), On model selection criteria in multimodel analysis, *Water Resour. Res.*, 44, W03428, doi:10.1029/2008WR006803.
- Ye, M., K. F. Pohlmann, and J. B. Chapman (2008b), Expert elicitation of recharge model probabilities for the Death Valley regional flow system, *J. Hydrol.*, 354, 102–115, doi:10.1016/j.jhydrol.2008.03.001.
- Ye, M., P. D. Meyer, Y.-F. Lin, and S. P. Neuman (2010a), Quantification of model uncertainty in environmental modeling, *Stochastic Environ. Res. Risk Assess.*, 24, 807–808, doi:10.1007/s00477-010-0377-0.
- Ye, M., K. F. Pohlmann, J. B. Chapman, G. M. Pohl, and D. M. Reeves (2010b), A model-averaging method for assessing groundwater conceptual model uncertainty, *Ground Water*, 48, 716–728, doi:10.1111/j.1745-6584.2009.00633.x.
- Ye, M., D. Lu, S. P. Neuman, and P. D. Meyer (2010c), Comment on “Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window” by Frank T.-C. Tsai and Xiaobao Li, *Water Resour. Res.*, 46, W02801, doi:10.1029/2009WR008501.

Appendix D

This appendix contains the paper of Lu et al. (2012), which was published in Mathematical Geosciences. This research helps answer the questions below explained in detail in the report:

Question 5: How to reduce predictive uncertainty by collecting data of maximum value of information or data-worth?

Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs

Dan Lu^a, Ming Ye^{a,*}, Shlomo P. Neuman^b, Liang Xue^b

^a Department of Scientific Computing, Florida State University, Tallahassee, FL 32306, United States

^b Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721, United States

ARTICLE INFO

Article history:

Received 18 June 2011

Received in revised form 12 October 2011

Accepted 18 October 2011

Available online 3 November 2011

Keywords:

Data worth

Value of information

Data uncertainty

Model uncertainty

Parameter uncertainty

Uncertainty reduction

ABSTRACT

To manage water resource and environmental systems effectively requires suitable data. The worth of collecting such data depends on their potential benefit and cost, including the expected cost (risk) of failing to take an appropriate decision. Evaluating this risk calls for a probabilistic approach to data-worth assessment. Recently we [39] developed a multimodel approach to optimum value-of-information or data-worth analysis based on model averaging within a maximum likelihood Bayesian framework. Adopting a two-dimensional synthetic example, we implemented our approach using Monte Carlo (MC) simulations with and without lead order approximations, finding that the former approach was almost equally accurate but computationally more efficient. Here we apply our methodology to pneumatic permeability data from vertical and inclined boreholes drilled into unsaturated fractured tuff near Superior, Arizona. In an attempt to improve computational efficiency, we introduce three new approximations that require less computational effort and compare results with those obtained by the original Monte Carlo method. The first approximation disregards uncertainty in model parameter estimates, the second does so for estimates of potential new data, and the third disregards both uncertainties. We find that only the first approximation yields reliable quantitative assessments of reductions in predictive uncertainty brought about by the collection of new data. We conclude that, whereas parameter uncertainty may sometimes be disregarded for purposes of analyzing data worth, the same does not generally apply to uncertainty in estimates of potential new data.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Hydrogeologists face a daunting challenge to help insure that contaminants in the subsurface do not pose unacceptable future risks to humans and the environment [11,12]. To quantify and manage such risks, one must understand their relations to existing pollution and to possible remediation schemes. This in turn requires collecting suitable data (hydrogeological, the focus of this study as well as physiological and other), assessing their uncertainty, incorporating them in subsurface flow and contaminant transport models, and using them to monitor system response to existing and future scenarios. As noted by Back [3], three strategies have traditionally been used to optimize a data collection effort: minimize cost for a specific level of analytic accuracy or precision, minimize analytic uncertainty for a given budget, and/or respond to regulatory demands for a given amount and quality of data. Many today prefer a fourth approach based on value-of-information or data-worth analysis [1,4,9–12,14,16,18,22,24–26,30,33,35,37,40,43,48–50,53,55,56,58,63]. Here value of information is measured by its cost-effectiveness. A data collection program is cost-effective if its expected benefit exceeds its cost, including the expected cost (risk) of failing to take the right decision. A major benefit

of new data would be its potential to help improve one's understanding of the system, in large part through a reduction in model predictive uncertainty. This benefit would be worth the cost only if it had the potential to impact decisions concerning system management. Corresponding methods of analysis include Bayesian decision-making [24] based on a decision tree [3] and graphic risk-driven approaches [11,12].

In the past, analyses of data-worth have commonly relied on a single model with known parameters. Predictive uncertainty has been attributed at best to estimation uncertainty of model parameters. A more recent trend has been to consider uncertainties in both model parameters and structure. This has been motivated by a growing recognition that environmental systems are open and complex, rendering them prone to multiple conceptualizations and mathematical descriptions, including parameterizations. This is true regardless of the quantity and quality of available and potential data. In the context of multi-model analysis, methods that explore how different sets of conditioning data impact the predictive uncertainty of multiple models include the Generalized Likelihood Uncertainty Estimation method (GLUE [5,6,17]), Bayesian Model Averaging (BMA; [7,15,20,28,31,41,49]), and a combination of GLUE and BMA [47]. The method described in [49] aims to assess the worth of disparate data measured on different scales.

Recently we [39] developed a multimodel approach to optimum value-of-information or data-worth analyses within a Bayesian

* Corresponding author. Tel.: +1 850 567 4488; fax: +1 850 567 0098.

E-mail address: mye@fsu.edu (M. Ye).

Model Averaging (BMA) framework. We focused on a maximum likelihood variant of BMA (MLBMA [38,59]) that (a) is compatible with both deterministic and stochastic models, (b) admits (but does not require) prior information about the parameters, (c) is consistent with modern statistical methods of hydrologic model calibration, (d) allows (but does not require) approximating lead predictive moments of any model by linearization, (e) updates model posterior probabilities as well as parameter estimates on the basis of potential new data both before and after such data become actually available (during so-called pre-posterior and posterior stages of data-worth analysis, respectively), and (f) allows modifying the models and their numbers at both the pre-posterior and posterior stages. We expect the proposed approach to be of help in designing the collection of hydrologic characterization and monitoring data in a cost-effective manner by optimizing the difference between their benefit and cost. Whereas benefits would accrue from optimum gain in information or reduction in predictive uncertainty, costs would include the risk of failing to take an appropriate decision. Evaluating this risk calls for a probabilistic approach to data-worth assessment as we propose.

Implementation of our proposed approach on a synthetic geostatistical problem in two space dimensions demonstrated a need to account for the impact of potential new data on model and parameter uncertainties. Though neither existing nor a potentially augmented set of data were sufficient to identify correctly the underlying geostatistical model (variogram) and its parameters, they nevertheless yielded self-consistent results and allowed identifying quite accurately the impacts of potential new data on the spatial distribution and magnitude of corresponding reductions in predictive variance. Approximating lead predictive moments associated with each model by linearization yielded results comparable to those obtained via Monte Carlo (MC) simulation with a much less expenditure of computational effort. The extent to which such linearization would work in strongly nonlinear situations was deemed an open question.

The purpose of this study is to (1) apply our multimodel Bayesian methodology of data-worth assessment to real data and (2) explore the possibility to improve computational efficiency by minimizing the need for Monte Carlo simulations. We achieve the latter by introducing three approximations that require less computational effort than does our original approach [39]: one approximation that disregards uncertainty in model parameter estimates, another that does so for estimates of potential new data, and a third approximation that disregards both uncertainties. Our test case entails pneumatic permeability data from six vertical and inclined boreholes drilled into unsaturated fractured tuff near Superior, Arizona. We ask two questions: (1) Given data from three boreholes and funds to drill and test one additional borehole at one of two possible locations, at which location should one drill and test so as to reduce predictive uncertainty at yet another target borehole? (2) How does the answer depend on which among the above four MLBMA approaches we adopt? In addition to addressing these questions in the context of real data distributed in three-dimensional space our analysis sheds new light on roles played by parameter and pre-posterior data uncertainties in our approach.

2. Methodology

We start by reviewing key background material of BMA, MLBMA, and the Bayesian method of data-worth analysis in Sections 2.1 and 2.2 summarized from [39]. Section 2.3 presents three new approximation schemes developed here in an attempt, evaluated later, to reduce the computational cost of our original approach.

2.1. Background of BMA and MLBMA

Consider a random vector, Δ , the multivariate statistics of which are to be predicted with a set $\mathbf{M}$ of K mutually independent models (a somewhat ambiguous concept discussed in [60]), M_k , each characterized by a vector of parameters θ_k , conditional on a discrete set of data, $\mathbf{D}$ (the case of correlated models has recently been considered in [51]). The joint posterior (conditional) distribution of Δ is [15,20]

$$p(\Delta|\mathbf{D}) = E_{M_k|\mathbf{D}} p(\Delta|\mathbf{D}, M_k) = \sum_{k=1}^K p(\Delta|\mathbf{D}, M_k) P(M_k|\mathbf{D}), \quad (1)$$

where $p(\Delta|\mathbf{D}, M_k)$ is the joint posterior distribution of Δ due to model M_k and $P(M_k|\mathbf{D})$ the posterior probability of this model. The latter are given by Bayes' rule [15,20]

$$P(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k)P(M_k)}{\sum_{l=1}^K p(\mathbf{D}|M_l)P(M_l)}, \quad (2)$$

where

$$p(\mathbf{D}|M_k) = \int p(\mathbf{D}|M_k, \theta_k) p(\theta_k|M_k) d\theta_k \quad (3)$$

is the integrated likelihood of model M_k , $p(\mathbf{D}|M_k, \theta_k)$ being the joint likelihood of this model and its parameters, $p(\theta_k|M_k)$ the prior density of θ_k under model M_k , and $P(M_k)$ the prior probability of M_k . Whereas the likelihood $p(\mathbf{D}|M_k, \theta_k)$ contains a statistical model of errors associated with $\mathbf{D}$ (due to measurement, stochastic interpolation or both), the prior density $p(\theta_k|M_k)$ may contain a model of parameter measurement errors [8]. All probabilities are implicitly conditional on the choice of models entering into the set $\mathbf{M}$, which are taken to be mutually independent (the case of correlated models has recently been considered in [51]). The posterior mean and covariance of Δ are given by Draper [15] and Hoeting et al. [20]

$$E(\Delta|\mathbf{D}) = E_{M_k|\mathbf{D}} E(\Delta|\mathbf{D}, M_k) = \sum_{k=1}^K E(\Delta|\mathbf{D}, M_k) P(M_k|\mathbf{D}), \quad (4)$$

$$\begin{aligned} Cov(\Delta|\mathbf{D}) &= E_{M_k|\mathbf{D}} Cov(\Delta|\mathbf{D}, M_k) + Cov_{M_k|\mathbf{D}} E(\Delta|\mathbf{D}, M_k) \\ &= \sum_{k=1}^K Cov(\Delta|\mathbf{D}, M_k) P(M_k|\mathbf{D}) \\ &\quad + \sum_{k=1}^K [E(\Delta|\mathbf{D}, M_k) - E(\Delta|\mathbf{D})][E(\Delta|\mathbf{D}, M_k) - E(\Delta|\mathbf{D})]^T P(M_k|\mathbf{D}), \end{aligned} \quad (5)$$

where T denotes transpose, $E_{M_k|\mathbf{D}} Cov(\Delta|\mathbf{D}, M_k)$ is the within-model component of $Cov(\Delta|\mathbf{D})$ and $Cov_{M_k|\mathbf{D}} E(\Delta|\mathbf{D}, M_k)$ is its between-model component. A scalar measure of the posterior variance of Δ is given by the trace

$$Tr[Cov(\Delta|\mathbf{D})] = Tr[E_{M_k|\mathbf{D}} Cov(\Delta|\mathbf{D}, M_k)] + Tr[Cov_{M_k|\mathbf{D}} E(\Delta|\mathbf{D}, M_k)] \quad (6)$$

which is of interest because, for $K > 1$, one generally has $Tr[Cov_{M_k|\mathbf{D}} E(\Delta|\mathbf{D}, M_k)] > 0$ so that $Tr[Cov(\Delta|\mathbf{D})] > Tr[E_{M_k|\mathbf{D}} Cov(\Delta|\mathbf{D}, M_k)]$. Hence the consideration of multiple models generally results in greater predictive uncertainty, as measured by $Tr[Cov(\Delta|\mathbf{D})]$, than the uncertainty associated with a single model, as measured by $Tr[Cov(\Delta|\mathbf{D}, M_k)]$.

MLBMA [38,59] is an approximation of BMA obtained through replacement of θ_k by an estimate, $\hat{\theta}_k$, which maximizes the likelihood $p(\mathbf{D}|M_k, \theta_k)$. Obtaining such ML estimates entails calibrating each model against (conditioning on) the data $\mathbf{D}$ using well-established statistical inverse methods. In MLBMA, the integrated likelihood, $P(M_k|\mathbf{D})$, in (2) is replaced by $P(M_k|\mathbf{D})_{ML}$, where the subscript indicates approximation based on ML estimation of θ_k , using the Laplace method [27,28,38,45,54,59,61,62] via

$$P(M_k|\mathbf{D}) \simeq P(M_k|\mathbf{D})_{ML} = \frac{\exp\left(-\frac{1}{2}\delta KIC_k^D\right)P(M_k)}{\sum_{l=1}^K \exp\left(-\frac{1}{2}\delta KIC_l^D\right)P(M_l)}, \quad (7)$$

where

$$\delta KIC_k^D = KIC_k^D - KIC_{\min}^D, \quad (8)$$

$$KIC_k^D = -2 \ln p(\mathbf{D}|M_k, \hat{\theta}_k^D)_{ML} - 2 \ln p(\hat{\theta}_k^D|M_k)_{ML} + N_k \ln \left(\frac{N^D}{2\pi}\right) + \ln |\mathbf{F}_k(\mathbf{D}|M_k)_{ML}|, \quad (9)$$

KIC_k^D being the Kashyap model selection (or information) criterion for model M_k [27], $KIC_{\min}^D$ its minimum value over all candidate models, and $-2 \ln p(\mathbf{D}|M_k, \theta_k)_{ML} - 2 \ln p(\theta_k|M_k)_{ML}$ a negative log likelihood incorporating prior measurements of the parameters (if available), evaluated at $\hat{\theta}_k^D$. Here N_k is the dimension of θ_k (number of adjustable parameters associated with model M_k), N^D is the dimension of $\mathbf{D}$ (number of discrete data points, which may include measured parameter values), and $\mathbf{F}_k$ is the normalized (by N^D) observed (as opposed to ensemble mean) Fisher information matrix having components [27]

$$F_{k,nm} = -\frac{1}{N^D} \left[\frac{\partial^2 \ln p(\mathbf{D}|M_k, \theta_k)}{\partial \theta_{kn} \partial \theta_{km}} \right]_{\theta_k = \hat{\theta}_k^D}. \quad (10)$$

The observed Fisher information provides a more reliable assessment of KIC than does the more commonly used expected Fisher information [29,32]. The ML approximation in MLBMA extends further to approximating $p(\Delta|\mathbf{D}, M_k)$ in (1) by $p(\Delta|\mathbf{D}, M_k)_{ML}$ [38,57], the accuracy of which is explored numerically in [32,39]. In the limit of large N^D/N_k , KIC_k^D reduces asymptotically to the Bayesian selection (or information) criterion, BIC_k^D (e.g. [38,44]). However, a recent study [32] shows that KIC_k^D yields more accurate approximations of the integrated likelihood and posterior model probability than does BIC_k^D .

2.2. Multimodel Bayesian data augmentation

Suppose that the original data set $\mathbf{D}$ is augmented by another hypothetical (pre-posterior) data set, $\mathbf{C}$, which has not yet been collected and is therefore uncertain (random). Assume that the multivariate statistics of Δ , predicted with model set $\mathbf{M}$, can be conditioned on the augmented data set $\{\mathbf{D}, \mathbf{C}\}$. Assume further that the multivariate statistics of $\mathbf{C}$, conditional on $\mathbf{D}$, can be predicted either via BMA or via MLBMA with a set $\mathbf{P}$ of I mutually independent statistical models, P_i , having parameters π_i . The models P_i may be independent of M_k , may form extensions of M_k or may coincide with the latter as in the computational example given later in this paper. For example, whereas M_k may represent flow and/or transport models, P_i may be geostatistical models of spatial parameter variability; in our example below the two models coincide. Both sets of models, and their parameters, may change with the collection of new data. In some special cases, such as the Matérn variogram model [34], a range of models may be represented by a single model with one or more structural parameters. In the following models $\mathbf{P}$ are taken to be equivalent to models $\mathbf{M}$. Analogy to (1) and the law of total probability imply

$$p(\Delta|\mathbf{D}) = E_{\mathbf{C}|\mathbf{D}} p(\Delta|\mathbf{D}, \mathbf{C}) = \int p(\Delta|\mathbf{D}, \mathbf{C}) p(\mathbf{C}|\mathbf{D}) d\mathbf{C}, \quad (11)$$

where

$$p(\Delta|\mathbf{D}, \mathbf{C}) = \sum_{k=1}^K p(\Delta|\mathbf{D}, \mathbf{C}, M_k) P(M_k|\mathbf{D}, \mathbf{C}), \quad (12)$$

and $\mathbf{C}$ is implicitly conditional on the choice of models $\mathbf{P}$ and on $\mathbf{D}$. Analogy to (4) and by virtue of the law of total expectation lead to

$$E(\Delta|\mathbf{D}) = E_{\mathbf{C}|\mathbf{D}} E(\Delta|\mathbf{D}, \mathbf{C}) = \int E(\Delta|\mathbf{D}, \mathbf{C}) p(\mathbf{C}|\mathbf{D}) d\mathbf{C}, \quad (13)$$

where

$$E(\Delta|\mathbf{D}, \mathbf{C}) = \sum_{k=1}^K E(\Delta|\mathbf{D}, \mathbf{C}, M_k) P(M_k|\mathbf{D}, \mathbf{C}). \quad (14)$$

Analogy to (5) and by virtue of the law of total variance lead to

$$\text{Cov}(\Delta|\mathbf{D}) = E_{\mathbf{C}|\mathbf{D}} \text{Cov}(\Delta|\mathbf{D}, \mathbf{C}) + \text{Cov}_{\mathbf{C}|\mathbf{D}} E(\Delta|\mathbf{D}, \mathbf{C}), \quad (15)$$

where

$$\begin{aligned} \text{Cov}(\Delta|\mathbf{D}, \mathbf{C}) &= \sum_{k=1}^K \text{Cov}(\Delta|\mathbf{D}, \mathbf{C}, M_k) P(M_k|\mathbf{D}, \mathbf{C}) + \sum_{k=1}^K [E(\Delta|\mathbf{D}, \mathbf{C}, M_k) \\ &\quad - E(\Delta|\mathbf{D}, \mathbf{C})][E(\Delta|\mathbf{D}, \mathbf{C}, M_k) - E(\Delta|\mathbf{D}, \mathbf{C})]^T P(M_k|\mathbf{D}, \mathbf{C}). \end{aligned} \quad (16)$$

Analogy to (6) gives

$$\text{Tr}[\text{Cov}(\Delta|\mathbf{D})] = \text{Tr}[E_{\mathbf{C}|\mathbf{D}} \text{Cov}(\Delta|\mathbf{D}, \mathbf{C})] + \text{Tr}[\text{Cov}_{\mathbf{C}|\mathbf{D}} E(\Delta|\mathbf{D}, \mathbf{C})]. \quad (17)$$

The term $\text{Tr}[E_{\mathbf{C}|\mathbf{D}} \text{Cov}(\Delta|\mathbf{D}, \mathbf{C})] = \text{Tr}[\text{Cov}(\Delta|\mathbf{D})] - \text{Tr}[E_{\mathbf{C}|\mathbf{D}} \text{Cov}(\Delta|\mathbf{D}, \mathbf{C})]$ represents the difference between the total trace conditional on $\mathbf{D}$ and the expected trace conditional jointly on $\mathbf{D}$ and $\mathbf{C}$. As this difference is positive, conditioning on $\mathbf{D}$ and $\mathbf{C}$ jointly results in a lower trace than conditioning on $\mathbf{D}$ alone. The difference could be viewed as an extended version of the A-criterion in optimal design, other measures of uncertainty reduction being possible. Appendix A reproduces the MLBMA procedure developed in [39] for prior conditioning on $\mathbf{D}$ and preposterior conditioning on both $\mathbf{D}$ and $\mathbf{C}$. The procedure accounts for data uncertainty through MC simulations in Step 8 and for parameter uncertainty through MC simulations in Step 8d.

2.3. Proposed new approximations

Implementing the above approach entails MC simulations with respect to both parameter and data uncertainties, rendering it computationally demanding. Here we complement lead-order approximations developed in [39] with three new approximations that reduce computational effort further. We start by rewriting (15) as

$$\begin{aligned} \text{Cov}(\Delta|\mathbf{D}) &= E_{\mathbf{C}|\mathbf{D}} E_{M_k|\mathbf{D}, \mathbf{C}} \text{Cov}(\Delta|M_k, \mathbf{D}, \mathbf{C}) \\ &\quad + E_{\mathbf{C}|\mathbf{D}} \text{Cov}_{M_k|\mathbf{D}, \mathbf{C}} E(\Delta|M_k, \mathbf{D}, \mathbf{C}) \\ &\quad + \text{Cov}_{\mathbf{C}|\mathbf{D}} E_{M_k|\mathbf{D}, \mathbf{C}} E(\Delta|M_k, \mathbf{D}, \mathbf{C}), \end{aligned} \quad (18)$$

where we applied the law of total covariance to $E_{\mathbf{C}|\mathbf{D}} \text{Cov}(\Delta|\mathbf{D}, \mathbf{C})$ and the law of total expectation to $\text{Cov}_{\mathbf{C}|\mathbf{D}} E(\Delta|\mathbf{D}, \mathbf{C})$ in a way that accounts explicitly for model uncertainty. Next we rewrite (18) as

$$\begin{aligned} \text{Cov}(\Delta|\mathbf{D}) &= E_{\mathbf{C}|\mathbf{D}} E_{M_k|\mathbf{D}, \mathbf{C}} E_{\theta_k|M_k, \mathbf{D}, \mathbf{C}} \text{Cov}(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C}) \\ &\quad + E_{\mathbf{C}|\mathbf{D}} E_{M_k|\mathbf{D}, \mathbf{C}} \text{Cov}_{\theta_k|M_k, \mathbf{D}, \mathbf{C}} E(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C}) \\ &\quad + E_{\mathbf{C}|\mathbf{D}} \text{Cov}_{M_k|\mathbf{D}, \mathbf{C}} E_{\theta_k|M_k, \mathbf{D}, \mathbf{C}} E(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C}) \\ &\quad + \text{Cov}_{\mathbf{C}|\mathbf{D}} E_{M_k|\mathbf{D}, \mathbf{C}} E_{\theta_k|M_k, \mathbf{D}, \mathbf{C}} E(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C}), \end{aligned} \quad (19)$$

where we applied the law of total expectation to $E(\Delta|M_k, \mathbf{D}, \mathbf{C})$ and the law of total covariance to $\text{Cov}(\Delta|M_k, \mathbf{D}, \mathbf{C})$ in a way that accounts explicitly for parameter uncertainty. The covariance decomposition indicates that uncertainties in data ($\mathbf{C}$), model (M), parameter (θ), and prediction sample (Δ) are quantified in (19). Here $\text{Cov}(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C})$, $\text{Cov}_{\theta_k|M_k, \mathbf{D}, \mathbf{C}} E(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C})$ and $\text{Cov}_{M_k|\mathbf{D}, \mathbf{C}} E_{\theta_k|M_k, \mathbf{D}, \mathbf{C}} E(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C})$ represent sampling, parameter and model uncertainties, respectively [45] whereas the last term in (19) represents uncertainty about the pre-posterior data $\mathbf{C}$.

Fig. 1. Spatial locations of 184 1-m-scale $\log_{10}k$ data along six boreholes at ALRS.

The first approximation eliminates parameter uncertainty from (19). The first term on the right hand side of (19) contains

Table 1

AIC- and KIC-based posterior probabilities (%) based on $\mathbf{D}$, $\{\mathbf{D}, \mathbf{C}_1'\}$ and $\{\mathbf{D}, \mathbf{C}_2'\}$. AICc- and BIC-based probabilities are similar to those of AIC-based.

Criterion	AIC			KIC		
	Pow0	Exp0	Sph0	Pow0	Exp0	Sph0
Case (%)	CV I					
$p(\mathbf{M}_k \mathbf{D})$	42.02	56.21	1.77	0.82	96.51	2.67
$p(\mathbf{M}_k \mathbf{D}, \mathbf{C}_1')$	65.80	34.02	0.18	2.37	97.16	0.47
$p(\mathbf{M}_k \mathbf{D}, \mathbf{C}_2')$	62.96	35.84	1.21	1.88	95.66	2.46
Case	CV II					
$p(\mathbf{M}_k \mathbf{D})$	75.82	13.09	11.09	4.26	56.25	39.49
$p(\mathbf{M}_k \mathbf{D}, \mathbf{C}_1')$	91.70	5.76	2.54	11.65	68.53	19.82
$p(\mathbf{M}_k \mathbf{D}, \mathbf{C}_2')$	34.03	65.21	0.77	0.73	98.55	0.72

$E_{\theta_k|\mathbf{M}_k, \mathbf{D}, \mathbf{C}} \text{Cov}(\Delta|\theta_k, \mathbf{M}_k, \mathbf{D}, \mathbf{C})$ which, by virtue of the mean integral value theorem, can be expressed as

$$\begin{aligned}
 E_{\theta_k|\mathbf{M}_k, \mathbf{D}, \mathbf{C}} \text{Cov}(\Delta|\theta_k, \mathbf{M}_k, \mathbf{D}, \mathbf{C}) \\
 &= \int \text{Cov}(\Delta|\theta_k, \mathbf{M}_k, \mathbf{D}, \mathbf{C}) p(\theta_k|\mathbf{M}_k, \mathbf{D}, \mathbf{C}) d\theta_k = \text{Cov}(\Delta|\xi_k, \mathbf{M}_k, \mathbf{D}, \mathbf{C}) \\
 &\times \int p(\theta_k|\mathbf{M}_k, \mathbf{D}, \mathbf{C}) d\theta_k = \text{Cov}(\Delta|\xi_k, \mathbf{M}_k, \mathbf{D}, \mathbf{C}), \quad (20)
 \end{aligned}$$

where ξ_k is an unknown value of θ_k . Replacing ξ_k with $\hat{\theta}_k^{D, C}$, the ML estimate of θ_k based jointly on $\{\mathbf{D}, \mathbf{C}\}$, gives $E_{\theta_k|\mathbf{M}_k, \mathbf{D}, \mathbf{C}} \text{Cov}$

Fig. 2. Sample and fitted variograms with number of data pairs per lag based on (a–b) $\mathbf{D}$, (c–d) $\mathbf{D} + \mathbf{C}_1'$, and (e–f) $\mathbf{D} + \mathbf{C}_2'$.

Fig. 3. Measured $\log k$ (+), predicted (solid) and 95% confidence intervals (dashed) for (a) Pow0, (b) Exp0, (c) Sph0 and (d) MLBMA along borehole V2 based on data D (black), $D + C_1$ (blue), and $D + C_2$ (green) in CV I. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 4. Measured $\log k$ (+), predicted (solid) and 95% confidence intervals (dashed) with (a) Pow0, (b) Exp0, (c) Sph0 and (d) MLBMA along borehole Y3 based on data D (black) and $D + C_1$ (blue) and $D + C_2$ (green) in CV II. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 5. Variation of $\text{Var}(\Delta|\mathbf{D})_{ML^D}$, $\text{Var}(\Delta|\mathbf{D}, \mathbf{C}_1')_{ML^{D,C_1}}$ and $\text{Var}(\Delta|\mathbf{D})_{ML^D} - \text{Var}(\Delta|\mathbf{D}, \mathbf{C}_1')_{ML^{D,C_1}}$ along predicted borehole V2 based on (a) $\mathbf{D} + \mathbf{C}_1$ and (b) $\mathbf{D} + \mathbf{C}_2$ in CV I.

Fig. 6. Variation of $\text{Var}(\Delta|\mathbf{D})_{ML^D}$, $\text{Var}(\Delta|\mathbf{D}, \mathbf{C}_1')_{ML^{D,C_1}}$ and $\text{Var}(\Delta|\mathbf{D})_{ML^D} - \text{Var}(\Delta|\mathbf{D}, \mathbf{C}_1')_{ML^{D,C_1}}$ along predicted borehole Y3 based on (a) $\mathbf{D} + \mathbf{C}_1$ and (b) $\mathbf{D} + \mathbf{C}_2$ in CV II.

$(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C}) \approx \text{Cov}(\Delta|\hat{\theta}_k^{D,C}, M_k, \mathbf{D}, \mathbf{C})$. This is equivalent to a first-order moment approximation; the second-order moment approximation of $E_{\theta_k|M_k, \mathbf{D}, \mathbf{C}} \text{Cov}(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C})$ is given in equation (B5) of [39]. Replacing θ_k with $\hat{\theta}_k^{D,C}$ in the second term of (19) eliminates parameter uncertainty by rendering this term equal to zero. Hence in our first approximation of (19) takes the partially linearized form

$$\begin{aligned} \text{Cov}(\Delta|\mathbf{D}) &\approx E_{\mathbf{C}|\mathbf{D}} E_{M_k|\mathbf{D}, \mathbf{C}} \text{Cov}(\Delta|\hat{\theta}_k^{D,C}, M_k, \mathbf{D}, \mathbf{C}) \\ &\quad + E_{\mathbf{C}|\mathbf{D}} \text{Cov}_{M_k|\mathbf{D}, \mathbf{C}} E(\Delta|\hat{\theta}_k^{D,C}, M_k, \mathbf{D}, \mathbf{C}) \\ &\quad + \text{Cov}_{\mathbf{C}|\mathbf{D}} E_{M_k|\mathbf{D}, \mathbf{C}} E(\Delta|\hat{\theta}_k^{D,C}, M_k, \mathbf{D}, \mathbf{C}) \end{aligned} \quad (21)$$

which eliminates the need for MC simulations in Step 8d of [Appendix A](#). By disregarding parameter uncertainty the approximation introduces an error of order $O(1/N)$ where N is the dimension of $\{\mathbf{D}, \mathbf{C}\}$, which may be acceptable when model uncertainty exceeds parameter uncertainty [45]. The corresponding equivalent of $\text{Tr}[\text{Cov}_{\mathbf{C}|\mathbf{D}} E(\Delta|\mathbf{D}, \mathbf{C})]$ in (17) is $\text{Tr}[\text{Cov}_{\mathbf{C}|\mathbf{D}} E(\Delta|\hat{\theta}_k^{D,C}, \mathbf{D}, \mathbf{C})]$ denoted by $\text{Tr}[\text{Cov}_{\mathbf{C}|\mathbf{D}} E(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}]$ in [Appendix A](#).

The second approximation removes the need for MC simulations over random realizations of the potential new data $\mathbf{C}$ in step 8 of our procedure ([Appendix A](#)) through their replacement by the corresponding conditional mean, $E(\mathbf{C}|\mathbf{D})$. This eliminates the second term and the mean operation in the first term at the right hand side of (15). Here the predictive uncertainty reduction cannot be defined as $\text{Tr}[\text{Cov}_{\mathbf{C}|\mathbf{D}} E(\Delta|\mathbf{D}, \mathbf{C})]$ because this term now vanishes. Instead, the reduction is defined as the difference between $\text{Tr}[\text{Cov}(\Delta|\mathbf{D})]$ and $\text{Tr}[\text{Cov}(\Delta|\mathbf{D}, E(\mathbf{C}|\mathbf{D}))]$. Following the derivation of (19) and using the law of total variance with respect to model uncertainty gives

$$\begin{aligned} \text{Cov}(\Delta|\mathbf{D}, E(\mathbf{C}|\mathbf{D})) &= E_{M_k|\mathbf{D}, E(\mathbf{C}|\mathbf{D})} \text{Cov}[\Delta|M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})] \\ &\quad + \text{Cov}_{M_k|\mathbf{D}, E(\mathbf{C}|\mathbf{D})} E[\Delta|M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})]. \end{aligned} \quad (22)$$

Applying the law of total variance to $\text{Cov}[\Delta|M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})]$ and the law of total expectation to $E[\Delta|M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})]$ in a way that accounts explicitly for parameter uncertainty yields

$$\begin{aligned} \text{Cov}(\Delta|\mathbf{D}, E(\mathbf{C}|\mathbf{D})) &= E_{M_k|\mathbf{D}, E(\mathbf{C}|\mathbf{D})} E_{\theta_k|M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})} \text{Cov}[\Delta|\theta_k, M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})] \\ &\quad + E_{M_k|\mathbf{D}, E(\mathbf{C}|\mathbf{D})} \text{Cov}_{\theta_k|M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})} E[\Delta|\theta_k, M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})] \\ &\quad + \text{Cov}_{M_k|\mathbf{D}, E(\mathbf{C}|\mathbf{D})} E_{\theta_k|M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})} E[\Delta|\theta_k, M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})], \end{aligned} \quad (23)$$

which is the predictive uncertainty given $\mathbf{D}$ and $E(\mathbf{C}|\mathbf{D})$.

The third approximation combines the first two approximations by disregarding both parameter and data uncertainties. Replacing θ_k in (23) by its ML estimate $\hat{\theta}_k$ obtained on the basis of $\mathbf{D}$ and $E(\mathbf{C}|\mathbf{D})$, thereby ignoring parameter uncertainty (and thus eliminating the need for MC simulations across random parameter realizations in step 8d of [Appendix A](#)), yields a third approximation of (19) which disregard both data and parameter uncertainties,

$$\begin{aligned} \text{Cov}(\Delta|\mathbf{D}) &\approx \text{Cov}(\Delta|\mathbf{D}, E(\mathbf{C}|\mathbf{D})) \\ &\approx E_{M_k|\mathbf{D}, E(\mathbf{C}|\mathbf{D})} \text{Cov}[\Delta|\hat{\theta}_k, M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})] \\ &\quad + \text{Cov}_{M_k|\mathbf{D}, E(\mathbf{C}|\mathbf{D})} E[\Delta|\hat{\theta}_k, M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})]. \end{aligned} \quad (24)$$

If $E(\mathbf{C}|\mathbf{D}) E[\Delta|\hat{\theta}_k, M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})]$ and $\text{Cov}[\Delta|\hat{\theta}_k, M_k, \mathbf{D}, E(\mathbf{C}|\mathbf{D})]$ can be estimated directly through geostatistical (as we do below) or stochastic modeling, step 8 in [Appendix A](#), which is computationally the most demanding, becomes redundant. The accuracy of approximations (23) and (24) diminishes with increasing variability of $\mathbf{C}$ and sensitivities of θ_k and Δ to $\mathbf{C}$. As in the second approximation, uncertainty reduction due to $E(\mathbf{C}|\mathbf{D})$ in the third approximation is measured by $\text{Tr}[\text{Cov}(\Delta|\mathbf{D})] - \text{Tr}[\text{Cov}(\Delta|\mathbf{D}, E(\mathbf{C}|\mathbf{D}))]$.

3. Application to air permeabilities of fractured tuff

We implement our procedure ([Appendix A](#)) and its three approximations, based on (21)–(24), on 1-m scale log air permeability ($\log k$) data from unsaturated fractured tuff at the former Apache Leap Research Site (ALRS) near Superior, Arizona. Spatially distributed $\log k$ data were obtained [19] based on a steady state interpretation of 184 pneumatic injection tests in 1-m-length intervals along six boreholes at the site ([Fig. 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. We consider two “cross-validation” cases: CV I, in which $\log k$ measured

Table 2

Actual and predicted trace variance reductions in CV I and CV II.

Case	CV I		CV II	
	C_1 (X2)	C_2 (Y2)	C_1 (V2)	C_2 (Z2)
Actual data (C')	2.31	1.01	3.11	2.84
Consider both parameter and data uncertainties				
Data estimates (C)	0.59	0.42	1.79	0.84
Approximation (21) disregarding parameter uncertainty (Eq. (21))				
Data estimates (C)	0.69	0.53	1.94	0.84
Approximation (22) disregarding data uncertainty				
Data estimates (C)	3.66	3.64	10.15	8.96
Approximation (23) disregarding both parameter and data uncertainties				
Data estimates (C)	4.10	4.05	6.25	5.94

Fig. 7. C' (red), $E(C|D)_{MLD}$ (solid blue), 200 realizations of predicted C (gray) and 95% confidence interval of all 200 realizations (dashed blue) for (a) C_1 along borehole X2 and (b) C_2 along borehole Y2 in CV I. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 9. Variation of $Var(D|D)_{MLD,C}$, $E_{C|D}Var(D|D,C)_{MLD,C}$ and $Var_{C|D}E(D|D,C)_{MLD,C}$ along borehole Y3 based on (a) $D + C_1$ and (b) $D + C_2$ in CV II.

Fig. 8. Variation of $Var(D|D)_{MLD,C}$, $E_{C|D}Var(D|D,C)_{MLD,C}$ and $Var_{C|D}E(D|D,C)_{MLD,C}$ along borehole V2 based on (a) $D + C_1$ and (b) $D + C_2$ in CV I.

in W2a, Y3 and Z2 play the role of existing data D , boreholes X2 and Y2 are the sites of potential new data C_1 and C_2 , the goal being to predict $\log k(\Delta)$ along V2; CV II, in which $\log k$ measured in W2a, X2 and Y2 play the role of D , boreholes V2 and Z2 are the sites of potential new data C_1 and C_2 , the goal being to predict $\log k(\Delta)$ along Y3 (true posterior data are denoted by C' and their pre-posterior estimates by C). Given that in each case one has funds to measure $\log k$ in only one borehole, along which among boreholes X2 and Y2 should one conduct such measurements in case CV I, and along which among V2 and Z2 in case CV II?

The issue is compounded by uncertainty about the correct geostatistical model and parameters to be employed in each case. To address it, the data (D and C) are viewed as a Gaussian random field with covariance described by three alternative variogram models: exponential (Exp0), spherical (Sph0) and power (Pow0) each having two parameters (sill and range in the case of Exp0 and Sph0, variance coefficient and power in the case of Pow0); our abbreviations are consistent with those in [59] where the 0 indicates no (zero-order) drift. In each case, variogram parameters are unknown and are estimated from the data using maximum likelihood. Uncertainty of the estimated parameters, θ , is quantified by a covariance matrix, Γ . The parameters are assumed to have a multivariate normal distribution, $N(\hat{\theta}, \Gamma)$, about their estimates, from which random samples are drawn by a Monte Carlo method (Appendix A, Step 4a). Details about these models,

Fig. 10. Variation of $\text{Var}(\Delta|D)_{MLB,C}$, $E_{C,D} \text{Var}(\Delta|D,C)_{MLB,C}$ and $\text{Var}_{C,D} E(\Delta|D,C)_{MLB,C}$ along predicted borehole V2 based on (a) $D + C_1$ and (b) $D + C_2$ disregarding parametric uncertainty in CV I.

parameter estimation, quantification of parameter uncertainty and estimation of model probabilities are given in [59–61]. Because of extensive previous studies of this dataset and the cross-validation nature of our study, no new models emerge in the pre-posterior and posterior stages. Consequently, we continue to use model set

M and set it equal to **P** in using data sets $\{D, C\}$ and $\{D, C'\}$ as we did in using data **D**. Fig. 2 shows how the three variogram models calibrated against **D**, $D + C_1$ and $D + C_2$ compare with sample variograms in CV I and CV II. Anisotropy and multi-modality are not exhibited in the sample variograms. The corresponding posterior probabilities based on the information theoretic model discrimination criterion AIC [2] and on KIC are listed in Table 1; those based on the information theoretic criterion AICc [21] and on BIC are similar to those based on AIC and are therefore not listed. Whereas KIC favors model Exp0 in all 6 cases listed in Table 1, AIC prefers Exp0 in 2 cases and Pow0 in 4 cases. Considering the Bayesian nature of our methodology we present below only results associated with KIC.

3.1. Effect of augmentation with actual data

To evaluate the predictive uncertainty of Δ associated with data sets **D**, $\{D, C_1'\}$ and $\{D, C_2'\}$ where primes indicate actual (posterior, measured) values we employ MC according to steps 1–5 in Appendix A. We generate $R_\theta = 1000$ realizations of θ_k according to step 4a while insuring that their mean and variance have stabilized for each variogram model. Values of Δ are predicted by using the KT3D package of [13], revised to produce a predictive (kriging) covariance according to [52]. Fig. 3 compares measured log *K* values along borehole V2 with predicted values and their 95% confidence intervals obtained with each geostatistical model and with MLBMA based on **D**, $\{D, C_1'\}$ and $\{D, C_2'\}$ in CV I and Fig. 4 does so for CV II. All measurements in Fig. 3 are contained within 95% confidence intervals, which are seen to narrow down slightly with data augmentation; Fig. 3b and d are similar to each other due to the high probability of model Exp0 obtained on the basis of KIC Table 1. While similar patterns are observed in Fig. 4, some measurements lie outside the 95% confidence intervals and predictions

Fig. 11. $E_{\theta_k|M_k,D,C} E(\Delta|\theta_k, M_k, D, C)$ (red) and $E(\Delta|\hat{\theta}_k^{D,C}, M_k, D, C)$ (blue) corresponding to predicted data index 13 (Fig. 10) in borehole V2 obtained for 200 realizations of C_1 in borehole X2, CV I, using model (a) Pow0, (b) Exp0 and (c) Sph0; $E_{M_k,D,C} E_{\theta_k|M_k,D,C} E(\Delta|\theta_k, M_k, D, C)$ (red) and $E_{M_k,D,C} E(\Delta|\hat{\theta}_k^{D,C}, M_k, D, C)$ (blue) obtained using (d) MLBMA. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 12. Coefficients of variation (CV) of 1,000 realizations of two parameters of models (a1–a2) Pow0, (b1–b2) Exp0 and (c1–c2) Sph0 for each realization in Fig. 11.

deviate from measurements to a greater extent than they do in Fig. 3. This confirms [61] that it is easier to predict log k along borehole V2 than along borehole Y3 on the basis of other site data. Below we focus exclusively on results of MLBMA analyses.

Figs. 5 and 6 plot variations of $\text{Var}(\Delta|\mathbf{D})_{ML^D}$, $\text{Var}(\Delta|\mathbf{D}, \mathbf{C}')_{ML^{D,C'}}$ and their difference, $\text{Var}(\Delta|\mathbf{D})_{ML^D} - \text{Var}(\Delta|\mathbf{D}, \mathbf{C}')_{ML^{D,C'}}$, along boreholes V2 and Y3 in CV I and CV II, respectively. The figures show that augmenting the samples reduces predictive variances along both boreholes. Table 2 lists actual trace variance reductions, $\text{Tr}[\text{Cov}(\Delta|\mathbf{D})_{ML^D}] - \text{Tr}[\text{Cov}(\Delta|\mathbf{D}, \mathbf{C}')_{ML^{D,C'}}]$, and their predicted counterparts, $\text{Tr}[\text{Cov}_{C_1|D}E(\Delta|\mathbf{D}, \mathbf{C}_1)]$, for the two cross-validation cases. In CV I, collecting new data along borehole X2 is seen to result in greater uncertainty reduction than doing the same along borehole Y2 regardless of whether one relies on actual data or their predictions; in CV II collecting new data along borehole V2 is more advantageous than doing so along borehole Z2 for actual data and their predictions.

3.2. Effect of augmentation with randomly generated data

Since in reality the actual data $\mathbf{C}'$ are unknown, we generate $R_c = 200$ pre-posterior realizations $\mathbf{C}$ of the same using available data $\mathbf{D}$ in accord with steps 6 and 7 of our procedure (Appendix A). Fig. 7 shows $\mathbf{C}'$, $E(\mathbf{C}|\mathbf{D})_{ML^D}$, the 200 realizations and their 95% confidence intervals in CV I. $E(\mathbf{C}|\mathbf{D})_{ML^D}$ is seen to represent a smoothed version of $\mathbf{C}'$, suggesting a slight bias in the generated

$\mathbf{C}$ values. Some values of $\mathbf{C}'$ lie outside the confidence intervals, indicating that the generated $\mathbf{C}$ values do not reflect all information contained in $\mathbf{C}'$. The same is true for CV II (result not shown).

Following Steps 8 and 9 allows us to compute and plot variations in $\text{Var}(\Delta|\mathbf{D})_{ML^{D,C}}$, $E_{C_1|D}\text{Var}(\Delta|\mathbf{D}, \mathbf{C}_1)_{ML^{D,C_1}}$ and $\text{Var}_{C_1|D}E(\Delta|\mathbf{D}, \mathbf{C}_1)_{ML^{D,C_1}}$ along borehole V2 in Fig. 8 and along borehole Y3 in Fig. 9. We verified that 200 realizations are enough for sample estimates of all three terms to stabilize. A comparison of Figs. 8 and 5 reveals that, although $\text{Tr}[\text{Cov}(\Delta|\mathbf{D})_{ML^{D,C_1}}] = 13.64$ and $\text{Tr}[\text{Cov}(\Delta|\mathbf{D})_{ML^{D,C_2}}] = 13.78$ exceed $\text{Tr}[\text{Cov}(\Delta|\mathbf{D})_{ML^D}] = 13.38$ by a slight amount (probably due to sampling errors stemming from size and accuracy limitations on $\mathbf{D}$), the spatial patterns in the two figures are similar. Likewise, though $E_{C_1|D}\text{Var}(\Delta|\mathbf{D}, \mathbf{C}_1)_{ML^{D,C_1}}$ and $E_{C_2|D}\text{Var}(\Delta|\mathbf{D}, \mathbf{C}_2)_{ML^{D,C_2}}$ in Fig. 8 tends to exceed $\text{Var}(\Delta|\mathbf{D}, \mathbf{C}')_{ML^{D,C'}}$ and $\text{Var}(\Delta|\mathbf{D}, \mathbf{C}_2)_{ML^{D,C_2}}$ in Fig. 5, these too have near identical spatial patterns. One therefore expects the estimated variance reduction $\text{Var}_{C_1|D}E(\Delta|\mathbf{D}, \mathbf{C}_1)_{ML^{D,C_1}}$ to exhibit a pattern similar to that of the true variance reduction $\text{Var}(\Delta|\mathbf{D})_{ML^D} - \text{Var}(\Delta|\mathbf{D}, \mathbf{C}')_{ML^{D,C'}}$. The predicted variance reduction measure $\text{Tr}[\text{Cov}_{C_1|D}E(\Delta|\mathbf{D}, \mathbf{C}_1)_{ML^{D,C_1}}] = 0.59$ Table 2 is smaller than its true (posterior) counterpart $\text{Tr}[\text{Cov}(\Delta|\mathbf{D})_{ML^D}] - \text{Tr}[\text{Cov}(\Delta|\mathbf{D}, \mathbf{C}_1)_{ML^{D,C_1}}] = 2.31$ while $\text{Tr}[\text{Cov}_{C_1|D}E(\Delta|\mathbf{D}, \mathbf{C}_2)_{ML^{D,C_2}}] = 0.42$ is smaller than $\text{Tr}[\text{Cov}(\Delta|\mathbf{D})_{ML^D}] - \text{Tr}[\text{Cov}(\Delta|\mathbf{D}, \mathbf{C}_2)_{ML^{D,C_2}}] = 1.01$; comparing Figs. 9 and 6 reveals similar relationships in the CV II case Table 2.

Fig. 13. Coefficients of variation (CV) of $E(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C})$ predicted on the basis of 1,000 realizations of two parameters of models (a1–a2) Pow0, (b1–b2) Exp0 and (c1–c2) Sph0 for each realization in Fig. 12.

As already noted, our generated $\mathbf{C}$ values do not contain all information about the actual data $\mathbf{C}^*$. This explains, for example, why actual data along boreholes V2 ($\mathbf{C}_1$) and Z2 ($\mathbf{C}_2$) help reduce predictive uncertainty along the entire borehole Y3 (Fig. 6) but generated values along boreholes V2 and Z2 help reduce uncertainty only along the bottom (Fig. 9a) and top (Fig. 9b) parts of borehole Y3. This makes intuitive sense considering that the bottom part of borehole Y3 is closer to V2 than to Z2 and its upper part is closer to Z2 than to V2.

Even though our generated $\mathbf{C}$ values underestimate the potential of such data to help reduce predictive uncertainty, Table 2 makes clear that they lead to correct choices of boreholes to sample (X2 in CV I and V2 in CV II).

4. Accuracies of the three approximations

Upon disregarding parameter uncertainty approximation (21) eliminates the need to simulate θ_k in Step 8d. Fig. 10 plots variations in $\text{Var}(\Delta|\mathbf{D})_{ML^{D,C}}$, $E_{C|D}\text{Var}(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}$ and $\text{Var}_{C|D}E(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}$

along borehole V2 in CV I computed in this manner. A comparison of Figs. 10 and 8 reveals that patterns of uncertainty reduction in CV I are not affected by the approximation; though we do not show it, the same holds true for CV II. Table 2 indicates that the approximation has minimal effect on computed measures of uncertainty reduction, causing them to increase on average by about 15% in the two cross validation cases and to remain unaffected along borehole Z2 in CV II. Table 2 suggests selecting boreholes X2 and V2 in CV I and CV II, respectively, as targets for collecting additional data regardless of whether or not parameter uncertainty is considered.

To gain insight into the question why parameter uncertainty has little impact on our results, we compare the predictive uncertainty reduction measure $\text{Tr}[\text{Cov}_{C|D}E_{M_k|\mathbf{D}, \mathbf{C}}E_{\theta_k|M_k, \mathbf{D}, \mathbf{C}}E(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C})]$, represented by the last term in (19) or equivalently by $\text{Tr}[\text{Cov}_{C|D}E(\Delta|\mathbf{D}, \mathbf{C})]$ in (17), with its approximation $\text{Tr}[\text{Cov}_{C|D}E_{M_k|\mathbf{D}, \mathbf{C}}E(\Delta|\hat{\theta}_k^{D,C}, M_k, \mathbf{D}, \mathbf{C})]$ in (21). Fig. 11a–c plot $E_{\theta_k|M_k, \mathbf{D}, \mathbf{C}}E(\Delta|\theta_k, M_k, \mathbf{D}, \mathbf{C})$ and $E(\Delta|\hat{\theta}_k^{D,C}, M_k, \mathbf{D}, \mathbf{C})$ corresponding to predicted data index 13 (Fig. 10) in borehole V2 (for which the difference between these measures is the largest)

obtained for 200 realizations of $\mathbf{C}_1$ in borehole X2 (CV I) using models Pow0, Exp0 and Sph0; Fig. 11d plots $E_{M_k, \mathbf{D}, \mathbf{C}} E_{\theta_k | M_k, \mathbf{D}, \mathbf{C}} E(\Delta | \theta_k, M_k, \mathbf{D}, \mathbf{C})$ (red) and $E_{M_k, \mathbf{D}, \mathbf{C}} E(\Delta | \hat{\theta}_k^{D, C}, M_k, \mathbf{D}, \mathbf{C})$ (blue) obtained using MLBMA. The exact and approximate measures are similar in all realizations, the sample variance of $E_{M_k, \mathbf{D}, \mathbf{C}} E_{\theta_k | M_k, \mathbf{D}, \mathbf{C}} E(\Delta | \theta_k, M_k, \mathbf{D}, \mathbf{C})$ being 0.042 and that of $E_{M_k, \mathbf{D}, \mathbf{C}} E(\Delta | \hat{\theta}_k^{D, C}, M_k, \mathbf{D}, \mathbf{C})$ 0.048. Fig. 12 plots coefficients of variation (CV) of 1000 realizations of the two parameters of models Pow0, Exp0, and Sph0 for each of the 200 realizations in Fig. 11. Fig. 13 plots coefficients of variation of $E(\Delta | \theta_k, M_k, \mathbf{D}, \mathbf{C})$ predicted on the basis of these 1000 parameter realizations for each of the 200 data realizations along X2. The coefficients of parameter variation in Fig. 12 are seen to be much larger than those of corresponding variations in $E(\Delta | \theta_k, M_k, \mathbf{D}, \mathbf{C})$, indicating that the latter is not very sensitive to the former. This explains why our first approximation works.

Upon disregarding data uncertainty approximation (23) eliminates the need to simulate $\mathbf{C}$ in Steps 7–9. Here uncertainty reduction is no longer measured by $\text{Tr}[\text{Cov}_{\mathbf{C}} E(\Delta | \mathbf{D}, \mathbf{C})]_{ML^D, \mathbf{C}}$ but by $\text{Tr}[\text{Cov}(\Delta | \mathbf{D})_{ML^D}] - \text{Tr}[\text{Cov}(\Delta | \mathbf{D}, E(\mathbf{C} | \mathbf{D}))]_{ML^D, E(\mathbf{C} | \mathbf{D})}$. Fig. 14 plots variations of $\text{Var}(\Delta | \mathbf{D})_{ML^D}$, $\text{Var}(\Delta | \mathbf{D}, E(\mathbf{C} | \mathbf{D}))_{ML^D, E(\mathbf{C} | \mathbf{D})}$ and their difference along borehole V2 in CV I. Comparing Fig. 14 with Figs. 8 and 10 shows a pattern of uncertainty reduction similar to those obtained earlier while taking data uncertainty into account; the same happens in CV II (results not shown). However, disregarding data uncertainty is seen to cause the predicted uncertainty reduction to be significantly exaggerated as evidenced further by corresponding measures of uncertainty reduction in Table 2. These measures nevertheless show a preference for boreholes X2 and V2 as new data collection targets in CV I and CV II, respectively, consistent with previous results. However, the estimated variance reductions for boreholes X2 and V2 in CV I are almost identical, implying that borehole selection is tenuous in this case.

To gain insight into the question why data uncertainty has major impact on our results, we compare the first three terms on the right hand side of (19), representing $E_{\mathbf{C} | \mathbf{D}} \text{Cov}(\Delta | \mathbf{D}, \mathbf{C})$, with the three terms on the right hand side of (23), representing $\text{Cov}(\Delta | \mathbf{D}, E(\mathbf{C} | \mathbf{D}))$ in our

Fig. 14. Variation of $\text{Var}(\Delta | \mathbf{D})_{ML^D}$, $\text{Var}(\Delta | \mathbf{D}, E(\mathbf{C} | \mathbf{D}))_{ML^D, E(\mathbf{C} | \mathbf{D})}$ and $\text{Var}(\Delta | \mathbf{D})_{ML^D} - \text{Var}(\Delta | \mathbf{D}, E(\mathbf{C} | \mathbf{D}))_{ML^D, E(\mathbf{C} | \mathbf{D})}$ along predicted borehole V2 based on (a) $\mathbf{D} + E(\mathbf{C}_1 | \mathbf{D})$ and (b) $\mathbf{D} + E(\mathbf{C}_2 | \mathbf{D})$ disregarding data uncertainty in CV I.

Fig. 15. Variation of three components of $E_{\mathbf{C} | \mathbf{D}} \text{Cov}(\Delta | \mathbf{D}, \mathbf{C})$ in (21) (solid) (a) $E_{\mathbf{C} | \mathbf{D}} E_{M_k | \mathbf{D}, \mathbf{C}} E_{\theta_k | M_k, \mathbf{D}, \mathbf{C}} \text{Cov}(\theta_k | M_k, \mathbf{D}, \mathbf{C})$, (b) $E_{\mathbf{C} | \mathbf{D}} E_{M_k | \mathbf{D}, \mathbf{C}} \text{Cov}(\theta_k | M_k, \mathbf{D}, \mathbf{C}) E(\Delta | \theta_k, M_k, \mathbf{D}, \mathbf{C})$, and (c) $E_{\mathbf{C} | \mathbf{D}} \text{Cov}(\theta_k | M_k, \mathbf{D}, \mathbf{C}) E_{\theta_k | M_k, \mathbf{D}, \mathbf{C}} E(\Delta | \theta_k, M_k, \mathbf{D}, \mathbf{C})$ and three components of $\text{Cov}(\Delta | \mathbf{D}, E(\mathbf{C} | \mathbf{D}))$ in (23) (dashed) (a) $E_{M_k | \mathbf{D}, E(\mathbf{C} | \mathbf{D})} E_{\theta_k | M_k, \mathbf{D}, E(\mathbf{C} | \mathbf{D})} \text{Cov}(\theta_k | M_k, \mathbf{D}, E(\mathbf{C} | \mathbf{D}))$, (b) $E_{M_k | \mathbf{D}, E(\mathbf{C} | \mathbf{D})} \text{Cov}(\theta_k | M_k, \mathbf{D}, E(\mathbf{C} | \mathbf{D})) E(\Delta | \theta_k, M_k, \mathbf{D}, E(\mathbf{C} | \mathbf{D}))$, and (c) $\text{Cov}(\theta_k | M_k, \mathbf{D}, E(\mathbf{C} | \mathbf{D})) E_{\theta_k | M_k, \mathbf{D}, E(\mathbf{C} | \mathbf{D})} E(\Delta | \theta_k, M_k, \mathbf{D}, E(\mathbf{C} | \mathbf{D}))$ along borehole V2 in CV I based on $\{\mathbf{D}, \mathbf{X}_2\}$.

second approximation. As shown in Fig. 15, all three terms, but particularly the first one, are seriously underestimated upon disregarding data uncertainty. The pattern and magnitude of the differences between $E_{\mathbf{C} | \mathbf{D}} E_{M_k | \mathbf{D}, \mathbf{C}} E_{\theta_k | M_k, \mathbf{D}, \mathbf{C}} \text{Cov}(\theta_k | M_k, \mathbf{D}, \mathbf{C})$ and $E_{M_k | \mathbf{D}, E(\mathbf{C} | \mathbf{D})} E_{\theta_k | M_k, \mathbf{D}, E(\mathbf{C} | \mathbf{D})} \text{Cov}(\theta_k | M_k, \mathbf{D}, E(\mathbf{C} | \mathbf{D}))$ shown in Fig. 15a are similar to those between $\text{Var}(\Delta | \mathbf{D})_{ML^D}$ and $\text{Var}(\Delta | \mathbf{D}, E(\mathbf{C} | \mathbf{D}))_{ML^D, E(\mathbf{C} | \mathbf{D})}$ in Fig. 14a. This is so because kriging variance conditioned on $\mathbf{C}$ is significantly larger than that conditioned on $E(\mathbf{C} | \mathbf{D})$ due to the smooth nature of the latter, as seen in Fig. 7. Correspondingly, disregarding data uncertainty results in underestimation of predictive uncertainty and overestimation of its reduction. The smoother is $E(\mathbf{C} | \mathbf{D})$ in comparison to $\mathbf{C}$ or $\mathbf{C}'$, the more pronounced this effect would be.

Disregarding both data and parameter uncertainties according to (24) eliminates the need to simulate both θ_k and $\mathbf{C}$ in Steps 7–9, rendering this approach most efficient computationally. Here again uncertainty reduction is measured by $\text{Tr}[\text{Cov}(\Delta | \mathbf{D})_{ML^D}] - \text{Tr}[\text{Cov}(\Delta | \mathbf{D}, E(\mathbf{C} | \mathbf{D}))_{ML^D, E(\mathbf{C} | \mathbf{D})}]$. Fig. 16 plots variations of $\text{Var}(\Delta | \mathbf{D})_{ML^D}$, $\text{Var}(\Delta | \mathbf{D}, E(\mathbf{C} | \mathbf{D}))_{ML^D, E(\mathbf{C} | \mathbf{D})}$ and the difference between them along borehole V2 in CV I. The corresponding pattern of uncertainty reduction is practically identical to that in Fig. 14; the same happens in CV II (results not shown). Corresponding trace variance reductions in Table 2 indicate a preference for boreholes X2 and V2 as new data collection targets in CV I and CV II, respectively, consistent with previous results. However, the estimated uncertainty reductions in CV I do not support an unambiguous preference for borehole X2.

Fig. 16. Variation of $\text{Var}(\Delta|\mathbf{D})_{ML^D}$, $\text{Var}(\Delta|\mathbf{D}, E(\mathbf{C}|\mathbf{D}))_{ML^{D,E(\mathbf{C}|\mathbf{D})}}$ and $\text{Var}(\Delta|\mathbf{D})_{ML^D} - \text{Var}(\Delta|\mathbf{D}, E(\mathbf{C}|\mathbf{D}))_{ML^{D,E(\mathbf{C}|\mathbf{D})}}$ along predicted borehole V2 based on (a) $\mathbf{D} + E(\mathbf{C}_1|\mathbf{D})$ and (b) $\mathbf{D} + E(\mathbf{C}_2|\mathbf{D})$ disregarding data and parametric uncertainties in CV I.

Table 3

Estimated values of coefficient (for Pow0) and sill (for Exp0 and Sph0) based on $\mathbf{C}'$ and $E(\mathbf{C}|\mathbf{D})$.

Model	Pow0	Exp0	Sph0	Pow0	Exp0	Sph0
Case	CV I			CV II		
$\mathbf{C}'_1$	0.28	0.68	1.20	0.18	0.67	0.86
$E(\mathbf{C}_1 \mathbf{D})$	0.23	0.61	0.97	0.14	0.41	0.60
$\mathbf{C}'_2$	0.26	0.78	1.07	0.26	0.59	0.82
$E(\mathbf{C}_2 \mathbf{D})$	0.23	0.62	0.98	0.13	0.36	0.56

We thus see that whereas disregarding parameter uncertainty has minimal effect on computed measures of uncertainty reduction, disregarding data uncertainty exaggerates these measures significantly regardless of whether or not parameter uncertainty is considered. This is so because disregarding data uncertainty renders the predictive variance in our geostatistical analysis equivalent to the kriging variance, which in turn is proportional to the coefficient of the power model and to the sills of the exponential and spherical models. As shown in Table 3, estimates of these parameters become suppressed when data uncertainty is ignored. This causes predictive uncertainty, $\text{Var}(\Delta|\mathbf{D}, E(\mathbf{C}|\mathbf{D}))_{ML^{D,E(\mathbf{C}|\mathbf{D})}}$, to be suppressed and the variance reduction measure $\text{Tr}[\text{Cov}(\Delta|\mathbf{D})_{ML^D}] - \text{Tr}[\text{Cov}(\Delta|\mathbf{D}, E(\mathbf{C}|\mathbf{D}))_{ML^{D,E(\mathbf{C}|\mathbf{D})}}]$ to become inflated.

5. Conclusions and discussion

1. Our multimodel Bayesian approach to data-worth analysis [39] appears to work well when applied to log air permeability data from vertical and inclined boreholes drilled into unsaturated fractured tuff. In particular, the selection of targets for additional data collection during the pre-posterior stage is validated against actual data collected during the posterior stage of the analysis.
2. In our case study, partial linearization of the posterior variance equations achieved by disregarding parameter uncertainty brings about an improvement in computational efficiency

without affecting the results in any significant way. Though parameter uncertainty is significant, our results are only marginally sensitive to this uncertainty.

3. Partial or complete linearization of our posterior variance equations, achieved by disregarding data uncertainty while either considering or disregarding parameter uncertainty, brings about a significant reduction in computational effort. Whereas in our case study such linearization has only a minor effect on predicted pattern of uncertainty reduction due to the potential collection of new data, it significantly overestimates the magnitude of this reduction. The overestimation is due to replacement of variable data by their smooth estimate without taking account of its uncertainty. Correspondingly, selection among alternative data collection targets becomes tenuous. It follows that whereas parameter uncertainty may sometimes be disregarded for purposes of analyzing the worth of data, the same does not generally apply to uncertainty in estimates of potential new data.
4. Our study was limited to predetermined borehole locations so as to allow cross-validating our results against known data. In practice one may want to employ an optimization algorithm to select among alternative candidate locations.
5. Though demonstration of our Bayesian data-worth analysis has been limited to geostatistical models, the approach is general enough to apply to any models including those of groundwater flow and contaminant transport. Considering that the choice of variogram model may affect flow and transport modeling [23,42,46], a study such as ours could be considered either a prelude to groundwater flow and transport modeling or an integral part thereof.

Acknowledgement

The first and second authors were supported in part by NSF-EAR Grant 0911074 and DOE-ERSP Grant DE-SC0002687. The second author was also supported by the COFRS project of the Florida State University. The third and fourth authors were supported in part through a contract between the University of Arizona and Vanderbilt University under the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) III, funded by the US Department of Energy.

Appendix A. Computational implementation of MLBMA framework

To assess the impact of data augmentation within the above MLBMA framework computationally we propose the following approach:

1. Postulate a set $\mathbf{M}$ of K mutually independent geostatistical, statistical or stochastic models, M_k , with parameters θ_k for the desired output vector, Δ ;
2. Obtain ML estimates $\hat{\theta}_k^D$ of θ_k by calibrating each M_k against available data $\mathbf{D}$ through minimization of the log likelihood $-2\ln p(\mathbf{D}|\mathbf{M}_k, \theta_k) - 2\ln p(\theta_k|\mathbf{M}_k)$, then compute the corresponding estimation covariance $\mathbf{\Gamma}_k^D$ and KIC_k^D ;
3. Compute $p(\mathbf{M}_k|\mathbf{D})_{ML^D} = \frac{\exp(-\frac{1}{2}\delta KIC_k^D)p(\mathbf{M}_k)}{\sum_{i=1}^K \exp(-\frac{1}{2}\delta KIC_i^D)p(\mathbf{M}_i)}$ where the subscript ML^D designates the ML estimation process in step 2;
4. For each model M_k estimate $E(\Delta|\mathbf{D}, M_k)_{ML^D}$ and $\text{Cov}(\Delta|\mathbf{D}, M_k)_{ML^D}$ either through second-order approximations of [36] or via Monte Carlo simulation (used in this study);
- a. Draw random samples (realizations) of θ_k from a multivariate Gaussian distribution with mean $\hat{\theta}_k^D$ and covariance $\mathbf{\Gamma}_k^D$;

- b. Estimate $E(\Delta|\mathbf{D}, M_k, \theta_k)_{ML^D}$ and $Cov(\Delta|\mathbf{D}, M_k, \theta_k)_{ML^D}$ for each realization of θ_k ;
- c. Average over all realizations of θ_k to obtain sample estimates of $E(\Delta|\mathbf{D}, M_k)_{ML^D} = E_{\theta_k} E(\Delta|\mathbf{D}, M_k, \theta_k)_{ML^D}$ and $Cov(\Delta|\mathbf{D}, M_k)_{ML^D} = E_{\theta_k} Cov(\Delta|\mathbf{D}, M_k, \theta_k)_{ML^D} + Cov_{\theta_k} E(\Delta|\mathbf{D}, M_k, \theta_k)_{ML^D}$.
5. Compute $E(\Delta|\mathbf{D})_{ML^D} = \sum_{k=1}^K E(\Delta|\mathbf{D}, M_k)_{ML^D} p(M_k|\mathbf{D})_{ML^D}$ and $Cov(\Delta|\mathbf{D})_{ML^D} = \sum_{k=1}^K Cov(\Delta|\mathbf{D}, M_k)_{ML^D} p(M_k|\mathbf{D})_{ML^D} + \sum_{k=1}^K [E(\Delta|\mathbf{D}, M_k)_{ML^D} - E(\Delta|\mathbf{D})_{ML^D}][E(\Delta|\mathbf{D}, M_k)_{ML^D} - E(\Delta|\mathbf{D})_{ML^D}]^T p(M_k|\mathbf{D})_{ML^D}$ and/or $Tr[Cov(\Delta|\mathbf{D})_{ML^D}]$;
6. Postulate a set $\mathbf{P}$ of I alternative geostatistical, statistical or stochastic models, P_i , with parameters π_i for a potential data set $\mathbf{C}$; the models P_i may be independent of M_k may form extensions of M_k or may coincide with the latter as in the computational examples given in this paper;
7. Predict multivariate statistics of $\mathbf{C}$, conditional on $\mathbf{D}$, either via BMA or via MLBMA by means of the model set $\mathbf{P}$; in the case of MLBMA the procedure would parallel that described for Δ in steps 2–6;
8. Estimate $E(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}$ and $Cov(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}$, where the subscript $ML^{D,C}$ designates the ML estimation process in step 2 but now with respect to an augmented data set $\{\mathbf{D}, \mathbf{C}\}$, either through second-order approximations or via Monte Carlo simulation by using the statistics of $\mathbf{C}$ from step 7 to generate random realizations of $\mathbf{C}$ (both options are explored in our synthetic example below); for each realization and for each model M_k :
 - a. Optionally linearize the residuals entering into the negative log likelihood $-2 \ln p(\mathbf{D}, \mathbf{C}|M_k, \theta_k) - 2 \ln p(\theta_k|M_k)$ about $\hat{\theta}_k^D$ (this option is not explored in this paper);
 - b. Obtain ML estimates $\hat{\theta}_k^{D,C}$ of θ_k by minimizing this negative log likelihood with respect to θ_k , then compute the corresponding estimation covariance $\Gamma_k^{D,C}$ and $KIC_k^{D,C}$;
 - c. Compute $p(M_k|\mathbf{D}, \mathbf{C})_{ML^{D,C}} = \frac{\exp(-\frac{1}{2} \hat{\theta}_k^{D,C} \Gamma_k^{D,C} \hat{\theta}_k^{D,C}) p(M_k)}{\sum_{i=1}^K \exp(-\frac{1}{2} \hat{\theta}_i^{D,C} \Gamma_i^{D,C} \hat{\theta}_i^{D,C}) p(M_i)}$;
 - d. For each model M_k estimate $E(\Delta|\mathbf{D}, \mathbf{C}, M_k)_{ML^{D,C}}$ and $Cov(\Delta|\mathbf{D}, \mathbf{C}, M_k)_{ML^{D,C}}$ via Monte Carlo simulation:
 - i. Draw random samples (realizations) of θ_k from a multivariate Gaussian distribution with mean $\hat{\theta}_k^{D,C}$ and covariance $\Gamma_k^{D,C}$;
 - ii. Estimate $E(\Delta|\mathbf{D}, \mathbf{C}, M_k, \theta_k)_{ML^{D,C}}$ and $Cov(\Delta|\mathbf{D}, \mathbf{C}, M_k, \theta_k)_{ML^{D,C}}$ for each realization of θ_k ;
 - iii. Average over all realizations of θ_k to obtain sample estimates of $E(\Delta|\mathbf{D}, \mathbf{C}, M_k)_{ML^{D,C}} = E_{\theta_k} E(\Delta|\mathbf{D}, \mathbf{C}, M_k, \theta_k)_{ML^{D,C}}$ and $Cov(\Delta|\mathbf{D}, \mathbf{C}, M_k)_{ML^{D,C}} = E_{\theta_k} Cov(\Delta|\mathbf{D}, \mathbf{C}, M_k, \theta_k)_{ML^{D,C}} + Cov_{\theta_k} E(\Delta|\mathbf{D}, \mathbf{C}, M_k, \theta_k)_{ML^{D,C}}$;
- e. Compute $E(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}} = \sum_{k=1}^K E(\Delta|\mathbf{D}, \mathbf{C}, M_k)_{ML^{D,C}} p(M_k|\mathbf{D}, \mathbf{C})_{ML^{D,C}}$ and $Cov(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}} = \sum_{k=1}^K Cov(\Delta|\mathbf{D}, \mathbf{C}, M_k)_{ML^{D,C}} p(M_k|\mathbf{D}, \mathbf{C})_{ML^{D,C}} + \sum_{k=1}^K [E(\Delta|\mathbf{D}, \mathbf{C}, M_k)_{ML^{D,C}} - E(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}][E(\Delta|\mathbf{D}, \mathbf{C}, M_k)_{ML^{D,C}} - E(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}]^T \cdot p(M_k|\mathbf{D}, \mathbf{C})_{ML^{D,C}}$ and/or $Tr[Cov(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}]$;
9. Average over all realizations of $\mathbf{C}$ to obtain sample estimates of $E(\Delta|\mathbf{D})_{ML^{D,C}} = E_{\mathbf{C}} E(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}$, $Cov(\Delta|\mathbf{D})_{ML^{D,C}} = E_{\mathbf{C}} Cov(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}} + Cov_{\mathbf{C}} E(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}$ and/or $Tr[Cov(\Delta|\mathbf{D})_{ML^{D,C}}]$; note that due to inevitable sampling errors and approximations associated with the ML estimation process $E(\Delta|\mathbf{D})_{ML^{D,C}}$, $Cov(\Delta|\mathbf{D})_{ML^{D,C}}$ and $Tr[Cov(\Delta|\mathbf{D})_{ML^{D,C}}]$ obtained at this step would generally differ, though ideally not by much, from $E(\Delta|\mathbf{D})_{ML^D}$, $Cov(\Delta|\mathbf{D})_{ML^D}$ and $Tr[Cov(\Delta|\mathbf{D})_{ML^D}]$ obtained at step 5;

10. Repeat steps 6–9 for different sets $\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3, \dots$ of potential data and select that set which maximizes the difference $Tr[Cov_{\mathbf{C}} E(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}] = Tr[Cov(\Delta|\mathbf{D})_{ML^{D,C}}] - Tr[E_{\mathbf{C}} Cov(\Delta|\mathbf{D}, \mathbf{C})_{ML^{D,C}}]$ between the trace conditional on $\mathbf{D}$ and the expected trace conditional on $\mathbf{D}$ and $\mathbf{C}$.

References

- [1] Abbaspour KC, Schulin R, Schlappi E, Fluhler H. A Bayesian approach for incorporating uncertainty and data worth in environmental projects. *Environ Model Assess* 1996;1:151–258.
- [2] Akaike H. A new look at statistical model identification. *IEEE Trans Autom. Control* 1974;AC-19:716–22.
- [3] Back P-E. A model for estimating the value of sampling programs and the optimal number of samples for contaminated soil. *Environ Geol* 2007;52:573–85. doi:10.1007/s00254-0488-6.
- [4] Ben-Zvi M, Berkowitz B, Kesler S. Pre-posterior analysis as a tool for data evaluation: application to aquifer contamination. *Water Resour Manage* 1988;2:11–20.
- [5] Beven KJ, Binley AM. The future of distributed models: model calibration and uncertainty prediction. *Hydrol Process* 1992;6:279–98.
- [6] Beven KJ, Freer J. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J Hydrol* 2001;249:11–29.
- [7] Box G. Sampling and Bayes' inference in scientific modelling and robustness. *J Roy Statist Soc Ser A* 1980;143(4):383–430.
- [8] Carrera J, Neuman SP. Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. *Water Resour Res* 1986;22(2):199–210.
- [9] Dausman AM, Doherty J, Langevin CD, Sukop MC. Quantifying data worth toward reducing predictive uncertainty. *Ground Water* 2010;48(5):729–40.
- [10] Davis DR, Kisiel CC, Duckstein L. Bayesian decision theory applied to design in hydrology. *Water Resour Res* 1972;8(1):33–41.
- [11] de Barros FPJ, Rubin Y. A risk-driven approach for subsurface site characterization. *Water Resour Res* 2008;44:W01414. doi:10.1029/2007WR006081.
- [12] de Barros FPJ, Rubin Y, Maxwell RM. The concept of comparative information yield curves and its application to risk-based site characterization. *Water Resour Res* 2009;45:W06401. doi:10.1029/2008WR007324.
- [13] Deutsch CV, Journel AG. *GSLIB: geostatistical software library and user's guide*. 2nd ed. New York: Oxford Univ. Press; 1998.
- [14] Diggle P, Lophaven S. Bayesian geostatistical design. *Scand J Statist* 2006;33(1):53–64.
- [15] Draper D. Assessment and propagation of model uncertainty. *J Roy Statist Soc B* 1995;57(1):45–97.
- [16] Feyen L, Gorelick SM. Framework to evaluate the worth of hydraulic conductivity data for optimal groundwater resources management in ecologically sensitive areas. *Water Resour Res* 2005;41:W03019. doi:10.1029/2003WR002901.
- [17] Freer J, Beven K, Ambrose B. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. *Water Resour Res* 1996;32(7):2161–73.
- [18] Freeze RA, Brude J, Massman J, Sperling T, Smith L. Hydrogeological decision analysis: 4. The concept of data worth and its use in the development of site investigation strategies. *Ground Water* 1992;30(4):288–574.
- [19] Guzman AG, Geddis AM, Henrich MJ, Lohrstorfer CF, Neuman SP. 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. Rep. NUREG/CR-6360, prepared for US Nuclear Regulatory Commission, Washington, DC, 1996.
- [20] Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Statist Sci* 1999;14(4):382–417.
- [21] Hurvich CM, Tsai C-L. Regression and time series model selection in small sample. *Biometrika* 1989;76(2):99–104.
- [22] IT-Corporation. Value of information analysis for corrective action unit no. 98, Frenchman Flat, US Department of Energy, USA; 1997.
- [23] Jafarpour B, Tarrahi M. Assessing the performance of the ensemble Kalman filter for subsurface flow data integration under variogram uncertainty. *Water Resour Res* 2011;47:W05537. doi:10.1029/2010WR009090.
- [24] James BR, Freeze RA. The worth of data in predicting aquitard continuity in hydrogeological design. *Water Resour Res* 1993;29(7):2049–65.
- [25] James BR, Gorelick SM. When enough is enough: the worth of monitoring data in aquifer remediation design. *Water Resour Res* 1994;30(12):2513–3499.
- [26] James BR, Gwo J-P, Toran L. Risk-cost decision framework for aquifer remediation design. *J Water Resour Plan Manage* 1996;122(6):220–414.
- [27] Kashyap RL. Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Trans Pattern Anal Machine Intell* 1982;4(2):99–104.
- [28] Kass R, Raftery A. Bayes factor. *J Am Statist Assoc* 1995;90(430):733–95.
- [29] Kitandis PK, Lane RW. Maximum likelihood parameter estimation of hydrologic spatial processes by the Gaussian-Newton method. *J Hydro* 1985;79:53–71.

- [30] Knopman DS, Voss CI. Multiobjective sampling design for parameter estimation and model discrimination in groundwater solute transport. *Water Resour Res* 1989;25(10):2245–58.
- [31] Leamer E. Specification searches: Ad hoc inference with nonexperimental data. first ed. New York: John Wiley & Sons; 1978. p. 384.
- [32] Lu D, Ye M, Neuman SP. Dependence of Bayesian model selection criteria and Fisher information matrix on sample size. *Math Geosci* 2011. doi:[10.1007/s11004-011-9359-0](https://doi.org/10.1007/s11004-011-9359-0).
- [33] Marchant VP, Lark RM. Optimized sample schemes for geostatistical surveys. *Math Geol* 2007;39(1). doi:[10.1007/s11004-006-9069-1](https://doi.org/10.1007/s11004-006-9069-1).
- [34] Matérn B. Spatial variation. Berlin: Springer; 1986.
- [35] McNulty G, Deshler B, Dove H. Value of information analysis. USA: Nevada Test Site; 1997.
- [36] Morales-Casique E, Neuman SP, Vesselinov VV. Maximum likelihood Bayesian averaging of airflow models in unsaturated tuff. *Stoch Environ Res Risk Assess* 2010;24:863–80. doi:[10.1007/s00477-010-0383-2](https://doi.org/10.1007/s00477-010-0383-2).
- [37] Müller WG. Collecting spatial data. Optimum design of experiments for random fields. Berlin: Springer; 2007.
- [38] Neuman SP. Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models. *Stoch Environ Res Risk Assess* 2003;17(5):291–305. doi:[10.1007/s00477-003-0151-7](https://doi.org/10.1007/s00477-003-0151-7).
- [39] Neuman SP, Xue L, Ye M, Lu D. Bayesian analysis of data-worth considering model and parameter uncertainty. *Adv Water Resour* 2011. doi:[10.1016/j.advwatres.2011.02.007](https://doi.org/10.1016/j.advwatres.2011.02.007).
- [40] Norberg T, Rosen L. Calculating the optimal number of contaminant samples by means of data worth analysis. *Environmetrics* 2006;17:705–19.
- [41] Nowak W, de Barros FPJ, Rubin Y. Bayesian geostatistical design: task-driven site investigation when the geostatistical model is uncertain. *Water Resour Res* 2010;46:W035535. doi:[10.1029/2009WR008312](https://doi.org/10.1029/2009WR008312).
- [42] Pardo-Iguzquiza E, Chico-Olmo M, García-Soldado MJ, Luque-Espinar JA. Using semivariogram parameter uncertainty in hydrogeological applications. *Ground Water* 2009;47(1):25–34.
- [43] Rada ES, Schultz GA. Worth of hydrological data in water resources projects. Application: Bolivians Amazonas zone, in *Water Management of the Amazon Basin* 1998; Internet Edition, UNESCO.
- [44] Raftery AE. Bayesian model selection in social research. *Sociol Methodol* 1995;25:111–63.
- [45] Raftery AE, Madigan D, Volinsky CT. Accounting for model uncertainty in survival analysis improves predictive performance. In: Bernardo J, Berger J, Dawid A, Smith A, editors. *Bayesian statistics*. United Kingdom: Oxford University Press; 1996. p. 323–49.
- [46] Riva M, Willmann M. Impact of log-transmissivity variogram structure on groundwater flow and transport predictions. *Adv Water Res* 2009;32:1311–22.
- [47] Rojas R, Feyen L, Batelaan O, Dassargues A. On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling. *Water Resour Res* 2010. doi:[10.1029/2009WR008822](https://doi.org/10.1029/2009WR008822).
- [48] Rosen L, LeGrand HE. An outline of a guidance framework for assessing hydrogeological risks at early stages. *Ground Water* 1997;35(2):195–204.
- [49] Rubin Y, Chen X, Murakami H, Hahn M. A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields. *Water Resour Res* 2010;46:W10523. doi:[10.1029/2009WR008799](https://doi.org/10.1029/2009WR008799).
- [50] Russell KT, Rabideau AJ. Decision analysis for pump-and-treat design. *Groundwater Monit Remed*, 2000;(Summer):159–268.
- [51] Sain SR, Furrer R. Combining climate model output via model correlations. *Stoch Environ Res Risk Assess* 2010;24(6):821–9. doi:[10.1007/s00477-010-0380-5](https://doi.org/10.1007/s00477-010-0380-5).
- [52] Samper FJ, Neuman SP. Estimation of spatial covariance structures by adjoint state maximum likelihood cross-validation: 1. Theory *Water Resour Res* 1989;25(3):351–62.
- [53] Sohn M, Small M, Pantazidou M. Reducing uncertainty in site characterization using Bayes Monte Carlo methods. *J Environ Eng - ASCE* 2000;126(10):893–902. doi:[10.1061/\(ASCE\)0733-9372\(2000\)126:10\(893\)](https://doi.org/10.1061/(ASCE)0733-9372(2000)126:10(893)).
- [54] Taplin RH. Robust likelihood calculation for time series. *J Roy Statist Soc Ser B* 1993;55:829–36.
- [55] Tiedeman CR, Hill MC, D'Agnese FA, Faunt CC. Methods for using groundwater model predictions to guide hydrogeologic data collection, with application to the death valley regional groundwater flow system. *Water Resour Res* 2003;39(1):1010. doi:[10.1029/2001WR001255](https://doi.org/10.1029/2001WR001255).
- [56] Tiedeman CR, Ely DM, Hill MC, O'Brien GM. A method for evaluating the importance of system state observations to model predictions, with application to the death valley regional groundwater flow system. *Water Resour Res* 2004;40:W12411. doi:[10.1029/2004WR003313](https://doi.org/10.1029/2004WR003313).
- [57] Volinsky CT, Madigan D, Raftery AE, Kronmal RA. Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *J Roy Statist Soc Ser C* 1997;46:433–48.
- [58] Wagner BJ. Evaluating data worth for ground-water management under uncertainty. *J Water Resour Plann Manage* 1999;125(5):281–8.
- [59] Ye M, Neuman SP, Meyer PD. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resour Res* 2004;40:W05113. doi:[10.1029/2003WR002557](https://doi.org/10.1029/2003WR002557).
- [60] Ye M, Neuman SP, Meyer PD, Pohlmann KF. Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff. *Water Resour Res* 2005;41:W12429. doi:[10.1029/2005WR004260](https://doi.org/10.1029/2005WR004260).
- [61] Ye M, Meyer PD, Neuman SP. On model selection criteria in multimodel analysis. *Water Resour Res* 2008;44:W03428. doi:[10.1029/2008WR006803](https://doi.org/10.1029/2008WR006803).
- [62] Ye M, Lu D, Neuman SP, Meyer PD. Comment on “Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window” by Frank T.-C. Tsai and Xiaobao Li. *Water Resour Res* 2010;46:W02801. doi:[10.1029/2009WR008501](https://doi.org/10.1029/2009WR008501).
- [63] Yokota F, Thompson K. Value of information literature analysis: a review of applications in health risk assessment. *Med Decis Mak* 2004;24:287–98.