Journal of Hydrology 529 (2015) 1859–1873

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Maximum likelihood Bayesian model averaging and its predictive analysis for groundwater reactive transport models

Dan Lu [a], Ming Ye [b], Gary P. Curtis [c,*]

[a] Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[b] Department of Scientific Computing, Florida State University, Tallahassee, FL, USA
[c] U.S. Geological Survey, Menlo Park, CA, USA

ARTICLE INFO

Article history:
Received 19 November 2014
Received in revised form 10 July 2015
Accepted 17 July 2015
Available online 1 August 2015
This manuscript was handled by Peter K. Kitanidis, Editor-in-Chief, with the assistance of Roseanna M. Neupauer, Associate Editor

Keywords:
Uncertainty analysis
Reactive transport
Maximum likelihood Bayesian model averaging

SUMMARY

While Bayesian model averaging (BMA) has been widely used in groundwater modeling, it is infrequently applied to groundwater reactive transport modeling because of multiple sources of uncertainty in the coupled hydrogeochemical processes and because of the long execution time of each model run. To resolve these problems, this study analyzed different levels of uncertainty in a hierarchical way, and used the maximum likelihood version of BMA, i.e., MLBMA, to improve the computational efficiency. This study demonstrates the applicability of MLBMA to groundwater reactive transport modeling in a synthetic case in which twenty-seven reactive transport models were designed to predict the reactive transport of hexavalent uranium (U(VI)) based on observations at a former uranium mill site near Naturita, CO. These reactive transport models contain three uncertain model components, i.e., parameterization of hydraulic conductivity, configuration of model boundary, and surface complexation reactions that simulate U(VI) adsorption. These uncertain model components were aggregated into the alternative models by integrating a hierarchical structure into MLBMA. The modeling results of the individual models and MLBMA were analyzed to investigate their predictive performance. The predictive logscore results show that MLBMA generally outperforms the best model, suggesting that using MLBMA is a sound strategy to achieve more robust model predictions relative to a single model. MLBMA works best when the alternative models are structurally distinct and have diverse model predictions. When correlation in model structure exists, two strategies were used to improve predictive performance by retaining structurally distinct models or assigning smaller prior model probabilities to correlated models. Since the synthetic models were designed using data from the Naturita site, the results of this study are expected to provide guidance for real-world modeling. Limitations of applying MLBMA to the synthetic study and future real-world modeling are discussed.

Published by Elsevier B.V.

1. Introduction

Model averaging (or multimodel analysis) has been applied in the last decade to groundwater modeling. Instead of using a single model for prediction, model averaging approaches combine predictions and associated predictive uncertainty of multiple competing models in a weighted average manner. The predictive performance of model averaging is expected to be better than that of a single model, because model averaging considers model uncertainty, which avoids the problem of underestimation of predictive uncertainty when using the single model with consideration of only parametric uncertainty. A variety of model averaging techniques have been developed (Ye et al., 2010a), and they are mainly different from each other in calculating the model averaging weights. The Bayesian model averaging (BMA) method is used in this study. BMA has been applied to a wide range of problems, including geostatistical problems (Ye et al., 2004, 2005, 2008a; Troldborg et al., 2007; Lu et al., 2011, 2012; Neuman et al., 2012), groundwater flow problems (Poeter and Anderson, 2005; Foglia et al., 2007, 2013; Ye et al., 2008b, 2010b; Rojas et al., 2008, 2009; Singh et al., 2010a,b; Riva et al., 2011; Chitsazan and Tsai, 2014), unsaturated flow problems (Wöhling and Vrugt, 2008), rainfall–runoff problems (Duan et al., 2007; Vrugt and Robinson, 2007; Dong et al., 2013) and geochemical reaction problems (Lu et al., 2013).

However, application of BMA to groundwater reactive transport problems has been infrequent for the following reasons. First, groundwater reactive transport modeling is complex with coupled

* Corresponding author. Tel.: +1 650 329 4553.
E-mail address: gpcurtis@usgs.gov (G.P. Curtis).

http://dx.doi.org/10.1016/j.jhydrol.2015.07.029
0022-1694/Published by Elsevier B.V.

components such as groundwater flow, solute transport, and biogeochemical reactions (Steefel et al., 2005). Because of the interdependence of the model components, the uncertainty sources at different levels need to be addressed in a hierarchical manner (Wainwright et al., 2014). Recently, Tsai and Elshall (2013) and Elshall and Tsai (2014) developed a hierarchical BMA approach for groundwater flow modeling. The hierarchical BMA is theoretically general, and can be extended to groundwater reactive transport modeling, although the extension has not been reported. Another reason that hampers BMA applications to groundwater reactive transport modeling is the long execution time due to the coupling of the hydro-bio-geochemical processes. The high computational cost in each model run poses difficulty for evaluating the model averaging weights. As discussed below, BMA requires evaluating high-dimensional integrals (in model parameter spaces) for individual models. When the run time of a model execution is long, the computational cost of the integrations is practically unaffordable, despite the increasing computer capabilities and development of parallel computing techniques. To resolve the problem, Neuman (2003) and Ye et al. (2004) developed MLBMA, a maximum likelihood version of BMA, which uses the Laplace approximation to evaluate the integrations around maximum likelihood parameter estimates by assuming that the likelihood function is concentrated about its maximum (Ye et al., 2010c). MLBMA may be inaccurate for multimodal and highly nonlinear problems. Therefore, the validity of applying MLBMA to groundwater reactive transport modeling is examined in this study.

This paper, for the first time, uses MLBMA to quantify model uncertainty in groundwater reactive transport modeling. The paper is focused on developing alternative reactive transport models, calibrating the models, and evaluating their probabilities. When developing the alternative models, three different levels of model uncertainty were considered, and each level corresponds to a model component with competing model propositions. The hierarchical BMA provides a systematic way of analyzing different sources of model uncertainty and of evaluating model probabilities. In addition to evaluating the feasibility of using MLBMA for groundwater reactive transport modeling, another goal of this study is to examine predictive performance of MLBMA and individual models. Although model averaging often produces robust predictions compared to individual model predictions, it is not always the case. Winter and Nychka (2010) showed in a mathematical analysis that the mean squared error of model averaging can be larger than that of the best model if certain criteria are not satisfied. Cavadias and Morin (1986) showed more than two decades ago that weighting of discharge simulations from several hydrological models resulted in reduced performance in comparison with the best model in 20% of cases considered. Duan et al. (2007) found that model averaging yielded reductions of daily root mean square error and daily absolute error of the ensemble mean for several state variables of prediction but not for all variables. This study conducted similar evaluation of predictive performance using predictive bias, root mean squared error (RMSE) and logscore. While predictive bias and RMSE consider only mean predictions, logscore considers both mean and variance of predictions.

Predictive performance of MLBMA depends directly on posterior model probability (i.e., model averaging weight), which in turn depends on prior model probability, a quantitative expression of a modeler's belief on relative model plausibility based on expert judgment. In comparison with other techniques of evaluating model averaging weights, MLBMA has a unique feature of using prior model probability to improve predictive performance (Ye et al., 2005). This is achieved by compromising the assumption that the model averaging weights estimated during model calibration are still valid for model prediction. This assumption is questionable, particularly when predicting conditions different from those of the calibration, which is typical in reality. In this case, prior information based on expert judgment can adjust the model weights by reducing the influence from the calibration. Another use of prior probability is to alleviate the impact of model correlation on the evaluation of model weights. The influence of model correlation on model averaging is an open question (Sain and Furrer, 2010; Bishop and Abramowitz, 2013), and there is no formal way to handle it. Ye et al. (2004) used an empirical way to assign relatively small prior probabilities to correlated models, but did not evaluate how this affects MLBMA predictive performance.

The MLBMA application and predictive analysis were conducted for a synthetic case of hexavalent uranium (U(VI)) reactive transport modeling based on previous work at a site near Naturita Colorado, where uranium contamination in groundwater has posed a threat to the environment (Kohler et al., 2004; Curtis et al., 2004, 2006, 2009). A synthetic 'true' model was first developed using the information and data for the site, and then used to generate the synthetic data used in this study for model calibration and predictive analysis. The development of alternative models considered three uncertainty sources commonly encountered in practice, i.e., uncertainty in parameterization of model parameters (e.g., hydraulic conductivity), uncertainty in configuration of model boundary, and uncertainty in formulation of geochemical reactions. By postulating three propositions for each of the three model components, a total of 27 models were developed. Model calibration and parametric uncertainty analysis of the individual models were conducted using the weighted least-squares based regression software, UCODE_2005 (Poeter et al., 2005), which can be used to implement MLBMA (Ye et al., 2008a).

The remaining part of the paper is organized as follows. Section 2 briefly describes the BMA and MLBMA methods. Section 3 introduces first the true model and the 27 alternative groundwater reactive transport models, then the model calibration and the evaluation of posterior model probabilities, last the examination of the validity of applying MLBMA. The comparison results of model averaging and individual models are presented in Section 4. Conclusions are drawn is Section 5.

2. Bayesian model averaging and its maximum likelihood version

The BMA and MLBMA methods are described briefly here to make the paper self-contained. Details of BMA are described in Draper (1995) and Hoeting et al. (1999), and specific information on MLBMA is provided by Neuman (2003) and Ye et al. (2004, 2008a).

2.1. Bayesian model averaging (BMA)

Consider an interested quantity, Δ, which can be predicted with a set of models $\mathbf{M} = (M_1, \ldots, M_K)$, each characterized by a vector of parameters $\boldsymbol{\theta}_k$, conditional on a discrete set of data, $\mathbf{D}$. The posterior distribution of Δ is (Hoeting et al., 1999)

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

i.e., the average over all models of the posterior distributions $p(\Delta|\mathbf{D}, M_k)$ associated with individual models, weighted by the model posterior probabilities $p(M_k|\mathbf{D})$. These weights are given by Bayes' rule in the form (Hoeting et al., 1999)

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

where

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

is the integrated likelihood of model M_k, $p(\mathbf{D}|M_k, \boldsymbol{\theta}_k)$ being the joint likelihood of this model and its parameters, $p(\boldsymbol{\theta}_k|M_k)$ the prior density of $\boldsymbol{\theta}_k$ under model M_k, and $p(M_k)$ the prior probability of M_k. All probabilities are implicitly conditional on the choice of models entering into the set $\mathbf{M}$.

For the specification of prior model probability, Kashyap (1982) suggested that in the absence of any contrary information, the models be assigned equal prior probabilities. Similarly, Hoeting et al. (1999) stated that assuming all models a priori equally likely is a reasonable neutral choice when there is insufficient prior reason to prefer one model over another. However, Draper (1999) and George (1999) questioned that, when several models yield nearly equivalent predictions, giving them equal prior model probability may tamper with the predictive power of BMA. In this study, we evaluated the impacts of prior model probability on predictive performance using three weighting strategies including uniform priors and adjusted priors to consider model correlation effects.

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

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

and

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

respectively. The first term on the right-hand side of Eq. (5) represents within-model variance, and the second term represents between-model variance which measures the model uncertainty.

2.2. Maximum likelihood Bayesian model averaging (MLBMA)

BMA defines the integrated likelihood $p(\mathbf{D}|M_k)$ of model M_k in Eq. (3) and requires computing the integral through exhaustive sampling of the prior parameter space $\boldsymbol{\theta}_k$ for each model, which is extremely computationally demanding especially for models with a large number of parameters. One way to resolve this issue is to use the Laplace approximation by evaluating the integration around the maximum likelihood estimates (MLE) of the parameters, $\widehat{\boldsymbol{\theta}}_k$. This yields MLBMA as first proposed by Neuman (2003) and employed by Ye et al. (2004, 2008a, and 2010c). In MLBMA, $p(\mathbf{D}|M_k)$ can be calculated using either of two model selection criteria, BIC or KIC. As KIC is more accurate than BIC (Ye et al., 2010c; Lu et al., 2011) to approximate Eq. (3), KIC is used in this study to approximate the model weight via

$$p(M_k|\mathbf{D}) = \frac{\exp(-\Delta KIC_k/2) p(M_k)}{\sum_{l=1}^{K} \exp(-\Delta KIC_l/2) p(M_l)}, \quad (6)$$

where $\Delta KIC_k = KIC_k - KIC_{\min}$ is the difference between the KIC of model M_k and the minimum KIC, $KIC_{\min}$. KIC is defined as (Ye et al., 2008a)

$$KIC_k = -2\ln[L(\widehat{\boldsymbol{\theta}}_k|\mathbf{D})] - 2\ln p(\widehat{\boldsymbol{\theta}}_k) - N_k \ln(2\pi) + \ln|\mathbf{F}_k|, \quad (7)$$

where N_k is the number of estimated parameters associated with model M_k; $-\ln[L(\widehat{\boldsymbol{\theta}}_k|\mathbf{D})]$ is the minimum of the negative log-likelihood function; and $p(\widehat{\boldsymbol{\theta}}_k)$ is the prior probability of $\boldsymbol{\theta}_k$ evaluated at $\widehat{\boldsymbol{\theta}}_k$; $\mathbf{F}_k$ is the observed Fisher information matrix (FIM), and its elements are defined as (Kashyap, 1982)

$$F_{k,ij} = -\left.\frac{\partial^2 \ln[L(\boldsymbol{\theta}_k|\mathbf{D})]}{\partial\theta_{ki}\partial\theta_{kj}}\right|_{\boldsymbol{\theta}_k=\widehat{\boldsymbol{\theta}}_k}. \quad (8)$$

Eq. (8) requires calculating the second order derivatives with respect to the parameters at the MLE. To save the computational time, the observed FIM is often approximated by the expected one, which is estimated as (Kitanidis and Lane, 1985; Ye et al., 2008; Lu et al., 2011),

$$F_{k,ij} \approx \langle F_{k,ij} \rangle = \mathbf{J}^T \boldsymbol{\omega} \mathbf{J}, \quad (9)$$

where $\mathbf{J}$ represents the Jacobian of observations with respect to the parameters, $\boldsymbol{\omega}$ is the weighting used in model calibration. The expected FIM can accurately approximate the observed FIM when the model is relatively linear in the region around the MLE of the parameters.

Models associated with smaller values of KIC are ranked higher than those associated with larger values. KIC is calculated by UCODE_2005 (Poeter et al., 2005) in this study. KIC is derived based on the Laplace approximation by assuming that the likelihood function $L(\boldsymbol{\theta}_k|\mathbf{D})$ is highly peaked near its maximum $\widehat{\boldsymbol{\theta}}_k$ (Ye et al., 2008a, 2010c). If the assumption is not satisfied (e.g., likelihood function having multiple peaks), KIC may not be accurate. In addition, the expected FIM may not be an accurate approximation in calculating KIC for highly nonlinear problems. The validity of using KIC to our reactive transport models is examined in Section 3.4.

3. Groundwater reactive transport models and model probabilities

The predictive performance of MLBMA was evaluated using a synthetic study designed based on the Naturita Uranium Mill Tailing Remedial Action (UMTRA) site. This site had a uranium mill where ore was processed to produce uranium concentrates. Water used at the mill was discharged to ponds, and tailings were stored on site. Although the mill tailings were removed in the 1980s and contaminated soils were removed in 1990s, the shallow groundwater had been previously contaminated (Curtis et al., 2006). This work developed a hypothetical 'true' model that was used to generate synthetic data and a set of simpler reactive transport models to simulate the uranium reactive transport.

3.1. True model of uranium reactive transport

In the synthetic study, the true model was developed for a three-dimensional, unconfined aquifer. It covers an area of 1.25×10^6 m^2 and has three layers. Using the uniform cell size of 7.62 m in width and 7.66 m in length, the model area is discretized into 310 rows and 69 columns. Within the model boundary (Fig. 1), about 9760 cells are active for each model layer, and a total of 29,280 active cells in the three layers. The thickness of layer one is not a constant in the simulation domain but it is thick enough to keep all cells wet during the simulation; the thickness of layer two is 0.3 m and layer three is 1.5 m. The unconfined aquifer is bounded on the west and bottom by no-flow boundaries and by the San Miguel River to the north, east and south. The aquifer is recharged by the river from the southeast of the simulation domain and discharges to the river north of the domain, as shown in Fig. 1. The aquifer is subjected to two kinds of recharge. One is the natural recharge from precipitation with recharge rate of 8.2×10^{-4} m/d distributed uniformly in space and time, which is about 2% of the annual precipitation of the Naturita area. The other is a constant flux of recharge from two pond sources with the rate of 1.64×10^{-3} m/d. The natural recharge is located in the area of mill

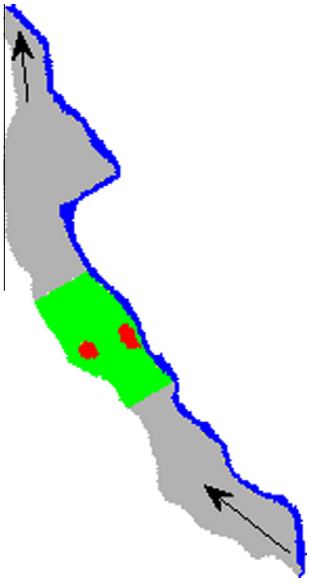

Fig. 1. Simulated domain of the true model where the blue area represents the river, the green area receives natural recharge, and the red area represents the two ponds sources of contaminant recharge. The arrows show flow direction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tailings in Naturita site, and the two ponds are located inside the natural recharge area, as shown in Fig. 1.

The hydraulic conductivity (K) fields of the aquifer have three hydrofacies. Facies one is the most permeable with a K value of 9.14 m/d; facies two is moderately permeable with a K value of 3.05 m/d; and facies three is the least permeable with a K value of 0.3 m/d. Based on available sample data from the Naturita study, the volume proportions of the three facies are 0.3, 0.2, and 0.5, respectively. Based on these volume proportions and the fitted transitional probabilities, we used the TPROGs software (Carle, 1999) to generate the K fields of the three layers for the true model. By keeping the total volume proportions of the three facies unchanged, we distributed different proportions of each facies in the three layers. For example, layer one has a large proportion of the most and moderate permeable materials for easy infiltration to lower layers; layer two, the thinnest layer, served as a semi-confining layer and has most of its volume occupied by the least permeable material. The generated K fields were then perturbed randomly to form the final K fields in the following procedure. First, a random field, R, was generated for the entire domain using sequential Gaussian simulation in GSLIB (Deutsch and Journel, 1998). The mean of the random field is zero, and its spatial correlation is quantified by an exponential variogram model $C(h) = 1.0\exp(-h/10.0)$, where h is lag distance between any two points. After the random field generation, the TPROGs-generated K fields and $\lambda \times R$ fields were added together, where λ equals to 1.8, 0.6, and 0.06 for the most, moderate, and the least permeable materials, respectively. These λ values were determined based on field observations of hydraulic conductivity to reflect that large K values have large variation. The spatial distributions of the final hydraulic conductivity fields of the three layers are shown in Fig. 2.

The transport simulation assumed three porosity values of 0.40, 0.35, and 0.25 corresponding to the three facies from the least permeable to the most permeable materials. As is common in stochastic subsurface hydrology, porosity is treated as a deterministic variable due to its small variability, and thus not perturbed. A longitudinal dispersivity of 3.0 m was used. This relatively small macrodispersion was used to avoid over-predicting mixing, which could drive many reactions and affect the adsorption and transport of U(VI) in the reactive transport modeling described below.

Synthetic surface complexation models (SCMs) were used in the true model to simulate the adsorption reactions of U(VI). Similar to the concept of reactive facies of Sassen et al. (2012), the U(VI) adsorption reactions for the three facies were simulated by different SCMs listed in Table 1. Each SCM in this synthetic model was adapted from studies of the Naturita site (Davis et al., 2004; Curtis et al., 2006) and from the Rifle UMTRA site (Hyun et al., 2009). The reactions in Table 1 were selected to provide a variable dependence of U(VI) adsorption on pH, carbonate activity and adsorption site affinity. For example, to simulate U(VI) adsorption reactions in facies one, the SCM consists of two reactions and two different site types, i.e., weak site denoted as Rw_OH and strong site denoted as Rs_OH. The values of the formation constant, $\log K$, for each reaction was manually adjusted to give U(VI) retardation factor of 2 for facies one, 4 for facies two, and 12 for facies three, using the average concentrations of the major ions in the uncontaminated groundwater and the U(VI) concentration of 1 μM. For different facies, the site concentrations were also different, and their values are shown in Table 1.

The model was simulated for two stress periods. Period one simulated 57 years with recharge from both the precipitation and the two pond sources; this simulated the Naturita site when the uranium mill was operated. Period two simulated 5 years with only recharge from the precipitation; this period simulated the Naturita site when the contaminated soils below the former mill tailings were excavated and transported offsite. The groundwater flow model was simulated using MODFLOW 2005 (Harbaugh, 2005); the river was simulated using the river package. Both the nonreactive and reactive transport models were simulated using PHT3D (Prommer, 2006).

The true model run generated a total of 360 Cl and 360 U(VI) concentrations as synthetic observations that were used for model calibration. They were collected from 4 different simulation times, year 57, 58, 60, and 62. At each time, they were collected from 30 locations (shown in Fig. 3) in each of the 3 layers, among which 12 locations were from existing observation wells at the Naturita site and an additional 18 locations were selected along the flow path lines, the upgradient area, and locations corresponding to different K values in the three layers. The 720 true values of Cl and U(VI) concentrations were then corrupted with measurement errors whose coefficient of variation is 5%. The 720 noisy data were finally used for the parameter estimation and multimodel analysis for the following defined alternative models. UCODE_2005 was used for the model calibration and uncertainty analysis such as evaluating prediction variances.

3.2. Alternative models of uranium reactive transport

Three model components were considered uncertain: the parameterizations of hydraulic conductivity (K) fields, the geometries of the north and west domain boundaries, and the surface complexation models to simulate the U(VI) adsorption. Three alternative parameterizations of the K fields were considered as the competing model propositions. The first parameterization (Heter) has heterogeneous fields. Using the true volume proportion of each facies in each layer, a TPROGs realization of the facies distributions was generated, and the spatial distribution of the K field in layer one is shown in Fig. 4a. The similarity between this K field and that of the true model (Fig. 2) presents a situation of detailed site characterization, which may not be very common in practice. In this parameterization, hydraulic conductivity is a constant within each facies so that the alternative models are simpler than the true model. The second parameterization (Zone) has two zones in layer

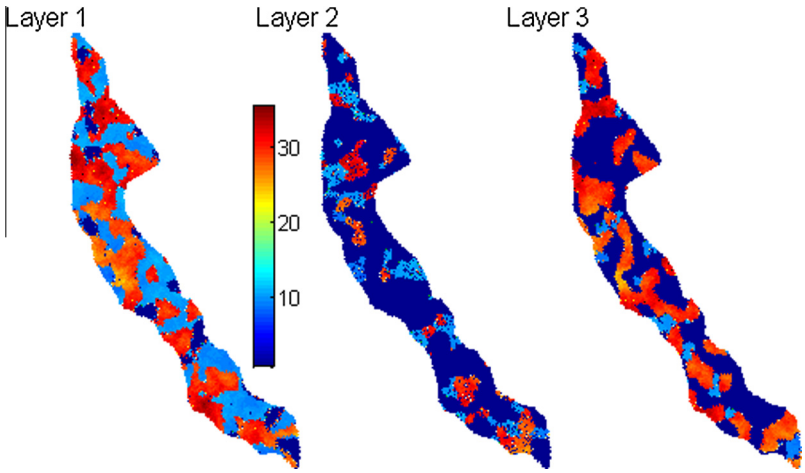

Fig. 2. Spatial distribution of hydraulic conductivity of the three layers of the true model.

Table 1
Surface complexation reactions and parameter values used in the true model.

Reactions	log K	Site concentration
Facies 1: 2 reactions and 2 sites		
$Rw\_OH + UO_2^{2+} = Rw\_OUO_2^{+} + H^{+}$	2.8	7.48×10^{-3}
$Rs\_OH + UO_2^{2+} + CO_3^{2-} = Rs\_OUO_2CO_3^{-} + H^{+}$	10.0	1.53×10^{-4}
Facies 2: 2 reactions and 2 sites		
$Sw\_OH + UO_2^{2+} = Sw\_OUO_2^{+} + H^{+}$	2.7	$3.21 \times 10^{2-}$
$Ss\_OH + UO_2^{2+} + 2CO_3^{2-} = Ss\_OHUO_2(CO_3)_2^{2-}$	22.0	9.92×10^{-4}
Facies 3: 4 reactions and 3 sites		
$Tw\_OH + UO_2^{2+} + CO_3^{2-} = Tw\_OUO_2CO_3^{-} + H^{+}$	7.4	$7.02 \times 10^{2-}$
$Ts\_OH + UO_2^{2+} + CO_3^{2-} = Ts\_OUO_2CO_3^{-} + H^{+}$	9.2	2.93×10^{-3}
$Ts\_OH + UO_2^{2+} + 2CO_3^{2-} = Ts\_OUO_2(CO_3)_2^{-} + H^{+}$	15.4	
$Tv\_OH + UO_2^{2+} + 2CO_3^{2-} = Tv\_OUO_2(CO_3)_2^{-3} + H^{+}$	16.4	1.17×10^{-4}

one and layer three as shown in Fig. 4b, but only one zone in layer two. While several zonation patterns were considered when designing the synthetic case, the one presented here produced the best overall results in model calibration. The third parameterization is denoted as Homo, in which each layer is homogeneous as shown in Fig. 4c.

Three alternative boundary configurations were considered. The first one (Bnd1) is the same as the true model (Fig. 4a), the second one (Bnd2) is the same as the one used in Curtis et al. (2006) which did not include the extended lobe in the north (Fig. 4b), and the third one (Bnd3) has an extension in the northwestern bound based on the second type of boundary as identified by the black line in Fig. 4c. The first boundary configuration has the largest area, because of the extended north and western boundaries of the downgradient domain. The three boundary configurations

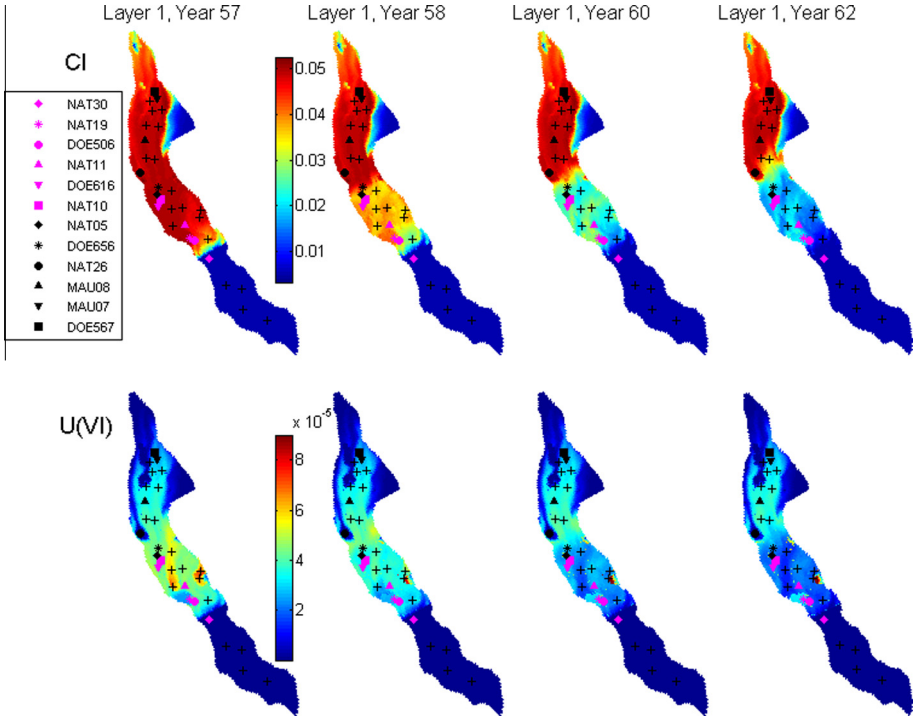

Fig. 3. Simulated concentration (M) of Cl (top) and U(VI) (bottom) of layer one after 57, 58, 60, and 62 years. The 30 symbols represent the 30 locations of calibration data. The 12 locations shown in the legend are existing observation wells from Naturita site, and the 18 plus symbols represent the new added locations.

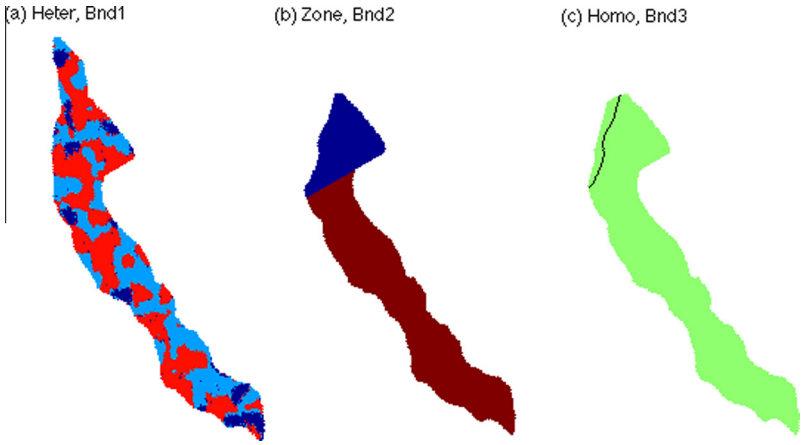

Fig. 4. Configuration of three alternative boundaries (Bnd1, Bnd2, and Bnd3) and hydraulic conductivity fields of layer one for the three alternative parameterization models (Heter, Zone and Home). Bnd3 has an extension in the northwestern bound based on Bnd2 as identified by the black line in (c).

representing different conceptualizations of the north boundary are based on the opinions of three hydrogeologists who have visited the Naturita site.

Three alternative surface complexation models were considered, and each model is applied to the entire simulation domain. As shown in Table 2, each SCM contains only one reaction and they are mainly different in the carbonate and proton stoichiometry of the adsorbed complex. The alternative models are simpler than the true model (Table 1) and model errors are potentially substantial. However, our approach attempts to balance errors that arise from both groundwater flow and geochemical reactions. Considering more complicated SCM models is possible and warranted in a future study.

A total of 27 alternative models were developed, and they are all simpler than the true model. These alternative models are a combination of the competing model propositions of the three uncertain model components. The models were named by the combination of their abbreviations of each proposition. For example, model Heter_bnd1_scm1 has heterogeneous hydraulic conductivity fields, boundary condition 1, and SCM1.

3.3. Model calibration and evaluation of model probability

For the 27 alternative models, the estimated parameters were hydraulic conductivities specific to different models (i.e., three K values corresponding to the three facies for the Heter model, five K values corresponding to the five zones for the Zone model, and three K values corresponding to the three layers for the Homo model), one dispersivity for all the models, and one formation constant ($\log K$) for each SCM. The parameters were calibrated sequentially in two steps. The first two types of parameters (K and dispersivity) were estimated from Cl concentration data in nonreactive transport modeling, and then the parameter $\log K$ of the SCM reactions was estimated based on U(VI) concentration data in reactive transport modeling. Porosity values were not estimated in the calibration process. For the Heter models, the true porosity values of the three facies were used. For the Zone and Homo models, the entire simulation domain has the same porosity value calculated as the power mean of the porosity values from the true model.

Fig. 5a shows the sum of squared weighted residuals (SSWR) of the 27 alternative models. The figure indicates that, overall,

Table 2
Alternative surface complexation reactions used in the alternative models. Each reaction is applied to the entire model domain.

Model	Reaction
SCM1	$S\_OH + UO_2^{2+} + 2CO_3^{2-} = S\_OHUO_2(CO_3)_2^{2-}$
SCM2	$S\_OH + UO_2^{2+} + CO_3^{2-} = S\_OUO_2CO_3^{-} + H^+$
SCM3	$S\_OH + UO_2^{2+} = S\_OUO_2^{+} + H^+$

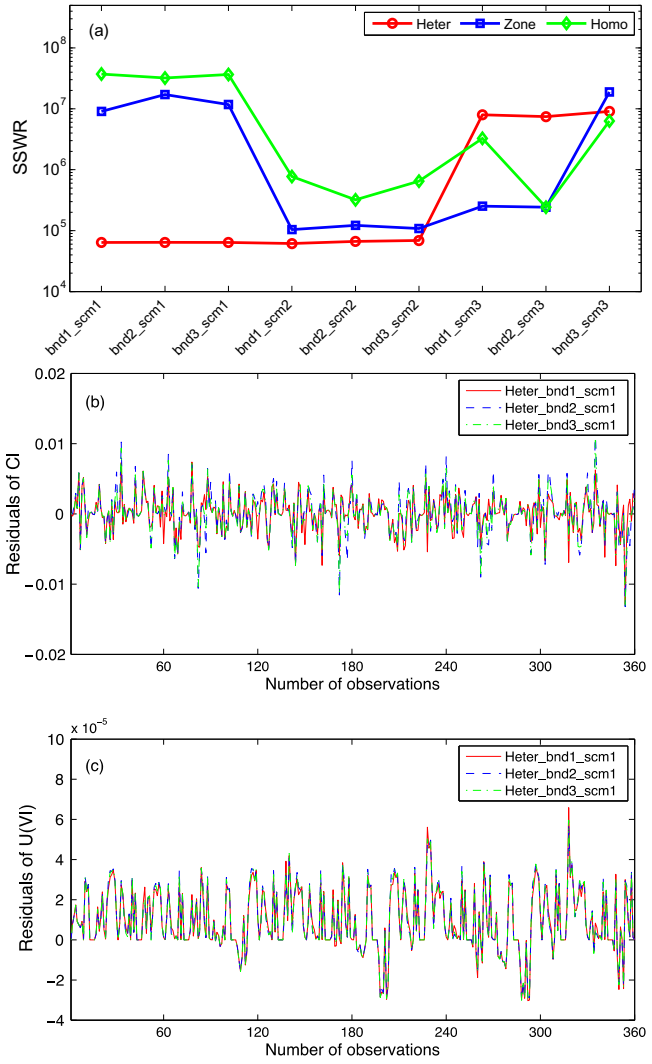

Fig. 5. (a) Sum of squared weighted residuals (SSWR) of all the 720 observations; (b) residuals of Cl, and (c) residuals of U(VI) for the three best models with the same Heter parameterization and SCM1.

heterogeneous models fit the calibration data better than the zone and homogeneous models. This is not surprising since the parameterization of the heterogeneous models is similar to that of the true model. However, not all of the heterogeneous models have better calibration performance than the zone and homogeneous models. For example, when the heterogeneous models were combined with SCM3, they even have worse fit than the simplest homogeneous models regardless of the model boundaries; the fits of Heter_bnd∗_scm3 models (∗ represents 1, 2, and 3 here) are also worse than the zone models when they were combined with Bnd1 and Bnd2. A possible reason is that SCM3 is the worst surface complexation model to simulate the U(VI) reactions without considering variable carbonate concentrations. When SCM3 was combined with the simple Zone and Homo models, the simple models may have more flexibility to compensate the model error in SCM3 as indicated by their physically unreasonable parameter estimates of K and dispersivity.

Fig. 5a also indicates that, when Heter was combined with SCM1 and SCM2, the good model fits are independent of the different boundaries. The Zone and Homo models give better calibration performance when they were combined with SCM2 than being combined with other SCMs and it was found again that the difference between the combinations with different boundaries is not very outstanding. This may imply that the boundary component is the most uncertain model component, because all conceptualizations perform comparably well and it is not possible to identify the better performing boundary condition without additional data, recalling that there were no calibration data near the extended boundary. In contrast, Heter and SCM2 are identified as the best model propositions. The specific model uncertainty contribution from each model component and its propositions is discussed in detail below through a model probability hierarchy.

Based on the calibration results, the calculated KIC values of the 27 models have the same performance as the SSWR (the plot was not shown here), i.e., the better fitting models with smaller SSWR values have smaller KIC values. By giving the models equal prior model probabilities (i.e., uniform prior), the KIC-based posterior model probabilities of the 27 alternative models are shown in Fig. 6 in a hierarchical way (Tsai and Elshall, 2013). The first level of uncertainty results from the parameterization of K fields; the second level results from the conceptualization of model boundaries; and the third level from the propositions of SCMs. Fig. 6 indicates that the best model is the combination of Heter, Bnd1, and SCM2 propositions; the resulting model has a probability of 66%. The only other models having significant model probability are Heter_bnd3_scm1, Heter_bnd1_scm1, and Heter_bnd2_scm1, with probabilities of 21.43%, 9.77%, and 2.8%, respectively, as summarized in Table 3. The branches starting from the Zone and Homo components all have zero probabilities and they were not extended to lower levels in Fig. 6.

Aggregation of model components in the hierarchical way provides a systematic representation of the competing propositions of different sources of model uncertainty, and allows recognizing better propositions. For example, the best model, Heter_bnd1_scm2, has the propositions of Heter, Bnd1, and SCM2; these propositions exhibit higher posterior model probabilities than other competing propositions, as exhibited in Fig. 6. The worst model Homo_bnd3_scm1, identified by the largest KIC value, does not have any propositions of the best model; the second worst model Homo_bnd1_scm1 has only the Bnd1 proposition of the best model. By analyzing the hierarchy of Fig. 6, the Heter component in level one has 100% model probability, because this parameterization is similar to that of the true model. The Bnd1 model in level two has 75.77% model probability, significantly higher than the 2.8% of Bnd2 and the 21.43% of Bnd3. This is expected for the following reasons: (1) Bnd1 is the true boundary configuration, (2) the northwest boundary of Bnd3 is close to the true boundary, and (3) Bnd2 deviates the most from the true boundary configuration. In level three, both SCM1 and SCM2 with carbonate reactions have significant model probabilities and SCM3 has zero probability, suggesting that carbonate plays an important role in the simulation of the U(VI) adsorption. This is consistent with the true SCM model (Table 1), whose reactions for the strong site for all the three facies have carbonate in the surface complex and relatively high values of formation constant ($\log K$).

Examination of the four best models with nonzero model probabilities listed in Table 3 shows that they all have the Heter component. It indicates that, in this study, the Heter parameterization outperforms the other two parameterizations. This is probably due to the similarity between the heterogeneous field of facies of the Heter parameterization and that of the true model. For a heterogeneous field of facies that is dramatically different from that of the true model, its performance may be worse than the Zone and Homo parameterizations. The four best models consider all the three boundary configurations, suggesting that the boundary configurations are not influential to the simulations of current U(VI) concentration data. As shown in Fig. 5b and c, the residuals of

Table 3
Five different sets of prior model probabilities and their associated posterior model probabilities for the four best models.

	Heter_bnd1 _scm2 (M1)	Heter_bnd3 _scm1 (M2)	Heter_bnd1 _scm1 (M3)	Heter_bnd2 _scm1 (M4)
Uniform prior	1/27	1/27	1/27	1/27
Posterior	66.00%	21.43%	9.77%	2.80%
Prior 1-1	1/2	1/2	–	–
Posterior	75.49%	24.51%	–	–
Prior 1-2	1/2	–	–	1/2
Posterior	95.93%	–	–	4.07%
Prior 2-1	1/2	1/6	1/6	1/6
Posterior	85.34%	9.24%	4.21%	1.21%
Prior 2-2	1/6	1/3	1/6	1/3
Posterior	53.13%	34.50%	7.87%	4.51%

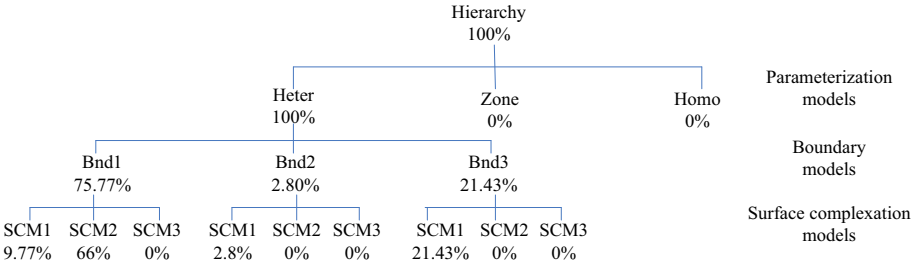

Fig. 6. Posterior model probabilities for the three uncertain model components with uniform prior model probabilities. The hierarchies under Zone and Homo models all have zero probabilities and are not shown in the figure.

the three best models with the same Heter and SCM1 components but different boundaries have similar magnitudes. This suggests that in this study only heterogeneous models should be pursued. If predictions are made at locations where boundary influence is significant, more effort should be spent to reduce model uncertainty in boundary conditions such as collecting more data that are sensitive to the boundary configurations. At last, three of the best models have the same model proposition of SCM1, implying that the three models may give similar U(VI) predictions.

3.4. Examine the validity of using MLBMA

The model probabilities calculated above are based on KIC that was derived using the Laplace approximation by assuming that the likelihood function is highly peaked at the calibrated parameter values. This is usually the case for large calibration datasets. For example, Kass and Raftery (1995) argued that, when the number of calibration data is greater than 20 times of the number of calibrated parameters, KIC is accurate for calculating model probabilities. Fig. 7 plots as an example the likelihood functions of the four best models with respect to the reaction parameter of log$K$. The figures indicate that the likelihood functions are highly peaked about their maxima, and decline fast as the values move away from the maxima. This suggests that KIC can be accurately used for our reactive transport problem.

Although the observed Fisher information matrix was not evaluated due to the computational cost, it is likely that the expected Fisher information matrix is an accurate approximation, due to relatively low nonlinearity of the models. We used the Linssen's modified version of Beales first measure (Seber and Wild, 2003; Hill and Tiedeman, 2007) to examine the nonlinearity of the reactive transport models. Beales measure tests the model linearity with respect to the parameter values. It evaluates the difference between the model-computed and the linearized estimates of the simulated values for the parameters generated on the edge of the 95% linear confidence region of the parameters. Fig. 8d shows that the values of Beales measure for the four best models are significantly below the threshold of linearity, indicating that the four models are effectively linear. Since the Beales measure may suffer from the problem of underestimating the model nonlinearity (Seber and Wild, 2003, p157), the issue of linearity is examined by plotting in Fig. 8 the relation between the model simulations and model parameter for the three best models. The figure shows that the models are essentially linear, confirming the conclusion based on the Beales measure. Because of the model linearity, KIC should be an accurate approximation of the integrated likelihood, and using the expected FIM to replace the observed FIM should give an accurate evaluation of KIC. It should be noted that not all groundwater reactive transport models are highly nonlinear, depending on geochemical conditions, nonlinearity of geochemical reactions, parameter combination in the reactions, and interactions between the processes of flow, transport, and geochemical reactions.

4. Assessment of predictive performance

The 27 calibrated models were used to predict the U(VI) concentrations for up to 200 years without the contaminant recharge from the two ponds sources. The assessment of predictive performance was conducted for the results of the best four models and MLBMA. For the convenience of discussion, the previous model

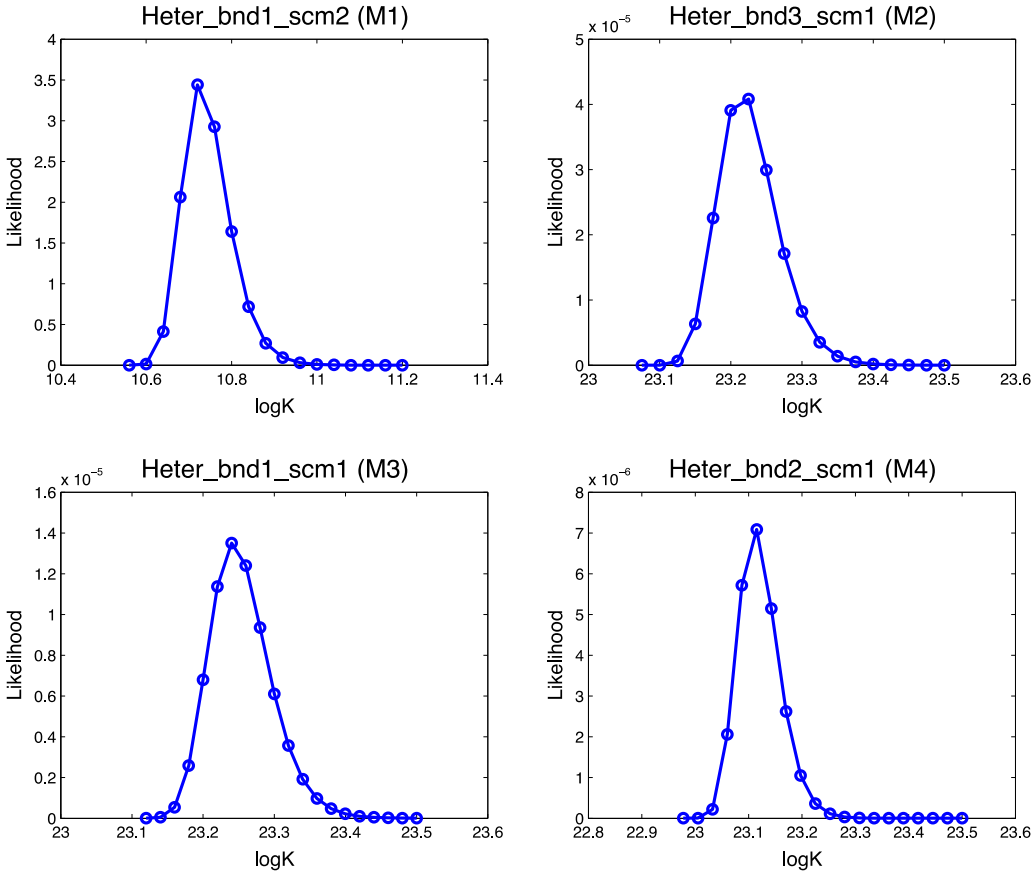

Fig. 7. Likelihood functions of the four best models with respect to parameter log$K$.

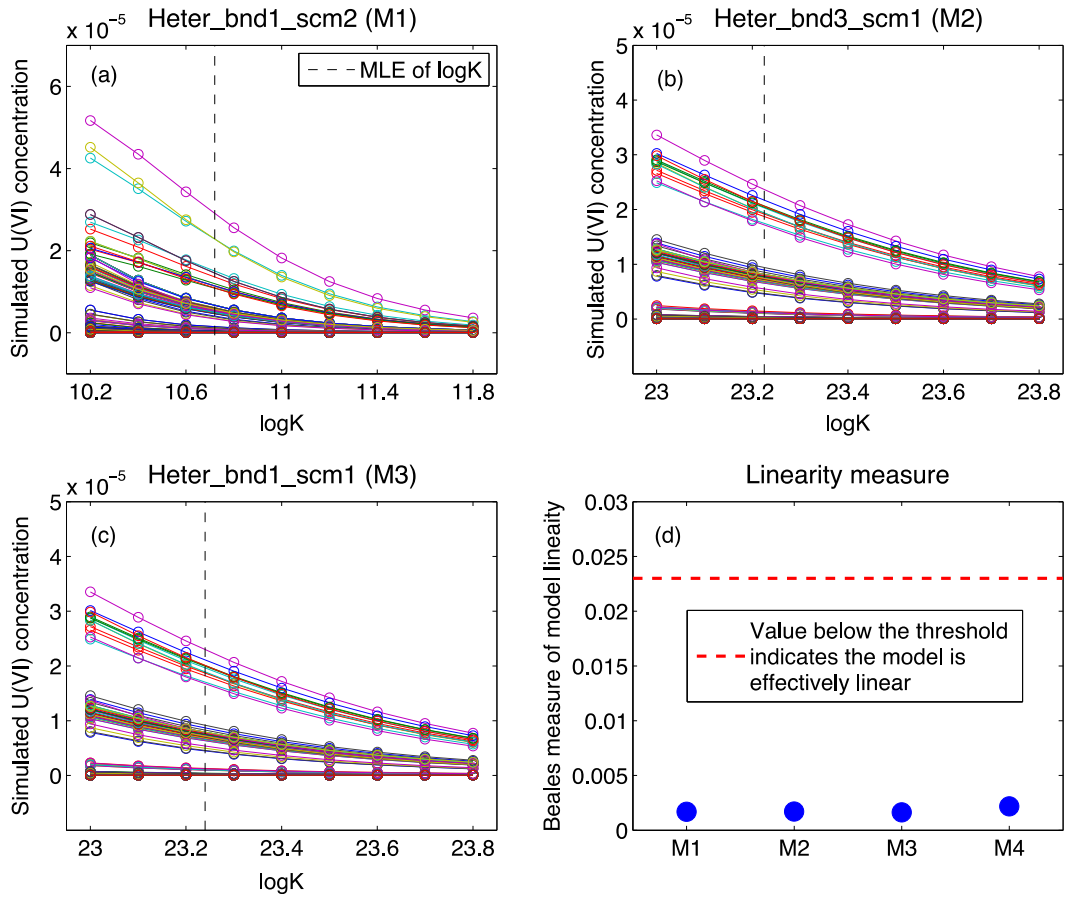

Fig. 8. Variation of simulated U(VI) concentration with parameter log$K$ for the three best models: (a) M1, (b) M2, and (c) M3 at observation locations. Beales measure of model linearity for the four best models is shown in (d).

notations are not used, and the four models are denoted as M1–M4 according to their model probabilities given in Fig. 6 from the largest to the smallest. We compared their performance for two prediction quantities: the U(VI) concentration in the top layer at all the cells of groundwater discharge zone (i.e., the gray area above the green area shown in Fig. 1), and the U(VI) concentration in four monitoring wells located in the top layer of the groundwater discharge zone (Fig. 9). The area of groundwater discharge zone is of special interest, because after 200 years this area contributes the highest U(VI) concentration in the entire simulation domain due to the joint influence of groundwater flow, solute transport, and geochemical reactions.

The predictive performance was assessed using two measures, predictive bias and predictive logscore (Good, 1952; Volinsky et al., 1997; Hoeting et al., 1999). The predictive bias, as a measure of predictive accuracy, is the difference between the simulations of the four alternative models and MLBMA and that of the true model. For evaluating the predictions of the entire groundwater discharge zone, the root mean squared error (RMSE) was calculated, because the predictive bias is a measure of point prediction and generally not good for the entire modeling domain especially when the domain areas are different like in this study with three different boundary configurations.

The predictive logscore of an individual model M_k is defined as the negative logarithm of the predictive density, i.e., (Hoeting et al., 1999; Ye et al., 2004)

$$-\ln p(\hat{\mathbf{y}}_k|M_k,\mathbf{D}) = -\sum_{\hat{y}_k\in\hat{\mathbf{y}}_k} \ln p(\hat{y}_k|M_k,\mathbf{D}) \tag{10}$$

where $\hat{\mathbf{y}}_k$ is the vector of predictions based on model M_k and $\mathbf{D}$ is the data for model calibration. The predictive logscore of model averaging is defined as (Hoeting et al., 1999; Ye et al., 2004)

$$-\ln p(\hat{\mathbf{y}}_k|\mathbf{D}) = -\sum_{\hat{y}_k\in\hat{\mathbf{y}}_k} \ln\left[\sum_{k=1}^{K} p(\hat{y}_k|M_k,\mathbf{D})p(M_k|\mathbf{D})\right]. \tag{11}$$

Based on Eqs. (10) and (11), the lower predictive logscore of model M_k or MLBMA indicates higher probability that M_k or MLBMA can predict $\mathbf{y}$ based on observation $\mathbf{D}$. Therefore, smaller logscore corresponds to better predictive performance. Logscore is negative if the density, p, is larger than one. To save computational time, we assume the probability density functions in Eqs. (10) and (11) are Gaussian and this assumption seems to be valid given that the models are effectively linear as discussed in Section 3.4. For the individual models, the mean and variance were evaluated using UCODE_2005; for MLBMA, the posterior mean and variance were estimated using Eqs. (4) and (5). The predictive logscore considers both the predictive bias and predictive uncertainty, and is better than the measure of predictive bias. For example, a model with small predictive bias and small predictive variance that does not enclose the true value may have a larger logscore than another model with relatively large bias but also relatively large variance that covers the true value. The discussion below presents the logscore averaged over the area of the groundwater discharge zone and at the four monitoring wells with different results of model predictions.

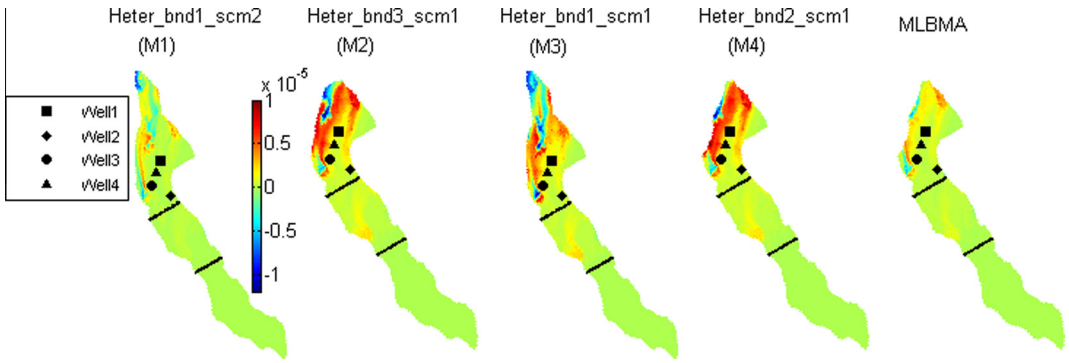

Fig. 9. Predictive bias of U(VI) concentration (M) in layer one of the four best models and by MLBMA. Well1 to Well4 are four monitoring wells for evaluation of predictive performance. The natural recharge zone is enclosed by the two black lines.

4.1. Prediction of U(VI) on an entire area

Fig. 9 shows the predictive bias of the simulated U(VI) fields in layer one after 200 years for the best four models and MLBMA. The area with the large difference for all the individual models and MLBMA is located at the groundwater discharge zone, close to the north boundary and downgradient from the natural recharge zone delineated by the two black lines in Fig. 9. Comparing the predictive bias of the four individual models and MLBMA, the figure indicates that the best-fitting model, M1, has the smallest overall predictive bias with the highest prediction accuracy. The prediction accuracy of MLBMA is between the best model and the other three models because the MLBMA mean prediction is the weighted average of the means of the four models. This finding is confirmed by their RMSE listed in Table 4. The table shows that the RMSE of predicted U(VI) concentration based on the best model is the smallest with the value of 2.27×10^{-8} M and is about one third of the RMSE of M4 that was the worst-predicting model. The MLBMA has a RMSE value of 3.36×10^{-8} M, a little larger than, but close to, that of the best model because the best model has a large model probability (66%).

However, as measured by the logscore values shown in Table 4 MLBMA gives the best predictive performance. These logscore values for the four models and MLBMA are averaged over the area of the groundwater discharge zone. The table indicates that MLBMA has the smallest logscore of −6.44 which is much smaller than those of all the individual models. This suggests that MLBMA can predict the integrated true value for the discharge zone with larger probability than any individual models.

4.2. Prediction of U(VI) at four monitoring wells

The locations of the four monitoring wells are shown in Fig. 9. The predictive bias and predictive logscore of the four individual models and MLBMA at the four wells are listed in Table 5. Table 5 indicates that the best model, M1, has smaller predictive bias than MLBMA for three of the four wells, but has larger predictive logscore than MLBMA except for Well 1. The reasons for the different predictive performance are illustrated in Fig. 10. The plots in the left column of Fig. 10 illustrate the predictive bias of the four individual models and MLBMA for the four wells. The plots in the right column are the predictive logscore of M1 and MLBMA.

For Well 1, M1 gives much better prediction than the three inferior models (Fig. 10a), making its predictive bias smaller than the MLBMA. In addition, M1 gives reasonably large uncertainty bounds which bound the true value; these two factors jointly make M1

Table 4
RMSE and logscore of the four best models and MLBMA averaged over the area of the groundwater discharge zone for predicted U(VI) concentration (M) after 200 years.

	M1	M2	M3	M4	MLBMA
RMSE	2.27×10^{-2}	6.36×10^{-2}	4.58×10^{-2}	6.69×10^{-2}	3.36×10^{-2}
Logscore	8.58	19.76	11.20	19.06	−6.44

Table 5
Predictive performance of the four best individual models and MLBMA for the four evaluated wells. The comparison is focused on the best model M1 and MLBMA results as highlighted in gray and the one giving better predictive performance has bold values.

Performance measure	M1	M2	M3	M4	MLBMA
Well 1					
Bias	$\mathbf{-6.07 \times 10^{-8}}$	9.17×10^{-7}	6.91×10^{-7}	6.86×10^{-7}	2.43×10^{-7}
Logscore	**−15.1**	−5.88	−10.3	−10.2	−13.5
Well 2					
Bias	-1.15×10^{-7}	2.19×10^{-7}	6.78×10^{-8}	3.06×10^{-7}	$\mathbf{-1.41 \times 10^{-8}}$
Logscore	−14.0	−12.5	−14.3	−9.86	**−14.7**
Well 3					
Bias	$\mathbf{1.99 \times 10^{-7}}$	1.14×10^{-6}	8.23×10^{-7}	1.33×10^{-6}	4.94×10^{-7}
Logscore	−3.43	−1.84	−5.71	−0.39	**−13.1**
Well 4					
Bias	$\mathbf{-1.18 \times 10^{-7}}$	6.62×10^{-7}	5.78×10^{-7}	5.37×10^{-7}	1.36×10^{-7}
Logscore	−13.7	−8.65	−8.77	−11.1	**−13.8**

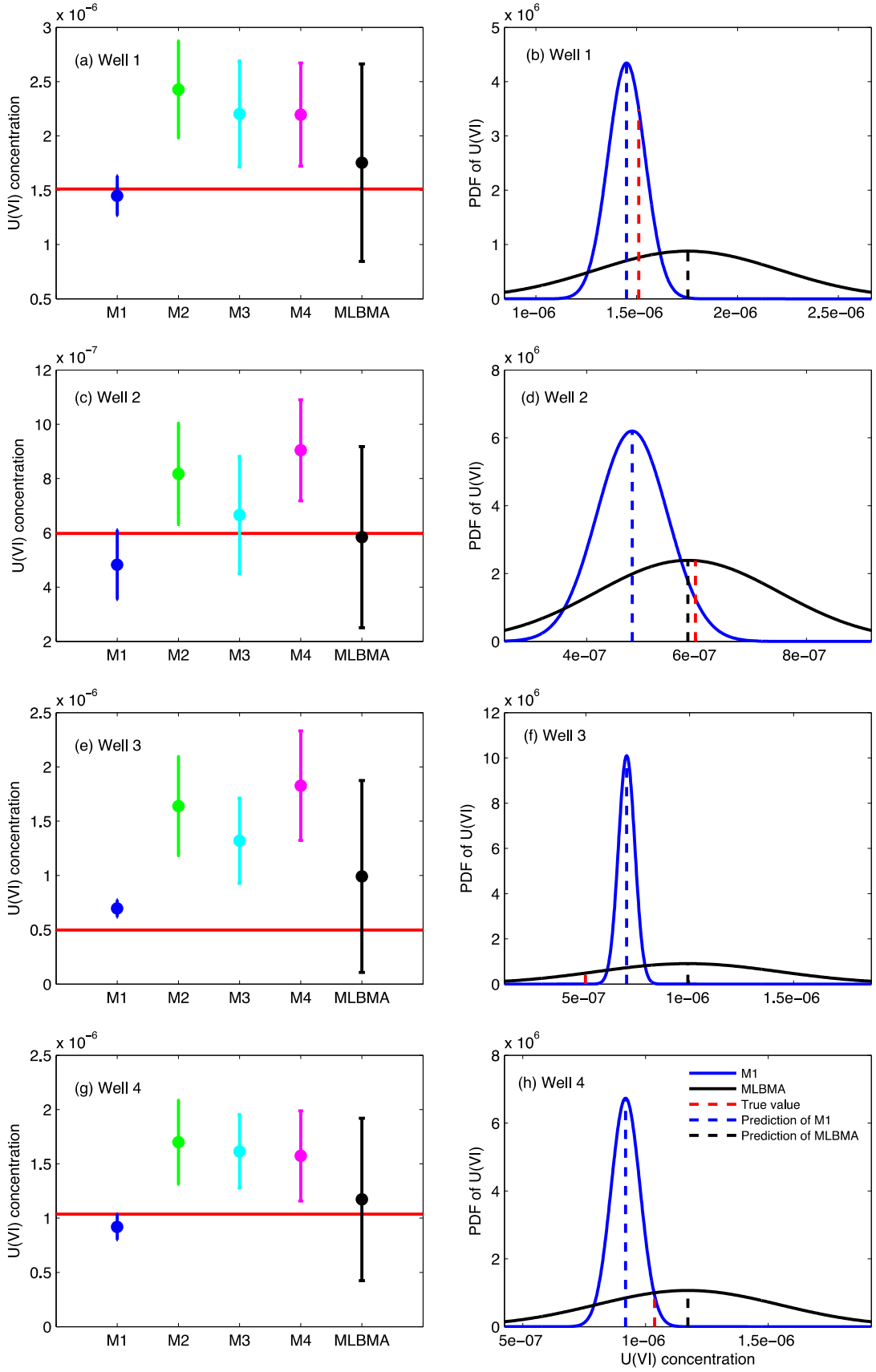

Fig. 10. Left column figures: predicted U(VI) concentrations (M) and their 95% linear confidence intervals of the four best models and MLBMA for the four monitoring wells; the horizontal red lines represent true values of the predictions. Right column figures: Probability density functions (PDF) of U(VI) based on model M1 and MLBMA at the four wells. The negative log of the PDF of the true value (red dashed line) measures the logscore. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Predictive performance of the best model (M1) and MLBMA with five sets of prior model probability for the four evaluated wells. For each well, the values of the best bias and logscore are in bold.

Performance measure	M1	MLBMA				
		Uniform prior	Prior 1-1	Prior 1-2	Prior 2-1	Prior 2-2
Well 1						
Bias	-6.07×10^{-8}	2.43×10^{-7}	1.79×10^{-7}	$\mathbf{-3.03 \times 10^{-8}}$	7.04×10^{-8}	3.70×10^{-7}
Logscore	**−15.1**	−13.5	−13.6	−14.6	−13.9	−13.3
Well 2						
Bias	-1.15×10^{-7}	-1.41×10^{-8}	-1.35×10^{-8}	-9.83×10^{-8}	$\mathbf{-1.17 \times 10^{-8}}$	3.32×10^{-8}
Logscore	−14.0	−14.7	−14.7	−14.7	**−14.8**	−14.6
Well 3						
Bias	$\mathbf{1.99 \times 10^{-7}}$	4.94×10^{-7}	4.30×10^{-7}	2.45×10^{-7}	3.26×10^{-7}	6.24×10^{-7}
Logscore	−3.43	−13.1	−13.2	**−13.8**	−13.5	−12.8
Well 4						
Bias	-1.18×10^{-7}	1.36×10^{-7}	7.33×10^{-8}	-9.12×10^{-8}	$\mathbf{-8.51 \times 10^{-9}}$	2.35×10^{-7}
Logscore	−13.7	−13.8	−14.0	**−14.6**	−14.2	−13.6

predict the true value with larger probability than MLBMA as illustrated in Fig. 10b and by the smaller logscore value as shown in Table 5. In summary, for Well 1, the best model gives better predictive performance than MLBMA according to the two measures.

For Well 2, the alternative models over- or under-predict the true value as shown in Fig. 10c, and M3 actually gives more accurate prediction than the best model M1 (Table 5). By averaging, MLBMA gives better predictive performance than M1 with smaller predictive bias and smaller predictive logscore as well, as shown in Fig. 10d. At this location, the model performing best in calibration did not perform best for prediction. It is therefore better to use MLBMA to consider all plausible models rather than using the best model selected after model calibration.

For Well 3, as shown in Fig. 10e, all models over-predict the true value and MLBMA has larger predictive bias than the best-predicting and also the best-fitting model M1. It is however noted that the 95% confidence interval of M1 is too narrow to cover the true value. As a result, the probability of M1 to predict the true value is negligible, as shown in Fig. 10f, and the predictive logscore of M1 is much larger than MLBMA as shown in Table 5. At this location, MLBMA gives better predictive performance in terms of predictive logscore but worse performance in terms of predictive bias. To improve the predictive accuracy of MLBMA, additional plausible models with different prediction abilities than the four existing ones need to be developed.

For Well 4, as shown in Fig. 10g and h and Table 5, MLBMA gives slightly larger predictive bias but smaller predictive logscore than M1. While the predictive performance for Well 4 is similar to that for Well 3, the reasons are different. As shown in Fig. 10g, the three inferior models have very similar predictions, but they were treated as separate equally likely models and equal prior model probabilities were assigned to them. This is tantamount to giving a triple weight to three slightly different versions of the same model, which weakens the predictive power of MLBMA. Specifically, the prediction accuracy of MLBMA is deteriorated by the three inferior models because of the high prior probabilities, and the predictive uncertainty of MLBMA is overestimated by treating the three correlated models as mutually exclusive ones. The three models are correlated due to the following reasons: (1) they have the same SCM1, (2) they have the same K field parameterization, and (3) the boundary conditions are not influential to the calibration and prediction data.

As suggested by Neuman (2003), the effect of correlated models on posterior model probabilities of MLBMA can be reduced by adjusting the prior model probability in two ways. The first one is to retain the structurally distinct models and to assign them equal prior model probability. We denote this strategy as Prior 1 (Tables 3 and 6). Among the four models listed in Table 3, model Heter_bnd1_scm2 is structurally distinct to the other three models that have the same reaction component SCM1. Among the three models with SCM1, model Heter_bnd1_scm1 is excluded, because it shares the boundary configuration of Bnd1 with model Heter_bnd1_scm2. For models Heter_bnd3_scm1 and Heter_bnd2_scm1, either of them is structurally distinct with Heter_bnd1_scm2, and the two cases are denoted as Priors 1-1 and 1-2, respectively, in Tables 3 and 6. For the two cases, only the two retained models were left, each having the prior probability of 1/2. This redistribution of prior probabilities brings an increase in the posterior probability of the best model M1 (Table 3). More importantly, it slightly improves predictive performance of MLBMA compared to the case of uniform prior at all the four wells with respect to both measures in that the values of predictive bias and logscore in Priors 1-1 and 1-2 are smaller than those of the uniform prior, as shown in Table 6. In addition, for Wells 2 and 4, the predictive bias of MLBMA is smaller than that of M1 (the best model) for the two cases of Priors 1-1 and 1-2. This is illustrated in Fig. 11 for Well 4 as an example. The figure shows that, when the posterior probability of M1 increases substantially in Prior 1-2, the posterior mean becomes closer to that of model M1 and smaller predictive bias is resulted.

The other way to address the correlated models is to keep the models but to reduce their prior model probabilities. Following Ye et al. (2004), we grouped the correlated models together, and the models in the same group were viewed equally likely. We denoted this strategy as Prior 2. As shown in Table 6, the four models are divided into two groups in two cases, denoted as Priors 2-1 and 2-2, based on different grouping criteria. In Prior 2-1, the grouping is based on predictive performance of the models. Since model Heter_bnd1_scm2 has different performance than the other three models (Fig. 10), it is included in the first group, and a prior probability of 1/2 is assigned to this model. Since the other three models (Heter_bnd3_scm1, Heter_bnd1_scm1, and Heter_bnd2_scm1) have similar predictive performance, they are included in the other group, and each of the models receives a prior probability of 1/6 (1/2 divided by 3). In Prior 2-2, the grouping criterion is different and based on the boundary configuration. The first group has models Heter_bnd1_scm2 and Heter_bnd1_scm2 that have the boundary configuration of bnd1. For the other two models, since they have different boundary configurations, they are included into two groups. Table 3 shows that, while the choice of priors in Prior 2-1 increases the posterior probability of M1, the choice of priors in Prior 2-2 decreases the posterior probability of M1. As shown in Table 6, Prior 2-1 improves the predictive performance of MLBMA, while Prior 2-2 decreases the predictive

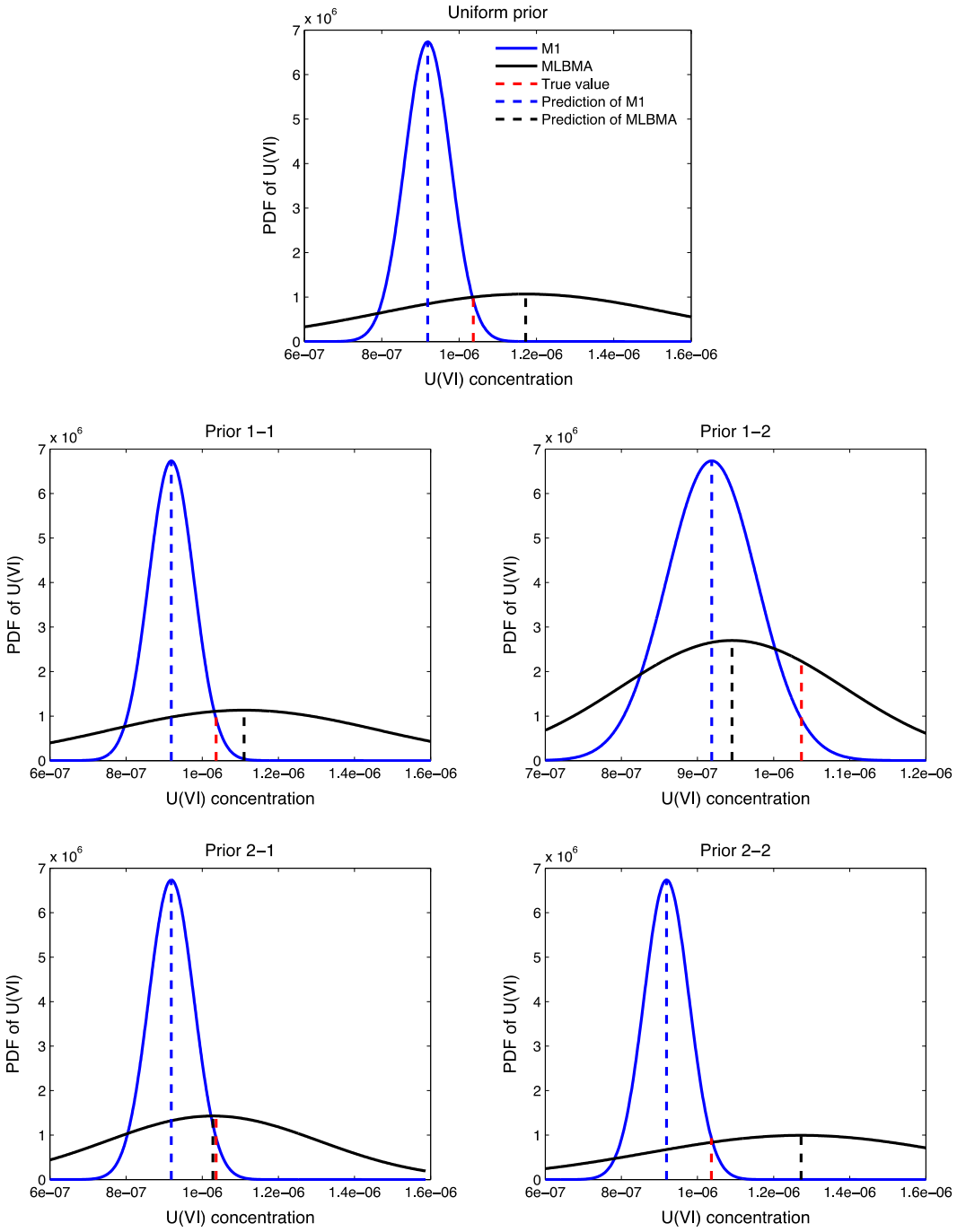

Fig. 11. Probability density functions (PDF) of U(VI) at Well 4 based on the best model M1 and MLBMA calculated with five different sets of prior model probabilities listed in Table 3. The negative log of the PDF of the true value (red dashed line) is the logscore. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

performance. This is illustrated in Fig. 11 for Well 4 as an example. The figure indicates that, using Prior 2-1, the mean of MLBMA becomes closer to the true value (i.e., a smaller predictive bias), which results in a higher probability prediction of the true value (i.e., a smaller predictive logscore). When using Prior 2-2, the MLBMA mean becomes significantly different from the true value, and predictive logscore becomes larger. This example suggests that grouping correlated models based on their predictive performance is better than grouping the models based on their structural difference for improving MLBMA predictive performance.

5. Conclusions and discussion

This work, for the first time, applied MLBMA to groundwater reactive transport modeling and assessed the predictive performance of MLBMA using 27 synthetic models of groundwater reactive transport designed using the data and information of the Naturita site. The 27 alternative models were a combination of three uncertain model components, each of which had three competing propositions. The three uncertain model components were parameterization of hydraulic conductivity (which had

heterogeneous, zonal, and homogeneous competing propositions), conceptualization of domain boundary which simulated the Naturita site with three different domains, and the surface complexation models which had three different reactions to simulate the U(VI) adsorption. The developed models represent three different sources of model uncertainty for assessing predictive performance of MLBMA in groundwater reactive transport modeling. The models were based on the synthetic observations whose locations were from existing observation wells in Naturita site and therefore the simulation and evaluation results presented in this study are expected to provide guidance for the research at the Naturita site.

We developed the alternative models and analyzed their probabilities in a hierarchical way by classifying the three model components into three levels. This aggregation of uncertain model components provides a systematic representation of the competing propositions of different sources of model uncertainty, and allows for recognizing the superior propositions. The results indicated that (1) heterogeneous parameterization was absolutely the best parameterization; (2) SCM1 and SCM2 with carbonate in the reactions were better than SCM3 that does not include carbonate in the U(VI) surface complex; and (3) each of the three tested boundary configuration was represented in the set of best models because the current calibration data were insufficient to discard any boundary configuration.

By comparing the predictive performance of MLBMA with that of the best model, the predictive logscore results showed that, in most cases, the predictive performance of MLBMA is better than the best model, suggesting that using MLBMA is a sound strategy to achieve a more robust assessment of predictive performance than using a single model. However, with regards to the measure of predictive bias, this study showed that MLBMA produced the larger predictive bias than the best model in most cases. For MLBMA to work the best, it is important that the alternative models are structurally distinct and have diverse model predictions. This conclusion is in line with the findings of Winter and Nychka (2010). If the model simulations are similar, it is possible that the models are correlated, e.g., the three alternative models in this study all having SCM1 to simulate uranium adsorption. This problem is empirically addressed in this study by adjusting the prior model probability to either eliminate certain correlated models or assign smaller prior probabilities to correlated models; the latter technique was shown to be better here. Using the new prior model probabilities dramatically improves the predictive performance of MLBMA in that MLBMA can predict the true value with higher probability with negligible bias. More research is warranted to resolve the issue of model correlation for model averaging analysis. In the future study, we will apply MLBMA to a real-world problem where the true model is unknown and the alternative models are calibrated against the real observation data.

This synthetic study is subject to several limitations. First of all, the high similarity between the heterogeneous facies of the Heter models and the true heterogeneous facies may not be practically realistic. If there is no sufficient data for site characterization, more advanced methods of model calibration (e.g., using level set methods) is needed to achieve the level of similarity shown in this study. It would be more interesting from a practical perspective to consider different facies fields to investigate to what extent the uncertainty in parameterization can be reduced. It is likely that using a wrong heterogeneous field may be worse than using a homogeneous or zonal field for numerical simulation. To test this hypothesis however requires using more advanced methods of model calibration to find appropriate spatial distributions of hydrofacies fields, which entails a large number of model executions. It is beyond the scope of this study and not pursued in this study. Another limitation of this study is that the low nonlinearity of the reaction models is not ideal to investigate whether MLBMA can be applied to highly nonlinear problems of groundwater reactive transport modeling. Although Lu et al. (2013) applied MLBMA to more complicated and nonlinear surface complexation models, their flow and transport models are one-dimensional only, and the interactions between flow models, transport models, and chemical reaction models are relatively low. More nonlinear reactive transport models will be developed in a future study to investigate the applicability of MLBMA to these models.

Acknowledgments

This work was supported in part by DOE-SBR Grants DE-SC0003681, DE-SC0002687 and DE-SC0000801, DOE Early Career Award, DE-SC0008272, NSF-EAR Grant 0911074, and National Natural Science Foundation of China Grants, 51328902. The first author performed part of the work when she was employed by the U.S. Geological Survey. We thank Alberto Guadagnini, Chris Green, and an anonymous reviewer for their helpful comments.

References

Bishop, C.H., Abramowitz, G., 2013. Climate model dependence and the replicate earth paradigm. Clim. Dyn. 41, 885–900. http://dx.doi.org/10.1007/s00382-012-1610-y.

Carle, S.F., 1999. T-PROGS: Transition probability geostatistical software. Version 2.1, University of California, Davis.

Cavadias, G., Morin, G., 1986. The combination of simulated discharges of hydrological models. Application to the WNO intercomparison of conceptual models of snowmelt runoff. Nordic. Hydrol. 17 (1), 21–32.

Chitsazan, N., Tsai, F.T.-C., 2014. A hierarchical Bayesian model averaging framework for groundwater prediction under uncertainty. Groundwater. http://dx.doi.org/10.1111/gwat.12207.

Curtis, G.P., Fox, P., Kohler, M., Davis, J.A., 2004. Comparison of in situ uranium K_D values with a laboratory determined surface complexation model. Appl. Geochem. 19 (10), 1643–1653.

Curtis, G.P., Davis, J.A., Naftz, D.L., 2006. Simulation of reactive transport of uranium(VI) in groundwater with variable chemical conditions. Water Resour. Res. 42 (4). http://dx.doi.org/10.1029/2005WR003979.

Curtis, G.P., Kohler, M., Davis, J.A., 2009. Comparing approaches for simulating the reactive transport of U(VI) in ground water. Mine Water Environ. 28, 84–93.

Davis, J.A., Meece, D.E., Kohler, M., Curtis, G.P., 2004. Approaches to surface complexation modeling of uranium(VI) adsorption on aquifer sediments. Geochim. Cosmochim. Ac. 68 (18), 3621–3641. http://dx.doi.org/10.1016/j.gca.2004.03.003.

Deutsch, C.V., Journel, A.G., 1998. GSLIB, Geostatistical Software Library and User's Guide, Second ed. Oxford University Press, New York.

Dong, L., Xiong, L., Yu, K., 2013. Uncertainty analysis of multiple hydrologic models using the Bayesian model averaging method. J. Appl. Math. 2013. http://dx.doi.org/10.1155/2013/346045 (article no 346045, 11 pages).

Draper, D., 1995. Assessment and propagation of model uncertainty. J. R. Stat. Soc. Ser. B 57 (1), 45–97.

Draper, D., 1999. Comment to "Bayesian model averaging: a tutorial". Stat. Soc. 14 (4), 405–409.

Duan, Q., Ajami, N.K., Gao, Z., Sorooshian, S., 2007. Multimodel ensemble hydrologic prediction using Bayesian model averaging. Adv. Water Resour. 30, 1371–1386.

Elshall, A.S., Tsai, F.T.-C., 2014. Constructive epistemic modeling of groundwater flow with geological structure and boundary condition uncertainty under the Bayesian paradigm. J. Hydrol. 517, 105–119.

Foglia, L., Mehl, S.W., Hill, M.C., Perona, P., Burlando, P., 2007. Testing alternative ground water models using cross validation and other methods. Ground Water 45 (5), 627–641.

Foglia, L., Mehl, S.W., Hill, M.C., Burlando, P., 2013. Evaluating model structure adequacy: the case of the Maggia Valley groundwater system, southern Switzerland. Water Resour. Res. 49. http://dx.doi.org/10.1029/2011WR011779.

George, E.I., 1999. Comment. Stat. Sci. 14 (4), 409–412.

Good, I.J., 1952. Rational decisions. J. R. Stat. Soc., Ser. B 57 (1), 107–114.

Harbaugh, A.W., 2005. MODFLOW-2005, The U.S. Geological Survey modular groundwater model – the groundwater flow process. U.S. Geol. Surv. Tech. Methods, A6–A16.

Hill, M.C., Tiedeman, C., 2007. Effective Calibration of Ground Water Models with Analysis of Data, Sensitivities, Predictions, and Uncertainty. John Wiely, New York.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14 (4), 382–417.

Hyun, S.P., Fox, P.M., Davis, J.A., Campbell, K.M., Hayes, K.F., Long, P.E., 2009. Surface complexation modeling of U(VI) adsorption by aquifer sediments from a former

mill tailings site at Rifle, Colorado. Environ. Sci. Technol. 43, 9368–9373. http://dx.doi.org/10.1021/es902164n.
Kashyap, R.L., 1982. Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Trans. Pattern Anal. Mach. Intell. 4 (2), 99–104.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90 (430), 773–795.
Kitanidis, P.K., Lane, R.W., 1985. Maximum likelihood parameter estimation of hydrologic spatial processes by the Gaussian–Newton method. J. Hydrol. 79, 53–71.
Kohler, M., Curtis, G.P., Meece, D.E., Davis, J.A., 2004. Methods for estimating adsorbed uranium (VI) and distribution coefficients of contaminated sediments. Environ. Sci. Technol. 38, 240–247.
Lu, D., Ye, M., Neuman, S.P., 2011. Dependence of Bayesian model selection criteria and Fisher information matrix on sample size. Math. Geosci. http://dx.doi.org/10.1007/s11004-011-9359-0.
Lu, D., Ye, M., Neuman, S.P., Xue, L., 2012. Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs. Adv. Water Res. 35, 69–82. http://dx.doi.org/10.1016/j.advwatres.2011.10.007.
Lu, D., Ye, M., Meyer, P.D., Curtis, G.P., Shi, X., Niu, X.-F., Yabusaki, S.B., 2013. Effects of error covariance structure on estimation of model averaging weights and predictive performance. Water Resour. Res. 49. http://dx.doi.org/10.1002/wrcr.20441.
Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models. Stoch. Environ. Res. Risk Assess. 17 (5), 291–305. http://dx.doi.org/10.1007/s00477-003-0151-7.
Neuman, S.P., Xue, L., Ye, M., Lu, D., 2012. Bayesian analysis of data-worth considering model and parameter uncertainties. Adv. Water Res. 36, 75–85. http://dx.doi.org/10.1016/j.advwatres.2011.02.007.
Poeter, E.P., Anderson, D.A., 2005. Multimodel ranking and inference in ground water modeling. Ground Water 43 (4), 597–605.
Poeter, E.P., Hill, M.C., Banta, E.R., Mehl, S.W., Christensen, S., 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.
Prommer, H., 2006. A reactive multicomponent transport model for saturated porous media, User's manual. Version 1.46, <http://www.pht3d.org>.
Riva, M., Panzeri, M., Guadagnini, A., Neuman, S.P., 2011. Role of model selection criteria in geostatistical inverse estimation of statistical data- and model-parameters. Water Resour. Res. 47, W07502. http://dx.doi.org/10.1029/2011WR010480.
Rojas, R., Feyen, L., Dassargues, A., 2008. Conceptual model uncertainty in groundwater modeling: combining generalized likelihood uncertainty estimation and Bayesian model averaging. Water Resour. Res. 44, W12418. http://dx.doi.org/10.1029/2008WR006908.
Rojas, R., Feyen, L., Dassargues, A., 2009. Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modeling. Hydrol. Process. 23 (8), 1131–1146.
Sain, S.R., Furrer, R., 2010. Combining climate model output via model correlations. Stoch. Env. Res. Risk Assess. 24 (6), 821–829. http://dx.doi.org/10.1007/s00477-010-0380-5.
Sassen, D.S., Hubbard, S.S., Bea, S.A., Chen, J., Spycher, N., Denham, M.E., 2012. Reactive facies: an approach for parameterizing field-scale reactive transport models using geophysical methods. Water Resour. Res. 48, W10526. http://dx.doi.org/10.1029/2011WR011047.
Seber, G.A.F., Wild, C.J., 2003. Nonlinear Regression. John Wiley & Sons, Hoboken, New Jersey.
Singh, A., Mishra, S., Rushauff, G., 2010a. Model averaging techniques for quantifying conceptual model uncertainty. Ground Water 48 (5), 701–715.
Singh, A., Walker, D.D., Minsker, B.S., Valocchi, A.J., 2010b. Incorporating subjective and stochastic uncertainty in an interactive multi-objective groundwater calibration framework. Stoch. Environ. Res. Risk Assess. 24, 881–898.
Steefel, C.I., DePaolo, S.J., Lichtner, P.C., 2005. Reactive transport modeling: an essential tool and a new research approach for the earth sciences. Earth Planet. Sci. Lett. 240, 539–558.
Troldborg, L., Refsgaard, J.C., Jensen, K.H., Engesgaard, P., 2007. The importance of alternative conceptual models for simulation of concentrations in multi-aquifer system. Hydrogeol. J. 15 (5), 843–860.
Tsai, F.T.-C., Elshall, A.S., 2013. Hierarchical Bayesian model averaging for hydrostratigraphic modeling: uncertainty segregation and comparative evaluation. Water Resour. Res. 49, 5520–5536. http://dx.doi.org/10.1002/wrcr.20428.
Volinsky, C.T., Madigan, D., Raftery, A.E., Kronmal, R.A., 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., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43, W01411. http://dx.doi.org/10.1029/2005WR004838.
Wainwright, H.M., Chen, J., Sassen, D.S., Hubbard, S.S., 2014. Bayesian hierarchical approach and geophysical data sets for estimation of reactive facies over plume scales. Water Resour. Res. 50 (4564–4584), 2013W. http://dx.doi.org/10.1002/R013842.
Winter, C.L., Nychka, D., 2010. Forecasting skill of model averaging. Stoch. Environ. Res. Risk Assess. http://dx.doi.org/10.1007/s00477-009-0350-.
Wöhling, T., Vrugt, J.A., 2008. Combining multi-objective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models. Water Resour. Res. 44, W12432. http://dx.doi.org/10.1029/2008WR007154.
Ye, M., Neuman, S.P., Meyer, P.D., 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resour. Res. 40, W05113. http://dx.doi.org/10.1029/2003WR002557.
Ye, M., Neuman, S.P., Meyer, P.D., Pohlmann, K.F., 2005. Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff. Water Resour. Res. 41, W12429. http://dx.doi.org/10.1029/2005WR004260.
Ye, M., Meyer, P.D., Neuman, S.P., 2008a. On model selection criteria in multimodel analysis. Water Resour. Res. 44, W03428. http://dx.doi.org/10.1029/2008WR006803.
Ye, M., Pohlmann, K.F., Chapman, J.B., 2008b. Expert elicitation of recharge model probabilities for the Death Valley regional flow system. J. Hydrol. 354, 102–115. http://dx.doi.org/10.1016/j.jhydrol.2008.03.001.
Ye, M., Meyer, P.D., Lin, Y.-F., Neuman, S.P., 2010a. Quantification of model uncertainty in environmental modeling. Stoch. Environ. Res. Risk Assess. http://dx.doi.org/10.1007/s00477-010-0377-0.
Ye, M., Pohlmann, K.F., Chapman, J.B., Pohll, G.M., Reeves, D.M., 2010b. A model averaging method for assessing groundwater conceptual model uncertainty. Ground Water. http://dx.doi.org/10.1111/j.1745-6584.2009.00633.x.
Ye, M., Lu, D., Neuman, S.P., Meyer, P.D., 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. http://dx.doi.org/10.1029/2009WR008501.