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. Yabusaki[2]

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, $\mathbf{C}_\varepsilon$, 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, $\mathbf{C}_{\mathbf{e}_k}$, 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 $\mathbf{C}_{\mathbf{e}_k}$ from the residuals during model calibration. The inferred $\mathbf{C}_{\mathbf{e}_k}$ 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 $\mathbf{C}_{\mathbf{e}_k}$ 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 $\mathbf{C}_{\mathbf{e}_k}$ 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 parameters [*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]. Multimodel 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, $\mathbf{M} = (M_1, ..., M_K)$, and denote $\hat{y}_k$ as a prediction (e.g., mean prediction or probability distribution) of model M_k for a

[1]Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA.
[2]Pacific Northwest National Laboratory, Richland, Washington, USA.
[3]U.S. Geological Survey, Menlo Park, California, USA.
[4]School of Earth Sciences and Engineering, Nanjing University, Nanjing, Jiangsu, China.
[5]Department 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)

0043-1397/13/10.1002/wrcr.20441

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

$$\hat{\bar{y}} = \sum_{k=1}^{K} w_k \hat{y}_k, \tag{1}$$

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

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

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]

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

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

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

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

where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood (ML) estimate of a vector $\boldsymbol{\theta}_k$ 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; $-\ln\left[L\left(\hat{\boldsymbol{\theta}}_k|\mathbf{D}\right)\right]$ is the minimum of the negative log likelihood (*NLL*) function, $-\ln[L(\boldsymbol{\theta}_k|\mathbf{D})]$, occurring, by definition, at $\hat{\boldsymbol{\theta}}_k$; $p\left(\hat{\boldsymbol{\theta}}_k\right)$ is the prior probability of $\boldsymbol{\theta}_k$ evaluated at $\hat{\boldsymbol{\theta}}_k$; and $\mathbf{F}_k$ is the observed (implicitly conditioned on the observations $\mathbf{D}$ and evaluated at the maximum likelihood parameter estimates $\hat{\boldsymbol{\theta}}_k$) Fisher information matrix having elements [*Kashyap*, 1982]

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

[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, $-2\ln\left[L\left(\hat{\boldsymbol{\theta}}_k|\mathbf{D}\right)\right]$, 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

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

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

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

[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

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

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

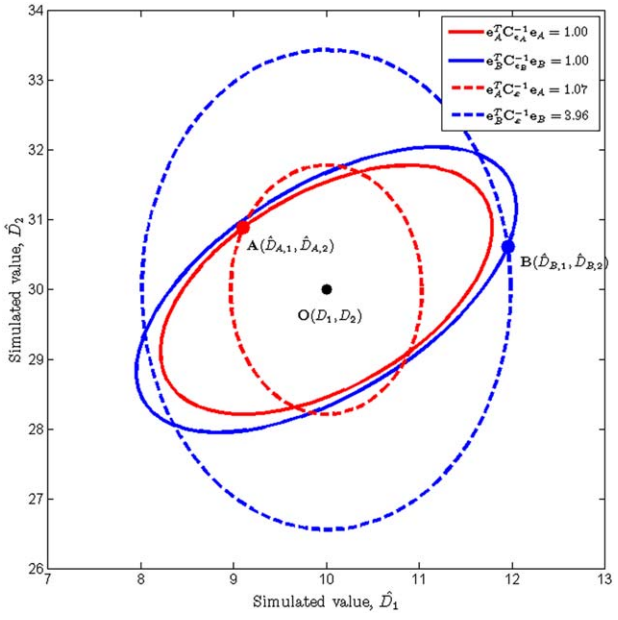

Figure 1. Illustration of effect of error covariance structure on the evaluation of *SSWR*, sum of squared weighted residuals. Point O represents the two observations D_1 and D_2; Points A and B are simulations of the observations by models A and B. The ellipses in solid red and blue lines are the *SSWR* contours based on matrices $\mathbf{C}_{\mathbf{e}_A}$ and $\mathbf{C}_{\mathbf{e}_B}$ 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 $\mathbf{C}_\varepsilon$ 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

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

[14] Assume that the joint probability distribution function of the total errors, $\mathbf{e}_k$, is multivariate Gaussian with zero mean and covariance matrix $\mathbf{C}_{\mathbf{e}_k}$. 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

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

[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 $\mathbf{C}_{\mathbf{e}_k}$ 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 $\mathbf{C}_{\mathbf{e}_k}$ 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, $\mathbf{C}_\varepsilon$, as

$$\mathbf{C}_\varepsilon = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}. \tag{13}$$

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

D_1 and D_2 are $\left(\hat{D}_{A,1} = 9.11,\ \hat{D}_{A,2} = 30.89\right)$ and $\left(\hat{D}_{B,1} = 11.96,\ \hat{D}_{B,2} = 30.62\right)$, 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:

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

[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 $\mathbf{C}_\varepsilon$ instead of $\mathbf{C}_{\mathbf{e}_k}$ on evaluating the goodness-of-fit of the two models as measured by $SSWR = \mathbf{r}_k^T \mathbf{C}_{\mathbf{e}_k}^{-1} \mathbf{r}_k$. When the full covariance matrix of total errors, $\mathbf{C}_{\mathbf{e}_k}$ is used, *SSWR* is

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

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

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

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, $\mathbf{C}_\varepsilon$, is used, *SSWR* is

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

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

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

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 $\mathbf{C}_{\mathbf{e}_k}$ is used, $BIC_A - BIC_B = \ln|\mathbf{C}_{\mathbf{e}_A}| - \ln|\mathbf{C}_{\mathbf{e}_B}| + \mathbf{r}_A^T \mathbf{C}_{\mathbf{e}_A}^{-1} \mathbf{r}_A - \mathbf{r}_B^T \mathbf{C}_{\mathbf{e}_B}^{-1} \mathbf{r}_B$, 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 $\ln|\mathbf{C}_{\mathbf{e}_k}|$. However, when $\mathbf{C}_\varepsilon$ is used, $BIC_A - BIC_B = \mathbf{r}_A^T \mathbf{C}_\varepsilon^{-1} \mathbf{r}_A - \mathbf{r}_B^T \mathbf{C}_\varepsilon^{-1} \mathbf{r}_B$, 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 $\ln|\mathbf{C}_{\mathbf{e}_k}|$ 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 $\left(\mathbf{X}_k^T \mathbf{C}_{\mathbf{e}_k}^{-1} \mathbf{X}_k\right)^{-1}$, 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, $D_t = \{D_1, D_2, \cdots, D_N\}$, measured at a sequence of discrete times $t = 1, 2, \cdots, N$ (e.g., observations of flows, heads, and/or concentrations collected over time), analogous to equation (11), we have

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

[24] Several discrete stochastic time series models are commonly used to simulate temporal correlation of the total errors $e_{k,t}$, including the pth-order autoregressive model, AR(p), the qth-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, $\mathbf{C}_{\mathbf{e}_k}$, of the total errors can be constructed. For example, if an AR(1) model is used, then $e_{k,t}$ can be quantified as $e_{k,t} = a e_{k,t-1} + \xi_t$, 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, $\mathbf{C}_{\mathbf{e}_k}$, can be expressed as

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

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

$$\mathbf{V} = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{N-1} \\ \rho_1 & & \ddots & \vdots \\ \vdots & & & \rho_1 \\ \rho_{N-1} & \cdots & \rho_1 & 1 \end{pmatrix} \tag{21}$$

with ρ_l (where $l = 1, 2, \cdots, N-1$) 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

$$\mathbf{V} = \begin{pmatrix} 1 & a & \ldots & a^{N-1} \\ a & & \ddots & \vdots \\ \vdots & & & a \\ a^{N-1} & \cdots & a & 1 \end{pmatrix}. \tag{22}$$

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

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

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 $\mathbf{C}_{\mathbf{e}_k}$. 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, $\mathbf{C}_{\mathbf{e}_k}$, of the total errors. Hereinafter, all the analyses are for residuals, with the ultimate goal of estimating $\mathbf{C}_{\mathbf{e}_k}$. 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 $\left\{r_{k,t}\right\} = D_t - f_k\left(\hat{\boldsymbol{\beta}}_{k,\varepsilon}\right), t = 1, 2, \cdots, N$, estimate the error variance $\hat{\sigma}^2$ by calculating the sample variance, $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{t=1}^{N}\left(r_{k,t} - \bar{r}_k\right)^2$ ($\bar{r}_k = \sum_{t=1}^{N} r_{k,t}/N$ is the sample mean), analyze the residuals serial correlation to determine the AR(p) model to simulate $\left\{r_{k,t}\right\}$, estimate $\hat{\mathbf{a}}_k$, and construct $\mathbf{C}_{\mathbf{e}_k}(\hat{\mathbf{a}}_k)$.

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

[30] 4. Replace $\hat{\boldsymbol{\beta}}_{k,\varepsilon}$ by $\hat{\boldsymbol{\beta}}_{\mathbf{e}_k}$ 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 $\mathbf{C}_{\mathbf{e}_k}(\hat{\mathbf{a}}_k)$ 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, $r_1, r_2, \ldots, r_N$, with constant intervals, the sample *ACF*, λ_l at lag l, is defined as

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

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 *ACF*s 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}, $r_{t-3}, \ldots, r_{t-l+1}$, 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 *PACF*s 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, **a**, of the model that is defined as

$$r_t = a_1 r_{t-1} + a_2 r_{t-2} + \ldots + a_p r_{t-p} + \xi_t, \tag{25}$$

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

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

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

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

[35] For $l = 1$,

$$\rho_1 = a_1 \rho_0 + a_2 \rho_{-1} = a_1 + a_2 \rho_1, \tag{28}$$

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

$$\rho_2 = a_1 \rho_1 + a_2. \tag{29}$$

[36] The method of moments replaces the theoretical *ACF*s ρ_1 and ρ_2 by the sample *ACF*s λ_1 and λ_2 calculated based on equation (24) to obtain

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

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

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

[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, **V**, defined in equation (21) can be specified. By virtue of equation (20), the covariance matrix, $\mathbf{C}_{\mathbf{e}_k}$, can be constructed after estimating the variance 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.

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

[a]Total site density used in this model is 1.3E-03 M/L. Summation of site fraction is one.

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	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	

K, is defined as $K_j = \left(S_jOUO_2OH\right)(H^+)^2/\left(\left(S_jOH\right)\left(UO_2^{2+}\right)\right)$, 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 *ACF*s 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 $(N\text{-}N_k)/N$ 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]

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

[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 $N\text{-}N_k$ 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.

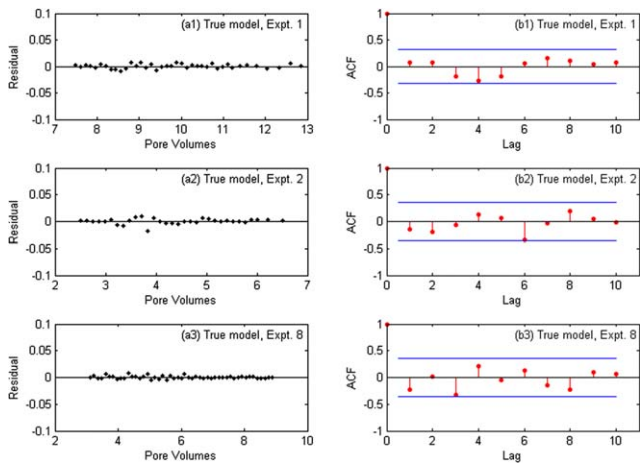

Figure 2. Plots of residuals of true model for (a1) Experiment 1, (a2) Experiment 2, and (a3) Experiment 8; (b1–b3) are sample *ACFs* plots of corresponding residuals.

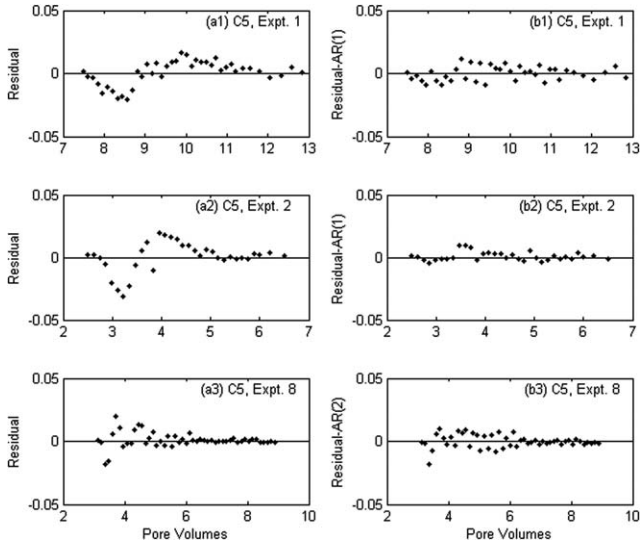

Figure 3. Plots of residuals (left) and residuals-AR(p) (right) of model C5 for (a1 and b1) Experiment 1, (a2 and b2) Experiment 2, and (a3 and b3) Experiment 8 in the synthetic study.

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 (logK1 and logK2) and the fraction of the strong site (logSite). 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 $\mathbf{Q}_k = \mathbf{C}_\varepsilon^{-1}$ as the weight matrix. Case II is the iterative two-stage model calibration using $\mathbf{Q}_k = \mathbf{C}_{\mathbf{e}_k}^{-1}$ 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 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, $\mathbf{C}_\varepsilon$, 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, $\mathbf{C}_{\mathbf{e}_k}$, 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.

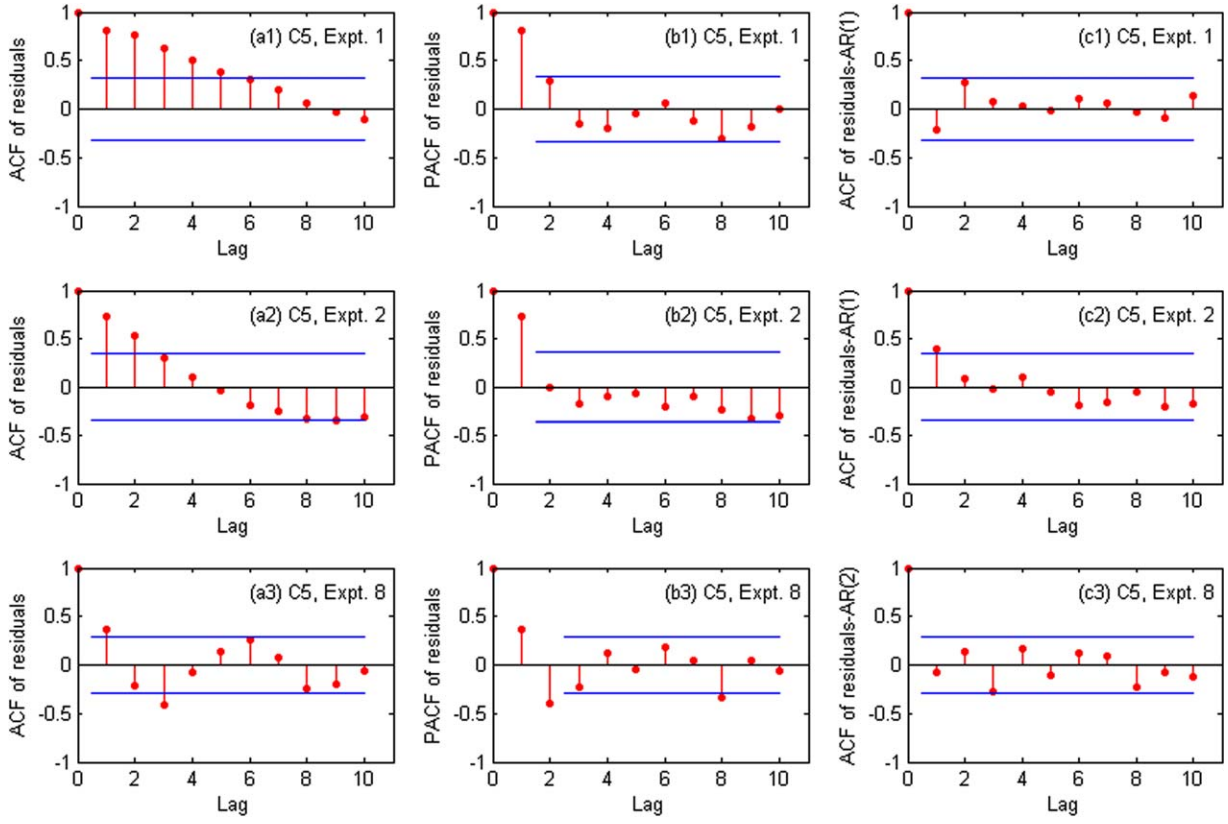

Figure 4. Sample *ACFs* plots of residuals of model C5 for (a1) Experiment 1, (a2) Experiment 2, and (a3) Experiment 8; (b1–b3) are sample *PACFs* plots of the corresponding residuals; (c1–c3) are sample *ACFs* plots of residuals-AR(p) in the synthetic study. The blue lines represent the 95% confidence 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, $\mathbf{C}_{\varepsilon_k}$, 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*, $\ln|\mathbf{C}|$, *NLL*, $\ln|\mathbf{F}|$, 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 $\mathbf{C}_\varepsilon^{-1}$ 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

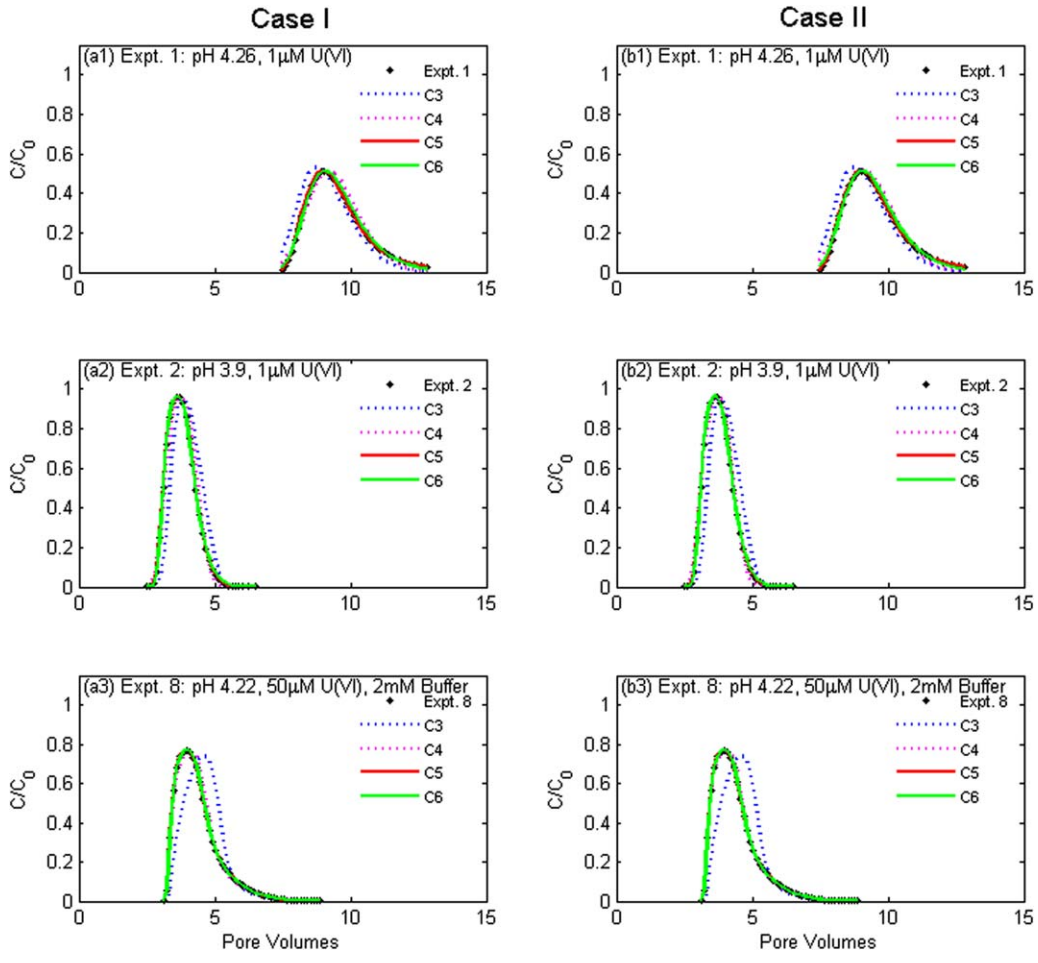

Figure 5. Comparison of observed and simulated breakthrough curves for (a1 and b1) Experiment 1, (a2 and b2) Experiment 2, and (a3 and b3) Experiment 8 for calibration Case I (left) and Case II (right) in the synthetic study.

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 $\mathbf{C}_\varepsilon^{-1}$ 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**|, Δ*NLL*, Δ*IC*, 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 $\mathbf{C}_\varepsilon$				Case II Using $\mathbf{C}_{\varepsilon_k}$			
	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
ΔAICc	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
αΔKIC[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

[a]Results 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.

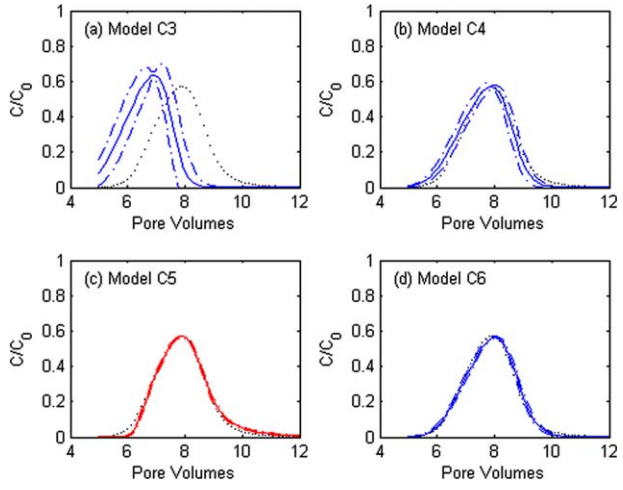

Figure 6. Predictions (solid blue lines) and their 95% linear confidence intervals (dashed blue lines) of the alternative models for Experiment 3 in Case I of the synthetic study. The *KIC*-based model averaging results are plotted in red lines in (c); the red lines are overlapped on the blue lines because 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 $\mathbf{C}_\varepsilon^{-1}$ 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 $\mathbf{C}_{\mathbf{e}_k}^{-1}$ 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, $\mathbf{C}_{\mathbf{e}_k}$, 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 $\ln|\mathbf{C}_{\mathbf{e}_k}|$ 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 $\ln|\mathbf{C}_{\mathbf{e}_k}|$ values are different for different models. Table 4 shows that in Case II model C3 has the largest $\ln|\mathbf{C}_{\mathbf{e}_k}|$ 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 temporal correlation of C6 is relatively larger than that of C5. The use of $\mathbf{C}_{\mathbf{e}_k}$ 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 $\mathbf{C}_\varepsilon^{-1}$ 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

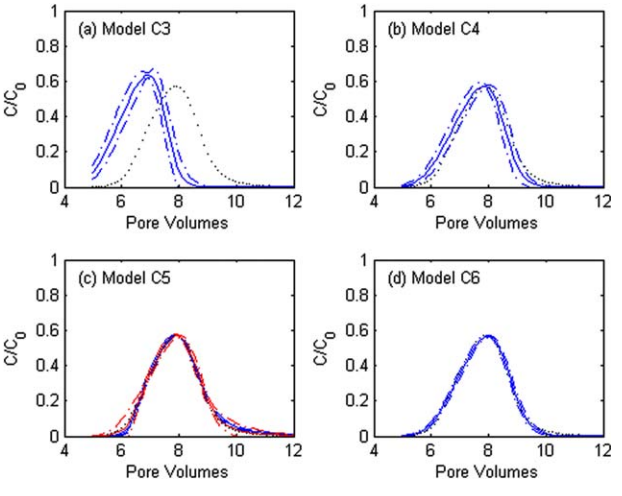

Figure 7. Predictions (solid blue lines) and their 95% linear confidence intervals (dashed blue lines) of the alternative models for Experiment 3 in Case II of the synthetic study. The *KIC*-based model averaging results are plotted in red lines in (c). Black dots represent the observed breakthrough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

						MLBMA	
	Model	C3	C4	C5	C6	*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]

[a]Results based on model averaging weights of $w_{\alpha KIC}$ (%) listed in Table 4.

[b]Results 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

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

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 weighting used in equation (32). The $s^2 \mathbf{Z}_k \left(\mathbf{X}_k^T \mathbf{Q}_k \mathbf{X}_k\right)^{-1} \mathbf{Z}_k^T$ term calculates prediction variance, a propagation of parameter estimate covariance $s^2 \left(\mathbf{X}_k^T \mathbf{Q}_k \mathbf{X}_k\right)^{-1}$ which in turn is a propagation of error covariance $\mathbf{Q}_k^{-1}$, i.e., $\mathbf{C}_\varepsilon$ in Case I and $\mathbf{C}_{\mathbf{e}_k}$ in Case II. The 95% confidence interval of model averaging is

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

where $\hat{\bar{y}}$ is the weighted average prediction calculated in equation (1), and $Var_{\bar{\hat{y}}}$ 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 $\mathbf{C}_{\mathbf{e}_k}$ 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 $\ln p(\hat{\mathbf{y}}_k | M_k, \mathbf{D}) = -\sum_{\hat{y}_k \in \hat{\mathbf{y}}_k} \ln p(\hat{y}_k | M_k, \mathbf{D})$, where $\mathbf{y}$ is the prediction 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

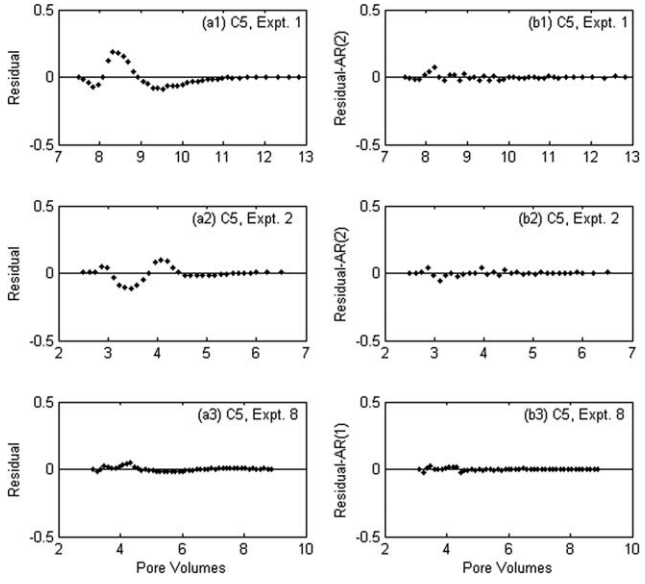

Figure 8. Plots of the residuals (left panel) and residuals-AR(p) (right panel) of model C5 for (a1 and b1) Experiment 1, (a2 and b2) Experiment 2, and (a3 and b3) Experiment 8 in the experimental study. The scale of y axis is 10 times larger than that of Figure 3, the counterpart of this figure.

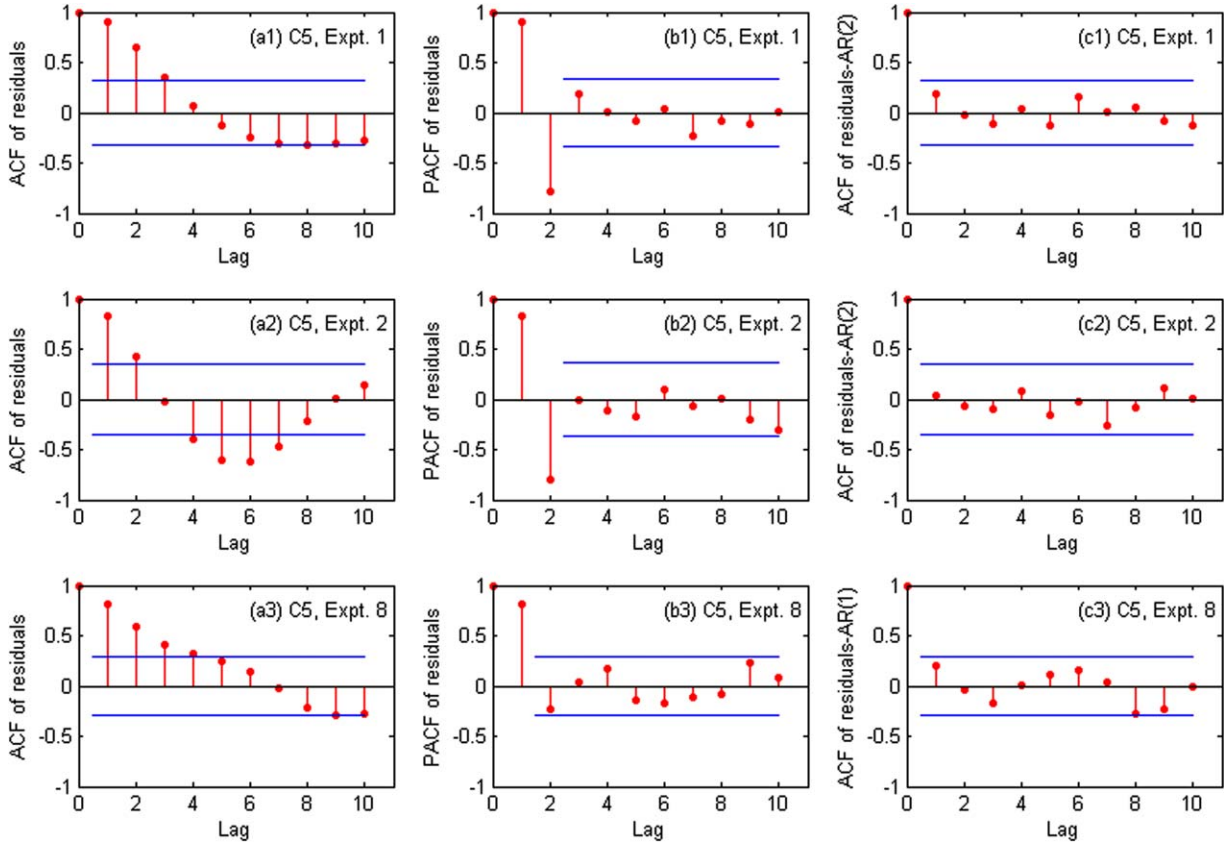

Figure 9. Sample *ACFs* plots of the residuals of model C5 for (a1) Experiment 1, (a2) Experiment 2, and (a3) Experiment 8; (b1)–(b3) are sample *PACFs* plots of the corresponding residuals; (c1)–(c3) are sample *ACFs* plots of residuals-AR(p) in the experimental study. The blue lines represent the 95%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

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

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

[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

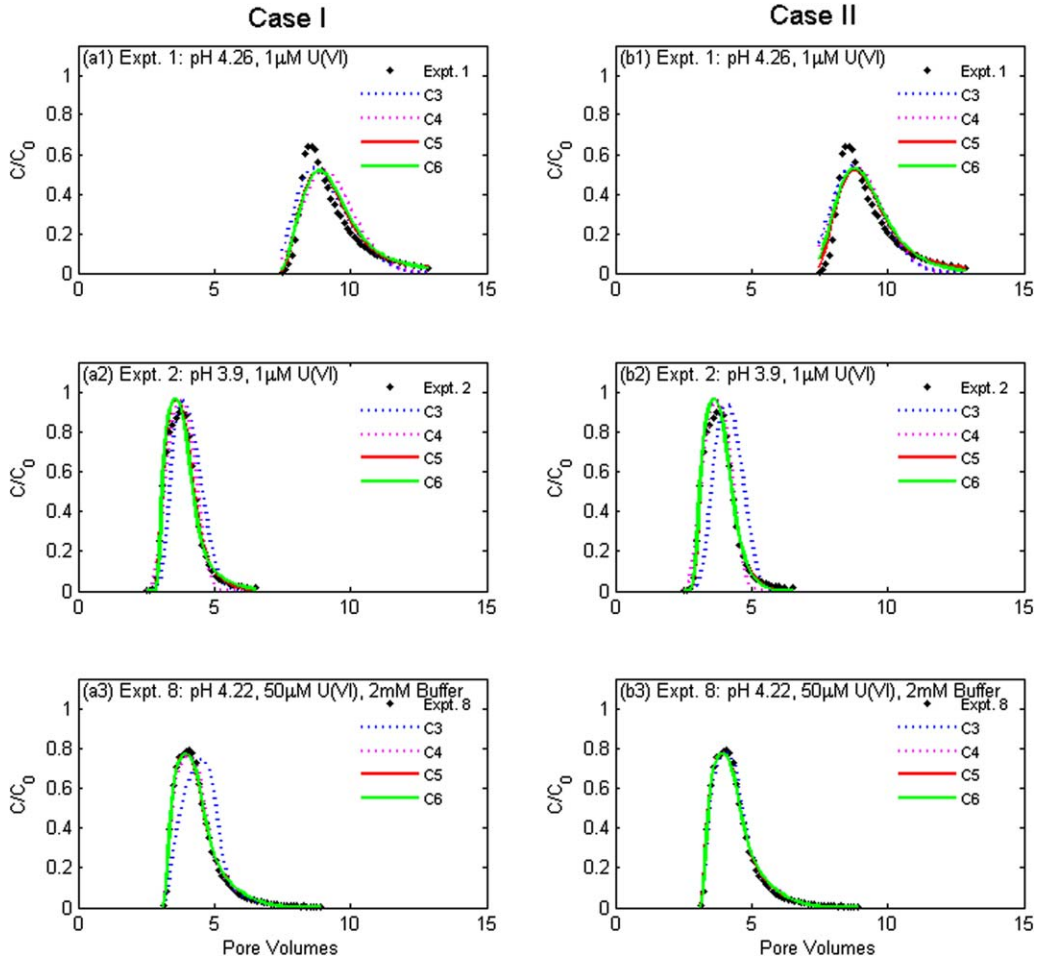

Figure 10. Comparison of observed and simulated breakthrough curves for (a1 and b1) Experiment 1, (a2 and b2) Experiment 2, and (a3 and b3) Experiment 8 for calibration Case I (left) and Case II (right) in the experimental study.

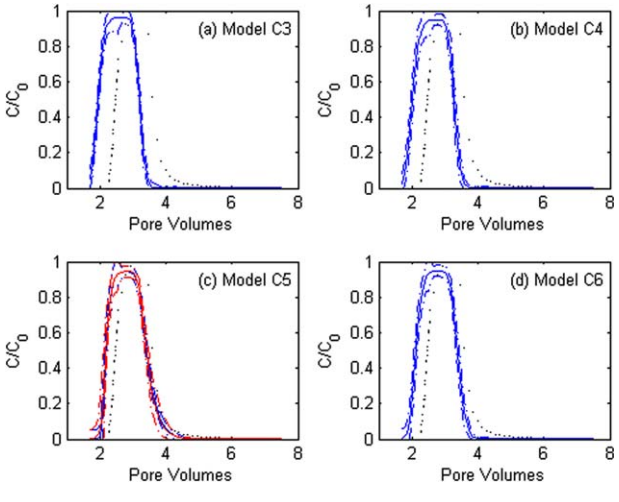

Figure 11. Predictions (solid blue lines) and their 95% linear confidence intervals (dashed blue lines) of the alternative models for Experiment 4 in Case II of the experimental study. The *KIC*-based model averaging results are plotted in red lines in (c). Black dots represent the observed breakthrough 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 $\mathbf{C}_\varepsilon^{-1}$ and $\mathbf{C}_{\varepsilon_k}^{-1}$ 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**|, Δ*NLL*, Δ*IC*, and Model Averaging Weights of Four Alternative Models Calculated for Case I and Case II Using *AICc, BIC*, *KIC* Averaging in the Experimental Study

	Case I Using $\mathbf{C}_\varepsilon$				Case II Using $\mathbf{C}_{\mathbf{e}_k}$			
	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
ΔAICc	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
αΔKIC[a]	8284.6	2115.6	0.0	249.4	13.9	4.4	0.0	0.3
w_{AICc} (%)	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

[a]Results based on 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**|, Δ*IC* 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 $\mathbf{C}_\varepsilon$. While this problem could not be resolved by using equation (8), it was resolved in Case II by using $\mathbf{C}_{\mathbf{e}_k}$ 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 underestimated 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, $\mathbf{C}_{\mathbf{e}_k}$, 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, $\mathbf{C}_\varepsilon$, of measurement errors, instead of the covariance matrix, $\mathbf{C}_{\mathbf{e}_k}$, 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 $\mathbf{C}_{\mathbf{e}_k}$ becomes the normalized Euclidian distance corresponding to $\mathbf{C}_\varepsilon$. 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, $\mathbf{C}_{\mathbf{e}_k}$, 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 $\mathbf{C}_\varepsilon$ 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 $\mathbf{C}_\varepsilon$ 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 $\mathbf{C}_{\mathbf{e}_k}$ gave more reasonable and realistic model averaging weights. This was supported by the calibration results and physical understanding of the alternative models. Using $\mathbf{C}_{\mathbf{e}_k}$ also improved predictive performance of the individual models by giving wider confidence intervals. Due to the reasonable model averaging weights obtained using $\mathbf{C}_{\mathbf{e}_k}$, 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 $\mathbf{C}_{\mathbf{e}_k}$ were almost identical to those obtained using $\mathbf{C}_\varepsilon$.

[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.

Bredehoeft, J. D. (2003), From models to performance assessment: The conceptualization problem, *Ground Water*, *41*(5), 571–577.

Bredehoeft, 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. Kwicklis, 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. Pohll (2007), Use of numerical groundwater 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. Troldborg (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. Pohll, 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.