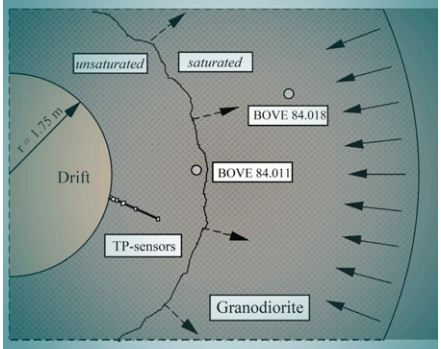
Special Section: Model-Data Fusion in the Vadose Zone

Xiaoqing Shi
Ming Ye*
Stefan Finsterle
Jichun Wu



Evaluating predictive performance of regression confidence intervals and Bayesian credible intervals is important for uncertainty quantification in model-data fusion. In this study numerical analyses showed that Bayesian intervals have better predictive performance and that MCMC simulation is computationally more efficient than regression analysis.

X. Shi and M. Ye, Dep. of Scientific Computing and Geophysical Fluid Dynamics Institute, Florida State Univ., Tallahassee, FL 32306, USA; S. Finsterle, Earth Sciences Division, Lawrence Berkeley National Lab., Berkeley, CA 94720, USA; X. Shi, J. Wu, Dep. of Hydrosciences, Nanjing Univ., Nanjing, 210098, China. *Corresponding author (mye@fsu.edu).

Vadose Zone J.
doi:10.2136/vzj2011.0147
Received 1 Nov. 2011.

Comparing Nonlinear Regression and Markov Chain Monte Carlo Methods for Assessment of Prediction Uncertainty in Vadose Zone Modeling

In vadose zone modeling, parameter estimates and model predictions are inherently uncertain, regardless of quality and quantity of data used in model-data fusion. Accurate quantification of the uncertainty is necessary to design future data collection for improving the predictive capability of models. This study is focused on evaluating predictive performance of two commonly used methods of uncertainty quantification: nonlinear regression and Bayesian methods. The former quantifies predictive uncertainty using the regression confidence interval (RCI), whereas the latter uses the Bayesian credible interval (BCI); neither RCI nor BCI includes measurement errors. When measurement errors are considered, the counterparts of RCI and BCI are regression prediction interval (RPI) and Bayesian prediction interval (BPI), respectively. The predictive performance is examined through a cross-validation study of two-phase flow modeling, and predictive logscore is used as the performance measure. The linear and nonlinear RCI and RPI are evaluated using UCODE_2005. The nonlinear RCI performs better than the linear RCI, and the nonlinear RPI outperforms the linear RPI. The Bayesian intervals are calculated using Markov Chain Monte Carlo (MCMC) techniques implemented with the differential evolution adaptive metropolis (DREAM) algorithm. The BCI/BPI obtained from DREAM has better predictive performance than the linear and nonlinear RCI/RPI. Different from observations in other studies, it is found that estimating nonlinear RCI/RPI is not computationally more efficient than estimating BCI/BPI in this case with low-dimensional parameter space and a large number of predictions. MCMC methods are thus more appealing than nonlinear regression methods for uncertainty quantification in vadose zone modeling.

Abbreviations: BCI, Bayesian credible interval; BPI, Bayesian prediction interval; DREAM, differential evolution adaptive metropolis; LHS, Latin hypercube sampling; MCMC, Markov Chain Monte Carlo; RCI, regression confidence interval; RPI, regression prediction interval.

Improving predictive accuracy and precision of vadose zone models requires accurate estimation of model parameters (e.g., water retention parameters) through inverse modeling, in which models and data are fused for parameter estimation (see the review article by Vrugt et al. (2008), and references therein). However, as explained in Abbaspour et al. (2004), parameter estimates and model predictions are inherently uncertain, regardless of quality and quantity of data used in model-data fusion. Accurate quantification of uncertainty is necessary to design future data collection for improving our understanding of vadose zone processes and for enhancing the predictive capability of vadose zone models to assess the performance of subsurface systems or to optimize prevention, mitigation, or cleanup measures (Abbaspour et al., 1996; Vrugt and Bouten, 2002; Yeh and Simunek, 2002; Wang et al., 2003; Liu and Yeh, 2004; Feyen and Gorelick, 2005; Nowak et al., 2010; Neuman et al., 2011). There are two kinds of widely used approaches for parameter estimation and uncertainty quantification. The first one comprises nonlinear regression methods, in which optimum model parameters are estimated, associated estimation uncertainties are evaluated using the least-square method, and predictive uncertainty is quantified using linear or nonlinear confidence intervals (Draper and Smith, 1981; Hill and Tiedeman, 2007). The second kind includes Bayesian methods, in which model parameters are treated as random variables and characterized by their probability density functions, and predictive uncertainty is quantified using linear or nonlinear credible intervals (Box and Tiao, 1992; Casella and Berger, 2002). While the linear confidence/credible intervals are evaluated for linear models or linearized nonlinear models, the nonlinear confidence/credible intervals are especially suitable for strongly nonlinear models.

The confidence and credible intervals have been widely used for quantifying parametric and predictive uncertainty in subsurface modeling (e.g., Vrugt et al. (2008), and references therein), including applications to vadose zone problems (Meyer et al., 1997; Schaap et al., 2001; Yeh and Simunek, 2002; Wang et al., 2003; Minasny and Field, 2005; Ye et al., 2007a, 2007b; Ye and Khaleel, 2008; Laloy et al., 2010a; Pan et al., 2009a, 2009b; Deng et al. 2009; Huisman et al., 2010). Confidence and credible intervals are conceptually different and often obtained using different numerical techniques. The confidence intervals are based on classical regression theories, in which true model parameters and predictions are considered to be fixed but unknown, and their estimates are random. The credible intervals are based on Bayesian theories, in which model parameters and predictions themselves are considered to be random variables. Lu et al. (2012) summarized literature that compared the two kinds of intervals in groundwater and vadose zone modeling. Their theoretical analysis showed that, for linear or linearized nonlinear models, linear confidence and credible intervals are mathematically equivalent when consistent prior parameter information is used. For nonlinear models, nonlinear confidence and credible intervals theoretically could be the same, but always differ in practice due to violation of assumptions used to derive the confidence intervals and/or numerical approximations used to calculate the credible intervals. This study considers linear and nonlinear confidence intervals as well as nonlinear credible interval.

While confidence and credible intervals have been widely used in predictive analysis of environmental modeling, little attention has been paid to comparison of predictive performance of the two kinds of intervals, i.e., how well the two kinds of intervals include future observations and which kind of intervals consistently assigns higher probability to future observations. For example, the study of Vrugt and Bouten (2002) was limited to the evaluation of the intervals of model parameters, not the intervals of model predictions. While Gallagher and Doherty (2007) and Lu et al. (2012) compared the intervals of model predictions, their comparisons were qualitative, and no definitive conclusions were drawn on predictive performance of the intervals. A comparative study on predictive performance is necessary, since the confidence/credible intervals are essential for defensible decision-making. For example, in the supplemental guidance of the USEPA for developing soil screening levels of superfund sites, the 95% upper confidence limit is recommended as a conservative estimate when performing a soil screening evaluation (USEPA, 2002). In the context of model-data fusion, predictive performance is an important success criterion. Therefore, one motivation of this study is to quantitatively evaluate predictive performance of the confidence and credible intervals.

In the statistical literature, equivalence between linear regression confidence and Bayesian credible intervals has been established for linear models (Box and Tiao, 1992); for nonlinear models, when model nonlinearity is very small, the same equivalence has also been established (Bates and Watts, 1988). However, discussion of the equivalence (especially for nonlinear confidence and credible intervals) in the context of groundwater and vadose zone modeling is rare. Cooley and Vecchia (1987) and Cooley and Naff (1990) are the pioneers of developing methods of calculating nonlinear confidence intervals. Use of nonlinear Bayesian credible intervals recently became popular due to the development of MCMC methods. In spite of these developments, differences and similarities between confidence and credible intervals have not been fully understood by groundwater and vadose zone modelers. We thus seek to contribute to the groundwater and vadose zone literature on the aspect of evaluating predictive performance of the two kinds of intervals.

While the confidence and credible intervals measure predictive uncertainty due to parametric uncertainty, they do not include measurement error. Since measurement error is inevitable in field observations, it should be considered in the evaluation of predictive performance of confidence and credible intervals. This requires calculation of regression and Bayesian prediction intervals (RPI and BPI defined in Section 2) based on the confidence and credible intervals. Therefore, this study evaluates predictive performance of not only confidence/credible intervals but also regression/Bayesian prediction intervals. Note that the prediction intervals are narrower than the total predictive uncertainty (e.g., that used in Vrugt et al., 2009a), because the latter includes errors in forcing, parameters, and model structure. When more sources of uncertainty are considered, model residuals (difference between observations and corresponding model simulations, a lump-sum of all errors) are more complicated, as they are likely to be correlated (e.g., Yang et al., 2008; Vrugt et al., 2009c; Laloy et al., 2010b) or exhibit non-Gaussian error distributions (Schoups and Vrugt, 2010). This study is limited to impact of measurement error in calculation of prediction intervals, and it is reasonable to assume that measurement errors are independent and Gaussian with zero mean. However, ignoring other uncertainty sources may result in underestimation of predictive uncertainty, as shown in the numerical example of this study.

This study is also focused on evaluation of computational efficiency of calculating the confidence and credible intervals. Since the credible intervals cannot be evaluated analytically except in idealized situations (e.g., integration in denominator of Bayes' theorem can be evaluated analytically; Hou and Rubin, 2005), the intervals need to be obtained numerically using, for example, computationally expensive MCMC methods. Due to the burn-in period of MCMC and slow convergence when inappropriate proposal distributions are used, calculating the nonlinear credible intervals is generally considered to be computationally more expensive than calculating the nonlinear confidence intervals. However, it is unknown whether this is still the case when more advanced MCMC techniques are used. In addition, this conclusion was mainly drawn when only one or a few predictions were made (e.g.,

Christensen et al., 2006; Gallagher and Doherty, 2007). Given that the nonlinear confidence intervals are evaluated for each individual prediction using a procedure that could be computationally expensive (Christensen and Cooley, 1999; Cooley, 2004), for a relatively large number of predictions, it is unknown whether calculating the nonlinear confidence intervals is still computationally more efficient than evaluating the nonlinear credible intervals.

The above questions are explored using a numerical experiment revised from the two-phase flow problem of Finsterle and Pruess (1995), which examines moisture inflow into a tunnel excavated from crystalline rock and the development of a dry-out zone in response to tunnel ventilation. The work of Finsterle and Pruess (1995) involves nine unknown parameters, which include permeability, porosity, five parameters related to the relative permeability and capillary pressure functions, and two parameters related to vapor diffusion in porous media. These parameters are estimated based on 162 water potential data taken at six distances from the tunnel wall. Additional data include two observations of gas pressure and one observation of evaporative water flux. In this study, the two least influential parameters identified by Finsterle and Pruess (1995), which are related to vapor diffusion in porous media, are not considered in the uncertainty analysis, and one more parameter related to the capillary pressure function is added. This leads to eight uncertain parameters in this study, i.e., permeability, porosity, and six parameters related to the relative permeability and capillary pressure functions. More discussion of the eight unknown parameters of this study is provided in Section 2.4 below. This problem is selected because it is considered to be representative for two-phase flow problems in terms of unknown parameters and quantities of prediction. For example, this kind of problem is typical in the unsaturated zone modeling of the Yucca Mountain site for geological storage of nuclear waste (e.g., Ye et al., 2007a; Pan et al., 2009a). In addition, the conclusions of this study are expected to be applicable to saturated zone models with similar or a larger number of model parameters and predictions. However, the conclusions may not be applicable to large-scale problems, since spatial variability of model parameters is not considered in this study.

The nonlinear regression analysis is undertaken using UCODE_2005 (Poeter et al., 2005), general-purpose computational software widely used in groundwater and vadose zone modeling. The code is based on nonlinear regression theories, and model calibration is performed using the Gauss-Marquardt-Levenberg method for estimating optimum parameters. UCODE_2005 consists of six computer toolboxes for model calibration, sensitivity analysis, and uncertainty evaluation. The linear confidence intervals and prediction intervals are estimated using the toolbox LINEAR_UNCERTAINTY; the nonlinear confidence and prediction intervals are obtained by running UCODE_2005 in nonlinear-uncertainty mode (Poeter et al., 2005, Chapter 17). Certain control parameters coefficients used for the algorithm of calculating the nonlinear intervals are not easy to determine, and a trial-and-error approach is used in the numerical example of this study.

The Bayesian analysis is conducted using a recently developed MCMC method, DREAM (Vrugt et al., 2009b, 2009c). Based on the Differential Evolution–Markov Chain (DE–MC) method of ter Braak (2006), DREAM was developed to improve sampling efficiency for complicated posterior parameter distributions with multiple modes by improving search efficiency of MCMC sampling. It runs multiple Markov chains in parallel to explore different regions of the parameter space. DREAM has been used widely in surface hydrology modeling, and more applications of the DREAM algorithm to groundwater and vadose zone modeling have been reported in the literature (e.g., Keating et al., 2010; Laloy et al., 2010a; Wohling and Vrugt, 2011).

Predictive performance is evaluated using the cross-validation methods in a manner patterned after Ye et al. (2004, 2008). In the cross-validation, since the confidence and credible intervals of model predictions are compared with field measurements, measurement errors are incorporated into the evaluation of the intervals. The resulting intervals are called prediction intervals (regression and Bayesian), which will be discussed in more detail in the next section. The predictive performance is quantified using predictive logscore defined below. Computational efficiency of calculating the confidence and credible intervals is evaluated using the total execution time. It is found that, for problems with large number of predictions, calculating the credible intervals is not necessarily more computationally expensive.

Materials and Methods

This section provides a brief overview of definitions and techniques of estimating confidence and credible intervals. Detailed discussions are referred to Draper and Smith (1981), Box and Tiao (1992), Hill and Tiedeman (2007), and Lu et al. (2012). For a random variable X, both confidence and credible intervals can be defined symbolically as

$$\text{Prob}(l \le X \le u) = 1 - \alpha \qquad [1]$$

where l and u are lower and upper interval limits and α is significance level. However, the definition is interpreted in different ways for the confidence and credible intervals, rendering the two kinds of intervals conceptually different. Take the intervals for a model prediction as an example. A confidence interval with a confidence level of, for example, 95% ($\alpha = 0.05$), is an interval that is expected to include the true value of the prediction 95% of the time in repeated sampling of observations used in regression (McClave and Sincich, 2000). A credible interval represents the posterior probability that the prediction lies in the interval (Box and Tiao, 1992; Casella and Berger, 2002), and the interval is determined via

$$\int_l^u p(g(\beta)|\mathbf{y})dg(\beta)=1-\alpha \quad [2]$$

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

$$p(g(\beta)\le l|\mathbf{y})=p(g(\beta)\ge u|\mathbf{y})=\alpha/2 \quad [3]$$

Other methods of estimating the credible intervals (e.g., highest posterior density interval) are also available (Box and Tiao, 1992; Chen and Shao, 1999; Casella and Berger, 2002) but not used in this work. Given that the linear confidence and credible intervals of linear models are mathematically equivalent (Lu et al., 2012), only linear and nonlinear confidence intervals and nonlinear credible intervals are considered in this study.

Linear and Nonlinear Confidence Intervals

Generally speaking, a nonlinear model, denoted by f, can be expressed as

$$\mathbf{y}=f(\beta)+\varepsilon \quad [4]$$

where $\mathbf{y}$ is a vector of n observations, β is a vector of p model parameters, and $\boldsymbol{\varepsilon}$ is a vector of statistically independent errors with zero expectation and covariance matrix $\mathbf{C}_\varepsilon=\sigma^2\omega^{-1}$, ω being an $n\times n$ known weight matrix and σ^2 a scalar, which is generally unknown but can be estimated (Carrera and Neuman, 1986). Theoretically speaking, ε includes all sources of errors such as measurement error, parameter error, and model structure error. However, this study assumes that model structure is correct and only considers measurement error and parameter error. This assumption is reasonable for the numerical problem of this study, as the cross-validation results show that the prediction intervals (defined in Section 2.3) cover the majority of observations.

To estimate the linear confidence interval, the nonlinear model is linearized by expanding it in a Taylor series and retaining only the first two terms, i.e., $f(\hat{\beta})\approx f(\beta^*)+\mathbf{X}_{\beta^*}(\hat{\beta}-\beta^*)$, where $\mathbf{X}_{\beta^*}=[\partial f/\partial\beta]_{\beta=\beta^*}$ is the sensitivity matrix, β^* is the true value of β, and $\hat{\beta}$ is the estimated parameter by minimizing the generalized least-squares objective function $S(\beta)=[\mathbf{y}-f(\beta)]^T\omega[\mathbf{y}-f(\beta)]$. For a linear or effectively linear model, $\hat{\beta}$ follows asymptotically the multivariate normal distribution, $\hat{\beta}\sim N_p(\beta^*,\sigma^2(\mathbf{X}_{\beta^*}^T\omega\mathbf{X}_{\beta^*})^{-1})$, because of the normality assumption of errors. If the nonlinear prediction function $g(\hat{\beta})$ is also approximated to the first order by $g(\hat{\beta})\approx g(\beta^*)+\mathbf{Z}_{\beta^*}^T(\hat{\beta}-\beta^*)$, where $\mathbf{Z}_{\beta^*}$ is prediction sensitivity vector $\mathbf{Z}_{\beta^*}=[\partial g/\partial\beta]_{\beta=\beta^*}$, then the $(1-\alpha)\times100\%$ linear confidence interval of $g(\beta^*)$ is (Seber and Wild, 2003)

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

where $t_{1-\alpha/2,n-p}$ is a t statistic with significance level α and degrees of freedom $n-p$, and $s^2=S(\hat{\beta})/(n-p)$ is estimated variance. In practice, the sensitivity matrices are approximated by replacing β^* by the estimate, $\hat{\beta}$ (Seber and Wild, 2003, p. 191). Effects of model nonlinearity on the approximation can be found in Hill and Tiedeman (2007, p. 393–398).

Calculating the nonlinear confidence interval for a nonlinear model does not require model linearization. As illustrated in Hill and Tiedeman (2007, p.178), the interval is determined as the maximum and minimum of model predictions intersecting a confidence region of model parameters (Vecchia and Cooley, 1987). The keys to estimating the nonlinear confidence interval are to evaluate the parameter confidence region and the maximum and minimum intersections. The approximate likelihood confidence region is defined as the set of parameter values whose corresponding objective function values, $S(\mathbf{b})$, satisfy (Christensen and Cooley, 1999; Cooley, 2004; Hill and Tiedeman, 2007, p. 178)

$$S(\beta)\le S(\hat{\beta})\left[\frac{1}{n-p}t_{\alpha/2,n-p}^2+1\right] \quad [6]$$

This parameter region contains the true model parameter with approximate probability of $(1-\alpha)\times100\%$. Estimation of nonlinear confidence intervals based on Eq. [6] can be done using UCODE_2005, following instructions of the user's manual (Poeter et al., 2005, Chapter 17). The estimation is based on an iterative method, and our experience shows that certain parameters of the method need to be carefully determined by trial and error to obtain stable solutions. In addition, accurate evaluation of the nonlinear confidence interval requires the following assumptions (Lu et al., 2012): (i) the model accurately represents the system, (ii) model predictions, $g(\beta)$, are sufficiently monotonic, (iii) there is a single minimum in the objective function, (iv) the residuals are multivariate normal distributed, and (v) model intrinsic nonlinearity is small (Cooley and Naff, 1990; Hill and Tiedeman, 2007). If the assumptions are not satisfied, Eq. [6] may not define the objective function value associated with the designated $1-\alpha$ confidence level, and the estimated nonlinear intervals may be smaller or larger than the true intervals.

Nonlinear Credible Intervals

Estimation of the nonlinear credible intervals requires evaluating the distribution of model prediction $g(\beta)$, which in turn requires knowledge of the distribution of model parameters β. Different from the regression theories, in which $\hat{\beta}$ and $g(\hat{\beta})$ are treated as random

variables and their distributions are estimated, Bayesian theories treat β and $g(\beta)$ as random variables and their distributions are estimated (Box and Tiao, 1992). The posterior parameter distribution, $p(\beta|\mathbf{y})$ (conditioned on $\mathbf{y}$), is estimated via the Bayes' theorem

$$p(\beta|\mathbf{y}) = \frac{p(\mathbf{y}|\beta)p(\beta)}{p(\mathbf{y})} \quad [7]$$

where $p(\beta)$ is the prior distribution, and $p(\mathbf{y}|\beta)$ is the likelihood function. The most commonly used likelihood function is the multivariate Gaussian

$$p(\mathbf{y}|\beta) = \frac{1}{(2\pi)^{p/2}|\mathbf{C}_\varepsilon|}\exp\left[-\frac{1}{2}(\mathbf{y}-f(\beta))^T\mathbf{C}_\varepsilon^{-1}(\mathbf{y}-f(\beta))\right] \quad [8]$$

which assumes that the residuals, $\mathbf{y}-f(\beta)$, follow the normal distribution. In this study, the posterior parameter distributions are estimated numerically using the DREAM algorithm. After the posterior parameter distribution is estimated, parameter samples can be drawn and the distribution of model prediction can be evaluated. Subsequently, the $(1-\alpha)\times 100\%$ credible interval can be determined using the equal-tail method of Eq. [3]. The interval, however, does not guarantee to bracket field observations, because only parametric uncertainty is considered in the evaluation. It happens often that model structure uncertainty dominates over parametric uncertainty; the numerical example below shows the importance of considering measurement errors.

Prediction Intervals

In the above ways of evaluating confidence and credible intervals of model predictions, measurement errors are not incorporated in the prediction. When comparing the confidence and credible intervals with field observations that are subject to measurement errors, incorporating measurement errors to the predictive intervals is necessary, and this results in prediction intervals. The linear prediction interval is evaluated based on Eq. [5] via (Cooley, 2004; Hill and Tiedeman, 2007)

$$g(\hat{\beta}) \pm t_{1-\alpha/2,n-p}\left[s^2\left(\mathbf{Z}_\beta^T(\mathbf{X}_\beta^T\boldsymbol{\omega}\mathbf{X}_\beta)^{-1}\mathbf{Z}_\beta + s_p^2\right)\right]^{1/2} \quad [9]$$

where s_p^2 is the variance of the error associated with a measured equivalent of the prediction. The nonlinear prediction intervals can be estimated by revising Eq. [6] as (Cooley, 2004; Christensen et al., 2006)

$$S(\beta) \le S(\hat{\beta})\left[\frac{1}{n-p}t_{\alpha/2,(n-p)}^2 + 1\right] - \omega_p\upsilon^2 \quad [10]$$

where υ is an estimate of prediction error and ω_p is the inverse of variance of the prediction. The predictive intervals are referred to as regression predictive intervals (RPI) (linear and nonlinear) to be distinguished from the regression confidence intervals (RCI). Evaluation of the linear RPI is performed using the UCODE_2005 toolbox of LINEAR_UNCERTAINTY, and the nonlinear RPI is obtained by running UCODE_2005 in nonlinear uncertainty mode (Poeter et al., 2005, Chapter 17).

Following Gallagher and Doherty (2007), the nonlinear BPI are computed by adding, to each realization of the DREAM model predictions, a random measurement error generated from a Gaussian distribution whose mean is zero and standard deviation equals to that of the measurement errors. The covariance of measurement errors is the same as $\mathbf{C}_\varepsilon$ used in Eq. [8], if the quantities of predictions are the same as the quantities (i.e., $\mathbf{y}$) used for estimation of posterior parameter distributions. Otherwise, the covariance matrix of the measurement errors differs from $\mathbf{C}_\varepsilon$. The resulting realizations of model predictions are then used to estimate the BPI in the manner of evaluating the Bayesian credible intervals (BCI) discussed above.

Numerical Example of Two-phase Flow Model

The RCI, BCI, RPI, and BPI are evaluated for a two-phase flow model developed by Finsterle and Pruess (1995). To determine the macro-permeability of crystalline rocks, starting on 26 Nov. 1991, a series of ventilation tests were conducted (Gimmi et al., 1997) at the Grimsel Rock Laboratory, Switzerland, a research facility operated by the Swiss National Cooperative for the Disposal of Radioactive Waste. As shown in Fig. 1, the experimental site was located in mildly deformed granodiorite that was considered homogeneous on the scale of interest. Boreholes BOVE 84.011 and BOVE 84.018 were drilled parallel to a tunnel and equipped with conventional pressure transducers to observe the pressure head. Thermocouple psychrometers sensors were installed at six different depths (2, 5, 10, 20, 40, and 80 cm from the drift wall) to measure negative water potentials in the partially saturated region as a

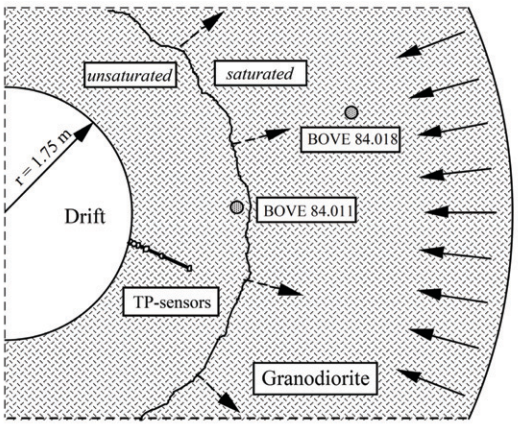


Fig. 1. Schematic of model domain, showing a vertical cross-section with boreholes for measuring gas pressure, and sensors for measuring water potential, adapted from Finsterle and Pruess (1995). Right-pointing arrows indicate the propagation of the dry-out zone against the prevalent groundwater flow (left-pointing arrows) toward the ventilation tunnel.

function of time. The total inflow to the drift was obtained from measurements of the moisture extracted from the circulated air in a cooling trap. These tests were interpreted using a two-phase, radial flow model implemented using TOUGH2 (Pruess et al., 1999); details of the model are described in Finsterle and Pruess (1995). In the model, for preventing capillary pressure from decreasing toward negative infinity as the effective saturation approaches zero, the relative permeability function and capillary pressure function of van Genuchten (1980) were revised as (Finsterle, 1999, 2007)

$$S_{ec} = \frac{S_l - S_{lr}}{1 - S_{lr}} \quad [11a]$$

$$S_{ek} = \frac{S_l - S_{lr}}{1 - S_{lr} - S_{gr}} \quad [11b]$$

$$S_{\varepsilon} = \frac{\varepsilon}{1 - S_{lr}} \quad [11c]$$

$$p_c = -\frac{1}{\alpha}\left[\left(S_{ec}\right)^{1/m} - 1\right]^{1/n} \quad \text{for } S_l \geq \left(S_{lr} - \varepsilon\right) \quad [12a]$$

$$p_c = -\frac{1}{\alpha}\left[\left(S_{ec}\right)^{1/m} - 1\right]^{1/n} - \beta\left(S_l - S_{lr} + \varepsilon\right) \quad \text{for } S_l < \left(S_{lr} - \varepsilon\right) \quad [12b]$$

$$\beta = -\left(\alpha n m (1 - S_{lr})\right)^{-1}\left(S_{\varepsilon}^{-1/m} - 1\right)^{(1/n)-1} S_{\varepsilon}^{-(1+m)/m} \quad [12c]$$

$$k_{rl} = S_{ek}^{\eta}\left[1 - \left(1 - S_{ek}^{1/m}\right)^{m}\right]^2 \quad [13a]$$

$$k_{rg} = \left(1 - S_{ek}\right)^{\zeta}\left[1 - S_{ek}^{1/m}\right]^{2m} \quad [13b]$$

where S_{ec} and S_{ek} are the effective saturations for the capillary pressure and relative permeability functions, respectively, S_{ε} is the effective saturation at a point near residual liquid saturation (i.e., at $S_l = S_{lr} + \varepsilon$, where ε is a small value specified as 0.01 in this study), β is the slope of the van Genuchten model (Luckner et al., 1989) evaluated at S_{ε}, which is used for a linear extrapolation of the capillary pressure curve to prevent it from decreasing toward negative infinity as the effective saturation S_{ec} approaches zero, S_{lr} is residual liquid saturation, S_{gr} is residual gas saturation, p_c is macroscopic capillary pressure, α is the van Genuchten parameter related to gas entry pressure [Pa], n is the van Genuchten parameter related to the pore size distribution index, and $m = 1 - 1/n$, and the exponents η and ζ are related to the tortuosity of the liquid- and gas-filled pore space, respectively. The two-phase model requires estimating six parameters for the relative permeability function and capillary pressure function (S_{lr}, S_{gr}, α, n, η, and ζ), as well as porosity ϕ and absolute permeability k of the homogeneous granodiorite matrix. For heterogeneous matrix, the assumption of homogeneity may result in underestimation of prediction uncertainty, because this assumption introduces model structure error (e.g., Pan et al., 2009b). The homogeneity assumption is valid from a physical point of view, and it is also confirmed in the modeling results below. Following Finsterle and Pruess (1995), the parameters k and $1/\alpha$ are log-transformed for estimation.

Nonlinearity of this model is examined by calculating Beale's measures of total nonlinearity and intrinsic nonlinearity using UCODE_2005. According to Cooley and Naff (1990) and Hill and Tiedeman (2007, p. 142–145), a model is effectively linear, moderately nonlinear, nonlinear, and highly nonlinear if the critical values for total and intrinsic model nonlinearity measure are less than 0.01, between 0.01 and 0.9, between 0.09 and 1.0, and larger than 1.0, respectively. For this two-phase problem, since the critical values for total nonlinearity and intrinsic nonlinearity measures are 0.007 and 0.004, respectively, the model is effectively linear. This leads to similar (but still different) linear and nonlinear confidence intervals, as shown in Results below.

Model Calibration and Sensitivity Analysis

The parameters are estimated in this study using the observations of the total inflow (q), the two gas pressures (h) at the two boreholes, and water potentials (p) at 27 logarithmically spaced points in time at the six different observation depths. Assuming that the observations are uncorrelated, the objective function for parameter estimation is

$$\text{SSWR} = \sum_{i=1}^{N_h} \omega_{h_i}\left(Y_{h_i} - \hat{Y}_{h_i}\right)^2 + \sum_{i=1}^{N_q} \omega_{q_i}\left(Y_{q_i} - \hat{Y}_{q_i}\right)^2 + \sum_{i=1}^{N_p} \omega_{p_i}\left(Y_{p_i} - \hat{Y}_{p_i}\right)^2 \quad [14]$$

where Y and $\hat{Y}$ are, respectively, observed and simulated values of the three kinds of observations, ω denotes inverse of variances of measurement errors of the observations and their corresponding values are given in Finsterle and Pruess (1995), and $N_h = 2$, $N_q = 1$, and $N_p = 162$ are the numbers of observations of the respective kinds. Following Finsterle and Pruess (1995), it is assumed that the standard deviations of the measurement errors are 10% of the measured values; this measurement error is also used for evaluation of prediction intervals. Given that the three kinds of observations are independent, this objective function is equivalent to the negative logarithm of the Gaussian likelihood function used in Eq. [8] for the Bayesian analysis (Hill and Tiedeman, 2007, Appendix A).

The most critical parameters to the objective function are selected using the Morris One-At-a-Time method implemented in the Design Analysis Kit for Optimization and Terascale Applications toolkit (DAKOTA) (Adams et al., 2010). This global sensitivity analysis method is based on the elementary effect calculated for the ith model input as (Morris, 1991)

$$d_i(\mathbf{P}) = \frac{f(P_1, \ldots, P_{i-1}, P_i + \Delta, P_{i+1}, \ldots, P_k) - f(\mathbf{P})}{\Delta} \quad [15]$$

where $\mathbf{P} = \{P_1, ..., P_k\}$ are model inputs, f is model output (the objective function in this study), and Δ is a predetermined multiplier of the ith model input. The mean effect measures the influence of parameter P_i on the model output; a high mean value indicates large overall influence. A high standard deviation of the mean effect suggests that the parameter is either interacting with other parameters or has a nonlinear effect on the output.

Logscore to Measure Predictive Performance

The Bayesian intervals are compared with the counterparts of regression intervals to evaluate their predictive performance, i.e., how well the intervals include future observations and which kinds of intervals consistently assign higher probability to future observations. The evaluation is done using cross-validation, in which the observations of water potentials at the six different depths are split into two parts: the observations at four depths (denoted as $\mathbf{D}^A$) are used for model calibration (the observations of total inflow and gas pressure measurements are also included in the calibration), and the observations at the remaining two depths (denoted as $\mathbf{D}^B$) are used for the predictive analysis. The prediction, $\hat{\mathbf{D}}^B$, and corresponding confidence and credible intervals are used for cross-validation. Repeating this process for a number of cross-validation cases can help gain insights into predictive performance. In this study, six cases of cross-validation are conducted. As shown below, the six cases reveal consistent results, suggesting that there is no need to consider more cases of cross-validation.

The predictive performance is evaluated using predictive logscore $-\ln p(\mathbf{D}^B \mid \mathbf{D}^A)$ (Good, 1952; Volinsky et al., 1997). The lower the predictive logscore based on data $\mathbf{D}^A$, the smaller the amount of information lost on eliminating $\mathbf{D}^B$ from the original dataset $\mathbf{D}$ (i.e., the higher the probability to reproduce the lost data, $\mathbf{D}^B$). When evaluating the logscore for the regression-based intervals, according to Eq. [5] and effective linearity of the model, model predictions are assumed to be Gaussian. For the Bayesian credible and prediction intervals, examining the model predictions indicates that the assumption of Gaussian distribution is also reasonable (results not shown). Therefore, the logscore is calculated via

$$-\ln p(\mathbf{D}^B \mid \mathbf{D}^A) = \frac{N_d}{2}\ln(2\pi) + \frac{1}{2}\sum_{i=1}^{N_d}\sigma_i^2 + \frac{1}{2}\sum_{i=1}^{N_d}\frac{(D_i^B - \bar{D}_i^B)^2}{\sigma_i^2} \quad [16]$$

where N_d is the dimension of $\mathbf{D}^B$, D_i^B are its components, and $\bar{D}_i^B$ and σ_i^2 are the corresponding mean prediction and prediction variance. When calculating logscore for the regression-based linear intervals, $\bar{D}_i^B$ and σ_i^2 are estimated based on Eq. [5] and [9] for confidence and prediction intervals, respectively. For the regression-based nonlinear intervals, $\bar{D}_i^B$ and σ_i^2 are estimated by assuming that the two interval values are the 2.5% and 97.5% quantiles of the normal distribution. For the Bayesian intervals, $\bar{D}_i^B$ and σ_i^2 are mean and variance of the predictions estimated from DREAM samples.

Results

In this section, we investigate predictive performance of the regression intervals evaluated using UCODE_2005 (Poeter et al., 2005), which can be done using other software such as PEST (Doherty, 2005) and iTOUGH2 (Finsterle, 2004; Finsterle and Zhang, 2011). The Bayesian intervals are estimated using DREAM.

Results of Sensitivity Analysis and Model Calibration

Table 1 lists the parameter ranges used for the Morris analysis; each of the ranges is divided into 99 levels, which leads to (8 + 1) × 100 = 900 forward model runs. The mean and variance of the elementary effect are plotted in Fig. 2. It shows that $\log(1/\alpha)$, $\log(k)$, and n are the three most influential parameters, consistent

Table 1. Estimates and standard deviations (SD) of the eight parameters obtained using UCODE_2005 and mean and SD of the parameters obtained using DREAM. The SD values in dark shading are unreasonable and caused by low sensitivity of the three parameters. The optimum values of the sum of squared weighted residuals are also listed.

		UCODE_2005			DREAM	
Parameter	Range	Initial	Estimate	SD	Mean	SD
ϕ	[0.003, 0.025]	0.01	0.0096	0.9	0.0143	0.01
$\log(k)$	[−15, −22]	−17	−18.5	0.1	−18.5	0.04
n	[1,6]	3	2.46	0.1	2.41	0.08
$\log(1/\alpha)$	[0,1]	0.1	0.222	0.03	0.234	0.01
S_{lr}	[0,0.05]	0.0	0.0500	0.2	0.0279	0.01
S_{gr}	[0,0.05]	0.0	0.0068	0.002	0.0059	0.004
η	[0,1]	0.0	0.425	0.4	0.670	0.2
ζ	[0,1]	0.0	0.0831	0.04	0.112	0.1
Sum of squared weighted residuals	–	–	129.5		124.6	

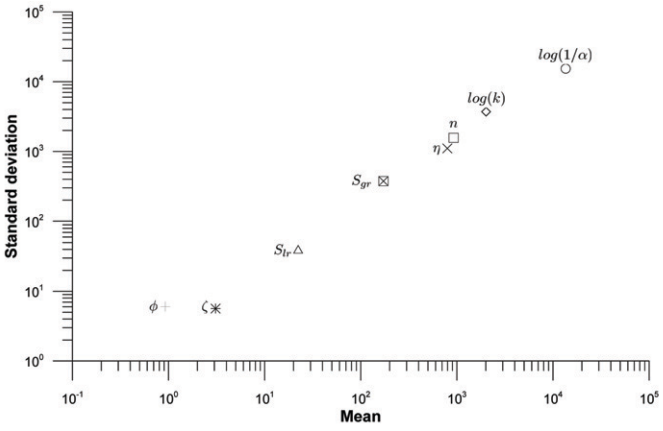


Fig. 2. Mean and standard deviation of elementary effect of the objection function estimated using the Morris One-At-a-Time method. The three most sensitive parameters are $\log(1/\alpha)$, $\log(k)$ and n.

with the results of the local sensitivity analysis of Finsterle and Pruess (1995). The three parameters were also found to be the most influential in Pan et al. (2011), in which a sampling-based regression method was used for the unsaturated zone at Yucca Mountain. While all eight parameters are calibrated in this section and included in the DREAM simulation of the "Results of DREAM Simulation" section, to reduce computational cost, only the three most important parameters are considered in "Evaluation of Predictive Performance using Cross-validation" in which predictive performance of the confidence and credible intervals are evaluated. As pointed out by Shi et al. (unpublished data, 2012), when computational cost is high and only a limited number of DREAM simulations is affordable, the estimation of posterior distributions is more reliable for influential parameters than for non-influential parameters.

All eight parameters are estimated using UCODE_2005. The initial parameter values and ranges used in the calibration are adopted from Finsterle and Pruess (1995) and listed in Table 1. The minimum SSWR values of the regression and DREAM are also listed in the table. The calibration SSWR is slightly better than that of calibrating the three most influential parameters in Finsterle and Pruess (1995), as is expected due to the larger number of adjustable parameters. The optimum SSWR values of DREAM are smaller than that of UCODE_2005, which is attributed to the gradient-based, local optimization method used by UCODE_2005. While the UCODE_2005 optimization may be improved by tuning its parameters such as tolerance of parameter changes, the default values of UCODE_2005 were used in this study.

Table 1 lists the parameter estimates and associated standard deviations obtained from UCODE_2005 together with the means and standard deviations of the parameter distributions using DREAM. Since the model is effectively linear, the parameter estimates are expected to approximately follow $N_p(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}_{\hat{\boldsymbol{\beta}}}^T \boldsymbol{\omega} \mathbf{X}_{\hat{\boldsymbol{\beta}}})^{-1})$. Although the parameter distributions obtained from DREAM are non-Gaussian as shown below, the means and standard deviations can be used as summary statistics for comparing the results of the regression and DREAM simulations. While the regression-based parameter estimates are similar to the mean parameters of DREAM, the standard deviations of regression are dramatically larger for parameters, ϕ, S_{lr}, and η. This is attributed to low sensitivity (i.e., small values of the sensitivity matrix, $\mathbf{X}$) of the available observations with respect to the three parameters as shown in Fig. 2. If one follows the nonlinear regression theories and assumes that the parameter estimates are Gaussian, the regression-based standard deviations of the three parameters are too large to be reasonable, because adding or subtracting 2 standard deviations around the estimates may lead to physically unreasonable parameter values, e.g., ϕ (porosity) less than zero or larger than one. This problem does not occur with DREAM results, because MCMC methods do not make any assumptions on parameter distributions. In this sense, MCMC methods are more suitable for estimation of the parameter distributions. The differences in parameter distributions between the regression and MCMC methods and their effects on predictive uncertainty will be discussed below.

Results of DREAM simulation

The DREAM simulation is conducted using uniform prior distributions with the ranges shown in Table 1. Eight Markov chains are run in parallel. Convergence of the DREAM simulation is monitored using the potential scale reduction factor, R, of Gelman and Rubin (1992). The R value becomes less than the critical value 1.1 (which suggests convergence) after 20,000 parameter realizations are sampled and the first 2000 realizations during the burn-in period are deleted. Figure 3 shows the histograms of the eight parameters obtained from DREAM. Although the model is effectively linear, the distributions of parameters ϕ, S_{lr}, and η, and ζ are non-Gaussian. This is a result of the fact that the prior distributions are informative because (i) the ranges of the parameters are narrow and (ii) the model outputs are not sensitive to these parameters (Fig. 2). In other words, the posterior distributions are not solely determined by the Gaussian likelihood function but jointly by the likelihood function and the prior distributions.

To verify the posterior parameter distributions obtained from DREAM, a conventional MC simulation is conducted to infer the true parameter distributions. The parameter samples are drawn using the Latin hypercube sampling (LHS) method, assuming that the parameters are independent and follow uniform distributions with the ranges listed in Table 1. A total of 700,000 model runs are conducted in parallel using seven processors, which takes about 11.6 d. Except the LHS method, no other sampling techniques are used so that their limitations do not affect the verification. The realizations that generate acceptable model goodness-of-fit between observations and corresponding simulations are retained to infer the parameter distributions. The acceptable goodness-of-fit is measured by the objective function; in a convenient way, the threshold value of objective function S_0 is determined based on the UCODE_2005 model calibration via (Christensen and Cooley, 1999)

$$S_0 = S(\hat{\boldsymbol{\beta}})\left[\frac{t^2_{\alpha/2,n-p}}{(n-p)}+1\right] \qquad [17]$$

where $\alpha = 0.05$ is the significance level. This threshold is used only to select realizations in which field observations are reasonably reproduced, and use of Eq. [17] should not be considered as a mixture of nonlinear regression and Bayesian analysis. Based on this threshold value (144.61), only 349 parameter realizations are retained, and the histograms of the eight parameters are plotted in Fig. 4. Although the histograms are not the posterior parameter distributions (because of the empirically determined threshold value and limited number of model runs), they provide insight into the parameter distributions. Comparing Fig. 3–4 shows that the histograms of the simple MC are similar to those of

DREAM for all the parameters. Given that only 20,000 model executions are conducted for DREAM, the numerical results suggest that DREAM is numerically efficient to infer the posterior parameter distributions.

Evaluation of Predictive Performance Using Cross-Validation

The evaluation is done using cross-validation, in which the observations of water potentials at the six different depths are split into two parts: the observations at four depths are used for model calibration, and the observations at the remaining two depths are used for the predictive analysis. To be consistent with Finsterle (1999) and to reduce computational cost, only the three most sensitive parameters ($\log(1/\alpha)$, $\log k$, and n) are considered during the cross-validation; other parameters are fixed at the values used by Finsterle (1999), i.e., the initial values listed in Table 1. The calibration results using all the data are listed as the reference case in Table 2. When only the three most influential parameters are calibrated, the goodness-of-fit slightly deteriorates, with the minimum objective function of regression increasing from 129.49 (Table 1) to 136.10 (Table 2, the reference case). However, the parameter estimates of the three parameters are almost identical, as expected. The optimum objective function of DREAM also increases from 124.63 to 130.04. Table 2 lists the estimated means and standard deviations of the three parameters for the six cross-validation cases. In each case, while the optimum parameters obtained from the nonlinear regression are almost identical to the mean parameters obtained from DREAM, the standard deviation of DREAM is on average 37% larger than that of the nonlinear regression. The differences in parameter distributions affect predictive performance of the nonlinear regression and DREAM; the effects are discussed below.

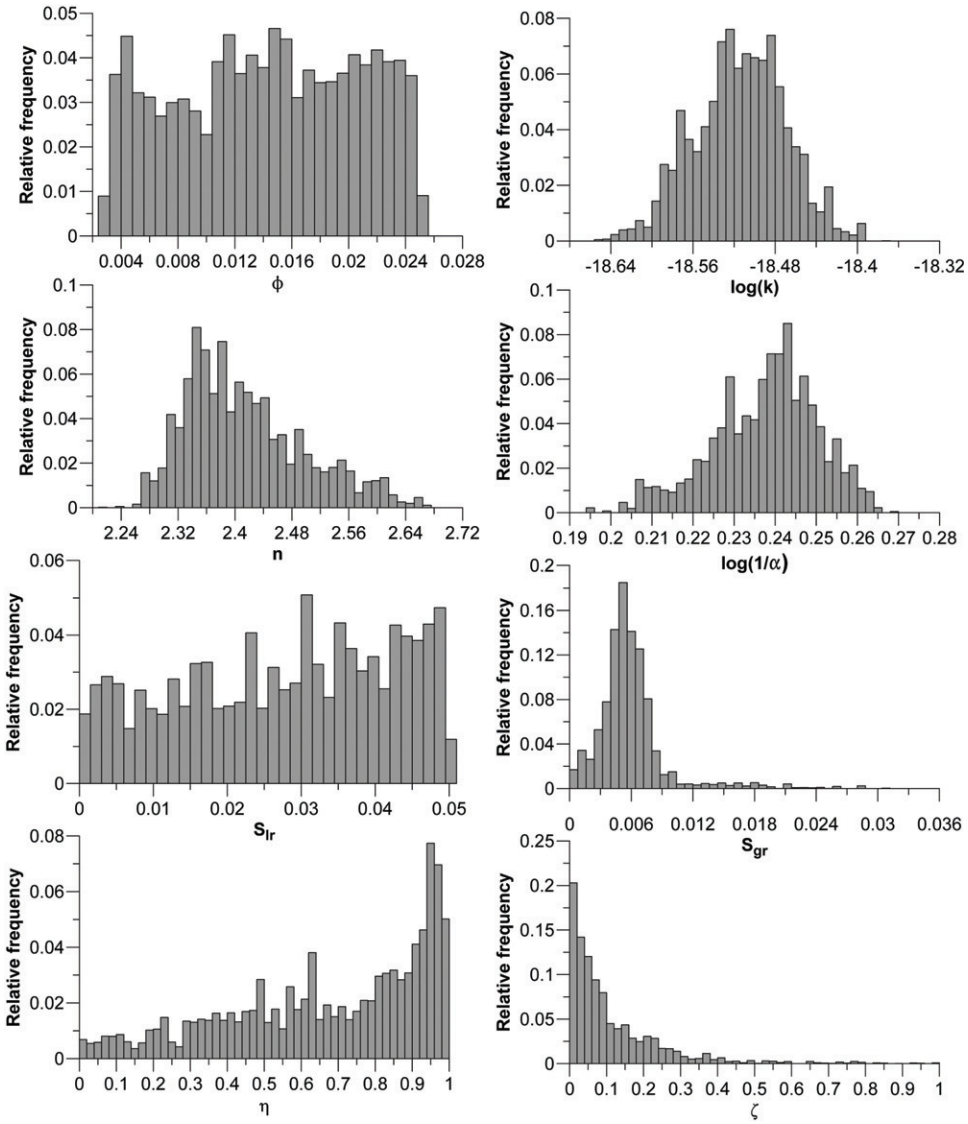


Fig. 3. Histograms of the eight model parameters obtained from simulations of differential evolution adaptive metropolis (DREAM).

The predictive performance is first evaluated for the 95% RCI and BCI without incorporating measurement errors. As shown in Fig. 5, the RCI (linear and nonlinear) and BCI are visually almost identical due to the large scales of the y axis. Table 3 shows that the predictive logscore of BCI is smaller than that of linear and nonlinear RCI for all six cross-validation cases, indicating that BCI has better predictive performance. The nonlinear RCI outperforms the linear RCI only slightly, which is not surprising given that the model is effectively linear. Table 3 shows significant differences between the logscore of RCI and BCI. This is caused by the differences in predictive variance of RCI and BCI, as shown in Fig. 6 which plots $p(\mathbf{D}^B \mid \mathbf{D}^A)$ and $-\ln p(\mathbf{D}^B \mid \mathbf{D}^A)$ for three observations; predictive accuracy is the worst in Fig. 6a and 6b but best in Fig. 6e and 6f. Figures 6c–6f show that the logscores of RCI and BCI are similar when the prediction variances of RCI and BCI are similar. Figures 6a and 6b show that the logscore becomes significantly different when the prediction variances are different. Figure 6a shows that, while the mean predictions corresponding to nonlinear RCI and BCI are very close (−1,070,250 and −1,069,496 Pa, respectively), the standard deviations of the predictions are significantly different (5841 and 10,312 Pa, respectively). As a result, Fig. 6b shows that the logscores corresponding to nonlinear RCI and

BCI are 39 and 19, respectively. These differences indicate that the logscore measures not only accuracy but also precision of model predictions.

The predictive performance of the prediction intervals is also evaluated in the manner similar to that for the predictive performance of confidence/credible intervals. Similar to Fig. 5, Fig. 7 plots the linear and nonlinear RPI and BPI together with the cross-validation data for the six cases. The prediction intervals bracket the majority of the observations, indicating that structural error in this numerical example is insignificant. Fluctuation in the nonlinear RPI is observed; the reason is that the nonlinear intervals are calculated independently for the individual predictions through an iterative process similar to that of model calibration using the Gauss-Marquardt-Levenberg method. Unlike Fig. 5, Fig. 7 shows that there are distinct differences between the RPI (linear and nonlinear) and BPI. The linear RPI always has the largest width, which may lead to conservative decision-making or management plans. However, predictive logscore listed in Table 3 shows that predictive performance of the linear RPI is the same as that of the nonlinear RPI and BPI in cases 2 and 5 but worse in the other four

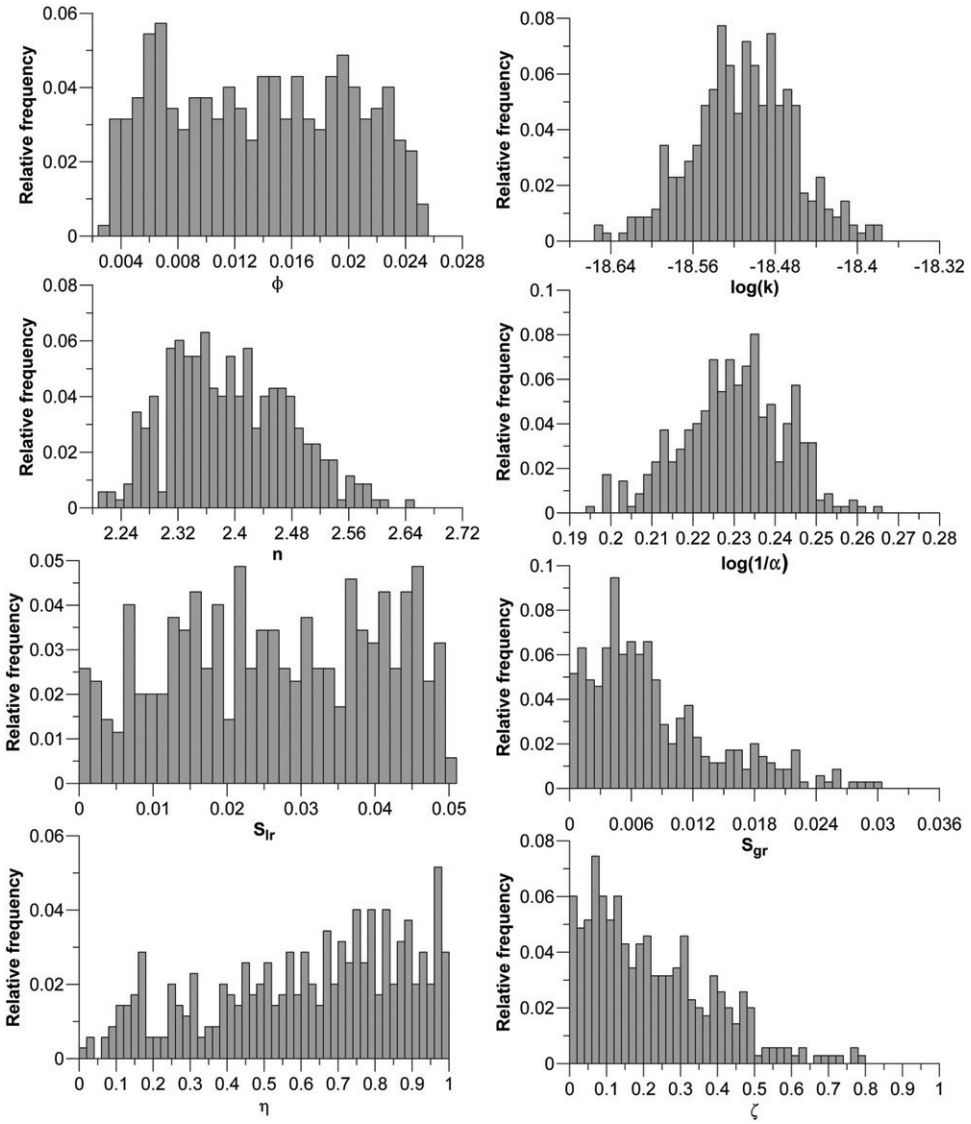


Fig. 4. Histograms of the eight parameters obtained from the simulations of brute-force, simple MC.

Table 2. Mean and standard deviations (SD) of three parameters obtained using DREAM and estimates and SD of the parameters obtained using UCODE_2005 for the reference and six cross-validation cases. The optimum values of the sum of squared weighted residuals are also listed.

	Calibration	Prediction		log(*k*) Mean ± SD	*n* Mean ± SD	log(1/α) Mean ± SD	Sum of squared weighted residuals
Reference	1, 2, 3, 4, 5, 6	–	Nonlinear	−18.5 ± 0.03	2.47 ± 0.03	0.232 ± 0.005	136.1
Case 1	2, 3, 4, 5	1, 6	DREAM	−18.6 ± 0.03	2.80 ± 0.08	0.227 ± 0.006	82.0
			Nonlinear	−18.6 ± 0.02	2.81 ± 0.05	0.225 ± 0.005	82.1
Case 2	1, 4, 5, 6	2, 3	DREAM	−18.6 ± 0.03	2.47 ± 0.04	0.234 ± 0.007	131.8
			Nonlinear	−18.5 ± 0.03	2.46 ± 0.03	0.234 ± 0.006	131.8
Case 3	1, 2, 3, 6	4, 5	DREAM	−18.6 ± 0.03	2.40 ± 0.03	0.233 ± 0.006	51.3
			Nonlinear	−18.6 ± 0.03	2.41 ± 0.04	0.233 ± 0.003	51.3
Case 4	2, 3, 5, 6	1, 4	DREAM	−18.6 ± 0.03	2.48 ± 0.04	0.228 ± 0.007	90.1
			Nonlinear	−18.6 ± 0.02	2.48 ± 0.03	0.229 ± 0.005	90.1
Case 5	1, 3, 4, 6	2, 5	DREAM	−18.5 ± 0.03	2.43 ± 0.04	0.235 ± 0.006	93.0
			Nonlinear	−18.5 ± 0.02	2.43 ± 0.03	0.234 ± 0.005	93.0
Case 6	1, 2, 4, 5	3, 6	DREAM	−18.6 ± 0.03	2.62 ± 0.06	0.240 ± 0.006	94.5
			Nonlinear	−18.6 ± 0.02	2.61 ± 0.04	0.239 ± 0.005	94.6

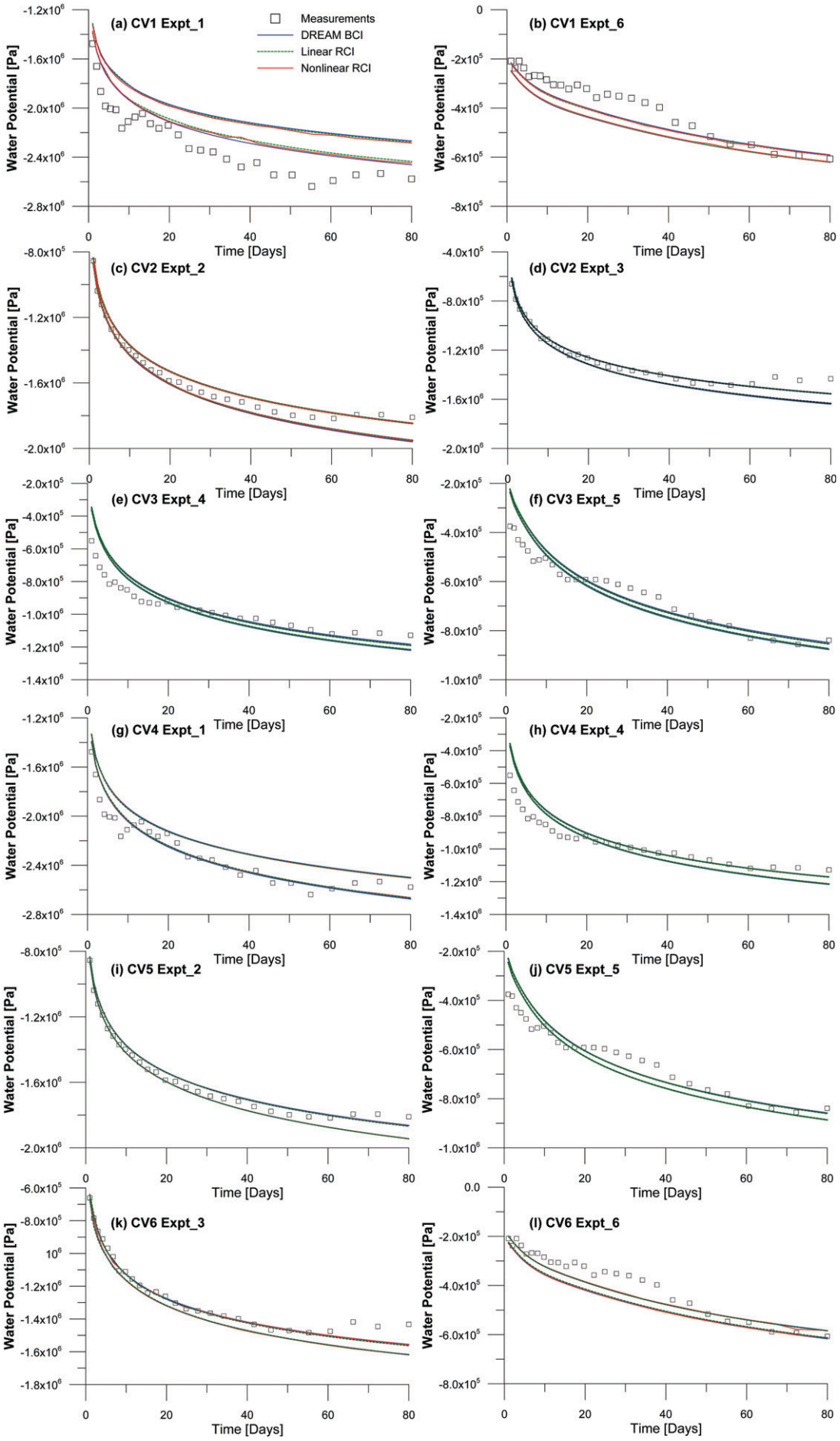


Fig. 5. Measurements of water potential and 95% linear and nonlinear regression confidence intervals (RCI) and Bayesian credible intervals (BCI) for the six cross-validation cases.

Table 3. Predictive logscore of the linear and nonlinear 95% regression confidence intervals (RCI) and regression prediction intervals (RPI) obtained using UCODE_2005, together with the Bayesian credible intervals (BCI) and Bayesian prediction intervals (BPI) obtained using DREAM for the six cross-validation cases.

	Linear		Nonlinear		DREAM	
	RCI	RPI	RCI	RPI	BCI	BPI
Case 1	2716	740	2676	737	1624	734
Case 2	723	690	721	690	703	690
Case 3	8023	3905	7944	753	3858	702
Case 4	4824	1782	4646	743	3268	716
Case 5	2717	691	2635	691	1927	691
Case 6	1573	686	1306	684	1137	684

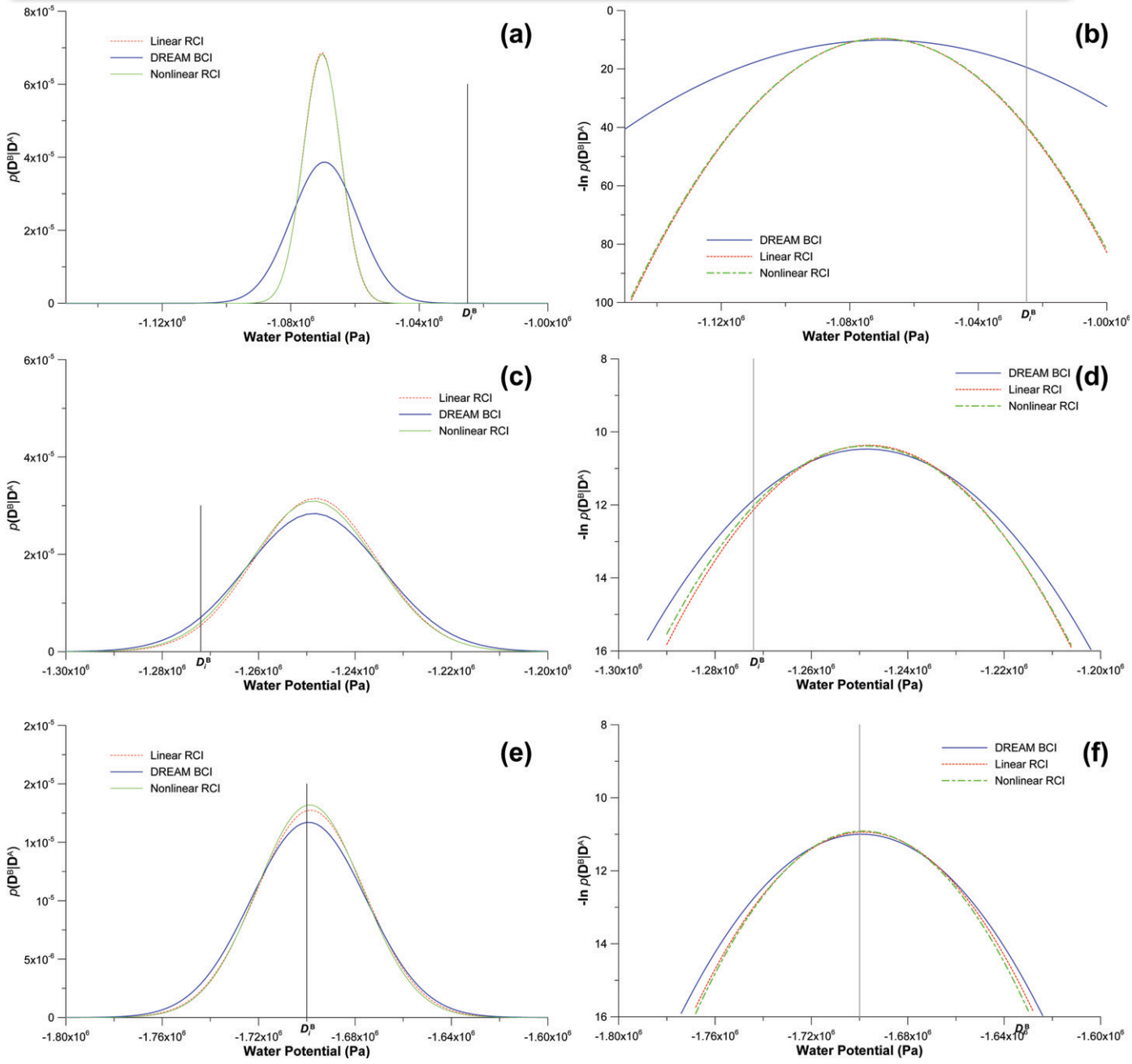


Fig. 6. (left) Predictive probability, $p(\mathbf{D}^B \mid \mathbf{D}^A)$, and (right) predictive logscore, $-\ln p(\mathbf{D}^B \mid \mathbf{D}^A)$, calculated based on the linear and nonlinear 95% regression confidence intervals and the Bayesian credible intervals for observations of water potential (a and b) at t = 41.68 d in experiment 4 of cross-validation case 3; (c and d) at t = 5.417 d in experiment 2 of cross-validation case 2, and (e and f) at t = 34.18 d of experiment 2 of cross-validation case 2. The vertical lines represent the locations of the corresponding observation D_i^B.

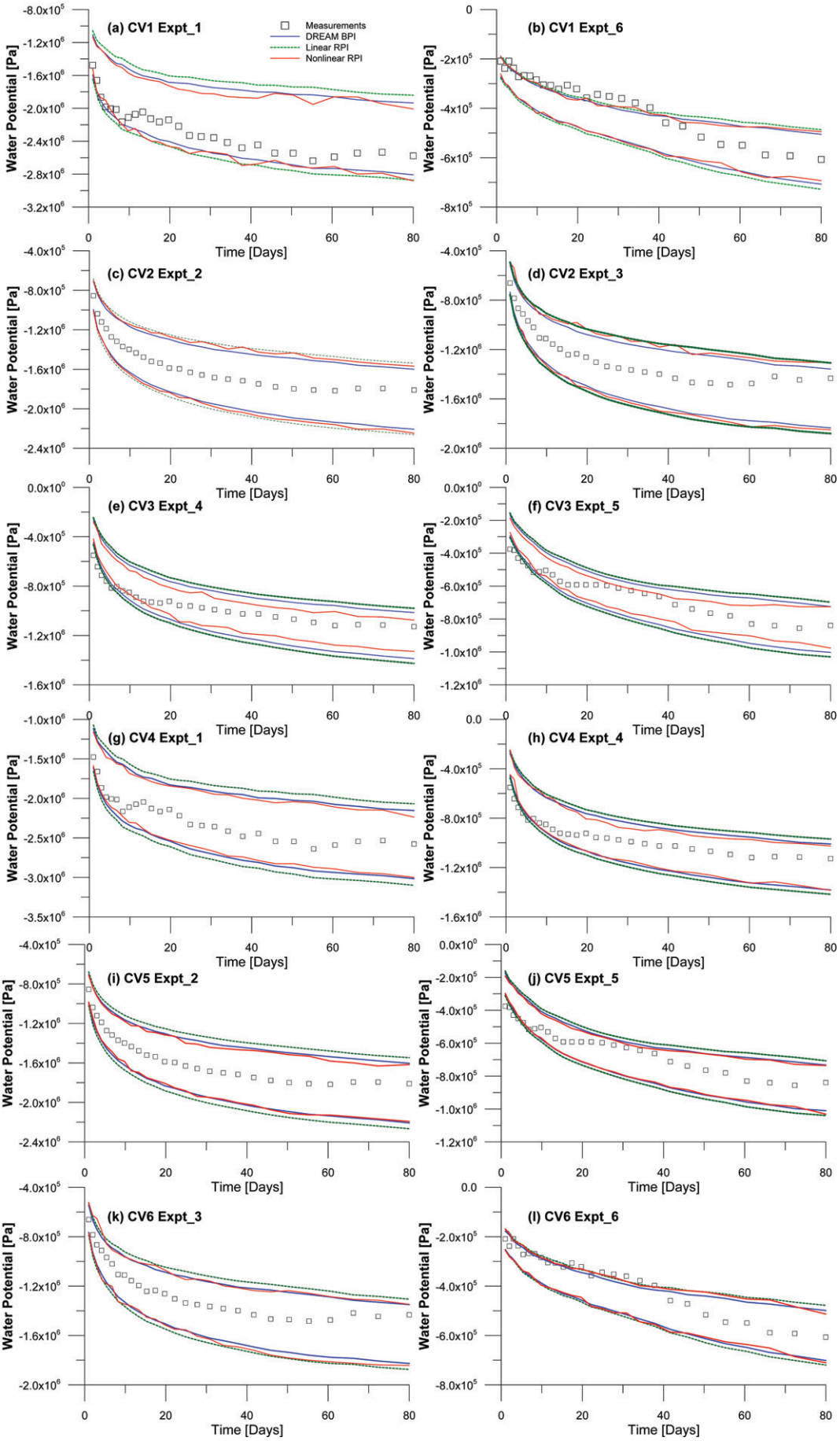


Fig. 7. Measurements of water potential and 95% linear and nonlinear regression prediction intervals and Bayesian prediction intervals for the six cross-validation cases.

cases. Nonlinear BPI outperforms nonlinear RPI in Cases 1, 3, and 4. Overall, the nonlinear BPI outperforms nonlinear RPI, and the linear RPI has the worst predictive performance.

Evaluation of Computational Cost

Computational cost of evaluating the regression-based nonlinear confidence/prediction intervals and the DREAM-based nonlinear credible/prediction intervals is evaluated using execution time. For the DREAM simulations, the execution time of calculating the BCI/BPI is almost the same (0.5 d) for each of the six cross-validation cases. However, the execution time of calculating the nonlinear RCI/RPI varies significantly for the different cases; the shortest is 1.5 d for Case 6 and the longest is 9.2 d for Case 2. The average computational time for nonlinear RCI/RPI is about 5.8 d, which is much longer than the 0.5 d for BCI/BPI. The reason is that the nonlinear RCI/RPI needs to be evaluated for each individual prediction in an iterative process similar to that of model calibration using the Gauss-Marquardt-Levenberg method (the iterative procedure may be time consuming). When the number of predictions is large (e.g., 54 for each cross-validation case), the computational time may be long. On the contrary, the nonlinear BCI/BPI can be estimated simultaneously for all the predictions based on the DREAM realizations. In addition, the computational time for RCI/RPI heavily depends on the UCODE_2005 parameters TolIntP and TolIntY that control convergence of the iterative process. When the values of the two parameters increase from 10^{-7} to 10^{-3}, the average computational time is reduced to 0.6 d, which is still slightly longer than that of DREAM. However, increasing these tolerance values renders the nonlinear RPI more fluctuating (not shown). As a result, it may not always be the case that MCMC simulations are computationally more expensive than nonlinear regressions in terms of calculating the nonlinear intervals, especially when the number prediction is large.

Conclusions

In vadose zone modeling, parameter estimates and model predictions are inherently uncertain, regardless of quality and quantity of data used in model-data fusion. This study is focused on predictive uncertainty, since accurate quantification of predictive uncertainty is necessary to design data collection systems for improving our understanding of vadose zone processes. In this study, we evaluate two commonly used methods for uncertainty quantification: nonlinear regression and MCMC methods. The former quantifies predictive uncertainty using the RCI, and the latter uses BCI; neither RCI nor BCI includes measurement errors. When measurement errors are considered, the respective methods are RPI and BPI. Relative predictive performance between RCI and BCI and between RPI and BPI is examined through a cross-validation study using predictive logscore as the performance measure. This study also investigates computational efficiency of estimating nonlinear credible intervals using nonlinear regression and MCMC methods. The following is a summary of our key findings in the numerical study:

1. When multiple model parameters are calibrated, although the mean parameter estimates of the nonlinear regression and DREAM are similar, the estimation variance of the nonlinear regression is too large to be reasonable for the three least influential parameters ϕ, S_{lr}, and η. For the DREAM results, although the variance of the parameters is reasonable, their posterior distributions do not differ significantly from their prior distributions, also because of relatively small sensitivity of the parameters. Therefore, parametric uncertainty analysis using either the nonlinear regression or MCMC methods requires conducting sensitivity analyses (global or local) to select the most influential parameters to reduce computational cost.

2. When measurement errors are not considered, the predictive logscore indicates that the nonlinear Bayesian credible interval has the best predictive performance. For all six cross-validation cases, the nonlinear Bayesian credible intervals have the smallest logscore, and the linear confidence intervals have the largest logscore.

3. When measurement errors are considered, the predictive logscore indicates that the nonlinear Bayesian prediction interval has the best predictive performance. For four out of six cross-validation cases, the Bayesian prediction intervals have the smallest logscore. Similarly, the nonlinear regression prediction interval performs better than the linear regression prediction interval.

4. For the numerical experiments of this study, calculating the nonlinear regression intervals is computationally less efficient than calculating the Bayesian intervals, which is different from conclusions reached in other studies. The reason is that the number of predictions is large in this study. Since the nonlinear regression intervals are calculated independently for individual predictions and the calculation for each prediction may be computationally expensive, calculating the nonlinear regression intervals for a large number of predictions may be computationally demanding. Calculation of the Bayesian intervals, however, is different, since they are simultaneously calculated for all predictions. It is worth mentioning that this study is special because the comparison of numerical efficiency is only conducted for three parameters. Computational cost of DREAM will dramatically increase for high-dimensional inverse problems due to the "curse of dimensionality," which may be resolved by using more computationally advanced methods (e.g., Laloy and Vrugt, 2012).

The numerical experiment conducted in this study may bring insight on selecting appropriate methods (e.g., regression and Bayesian) for uncertainty quantification with consideration of computational cost and predictive performance.

Acknowledgments

This work was supported in part by NSF-EAR grant 0911074, DOE-SBR grant DE-SC0002687, and ORAU/ORNL High Performance Computing Grant. The first and last authors are also supported by the National Science Foundation of China (No. 41172206, 40725010 and 41030746). The third author was supported, in part, by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors thank Jasper Vrugt for providing the DREAM code and Mary Hill for discussion on calculation of nonlinear confidence intervals.

References

Abbaspour, K.C., C.A. Johnson, and M.T. van Genuchten. 2004. Estimating uncertain flow and transport parameters using a sequential uncertainty fitting procedure. Vadose Zone J. 3:1340–1352.

Abbaspour, K.C., R. Schulin, E. Schlappi, and H. Fluhler. 1996. A Bayesian approach for incorporating uncertainty and data worth in environmental projects. Environ. Model. Assess. 1:151–158. doi:10.1007/BF01874902.

Adams, B.M., W.J. Bohnhoff, K.R. Dalbey, J.P. Eddy, M.S. Eldred, D.M. Gay, K. Haskell, P.D. Hough, and L.P. Swiler. 2010. DAKOTA, A multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis. Sandia Tech. Rep. SAND2010-2183, December 2009. Sandia Natl. Lab. Albuquerque, NM.

Bates, D.M., and D.G. Watts. 1988. Nonlinear regression analysis and its applications. John Wiley and Sons, New York.

Box, E.P., and G.C. Tiao. 1992. Bayesian inference in statistical analysis. Wiley, 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. doi:10.1029/WR022i002p00199.

Casella, G., and R.L. Berger. 2002. Statistical inference. 2nd ed. Duxbury Press, New York.

Chen, M.H., and Q.M. Shao. 1999. Monte Carlo estimation of Bayesian Credible and HPD intervals, J. Comput. Graphical Stat. 8:69–92.

Christensen, S., and R.L. Cooley. 1999. Evaluation of confidence intervals for a steady-state leaky aquifer model. Adv. Water Resour. 22:807–817. doi:10.1016/S0309-1708(98)00055-4.

Christensen, S., C. Moore, and J. Doherty. 2005. Comparison of stochastic and regression based methods for quantification of predictive uncertainty of model-simulated wellhead protection zones in heterogeneous aquifers. ModelCARE 2005, 5th Int. Conf. on Calibration and Reliability in Groundwater Modelling—From Uncertainty to Decision Making, the Hague (Scheveningen), the Netherlands, 6–9 June 2005. p. 440–447.

<bok>Christensen, S., C. Moore, and J. Doherty. 2006. Comparison of stochastic and regression based methods for quantification of predictive uncertainty of model-simulated wellhead protection zones in heterogeneous aquifers. In: Calibration and reliability in groundwater modelling: From uncertainty to decision making. IAHS Publ. 304:202–208.</bok>

Cooley, R.L. 2004. A theory for modeling ground-water flow in heterogeneous media. Prof. Pap. 1679. U.S. Geol. Surv., Reston, VA.

Cooley, R.L., and R.L. Naff. 1990. Regression modeling of ground-water flow. In: USGS Techniques in Water-Resources Investigations, Book 3, Chapter B4, USGS, Reston, VA.

Cooley, R.L., and A.V. Vecchia. 1987. Calculation of nonlinear confidence and prediction intervals for groundwater flow models. Water Resour. Bull. 23:581–599. doi:10.1111/j.1752-1688.1987.tb00834.x.

Deng, H., M. Ye, M.G. Schaap, and R. Khaleel. 2009. Quantification of uncertainty in pedotransfer function-based parameter estimation for unsaturated flow modeling. Water Resour. Res. 45:W04409. doi:10.1029/2008WR007477.

Doherty, J. 2005. PEST: Software for model-independent parameter estimation, Watermark Numerical Computing, Australia.

Draper, N.R., and H. Smith. 1981. Applied regression analysis. 2nd ed. John Wiley and Sons, New York.

USEPA. 2002. Supplemental guidance for developing soil screening levels for superfund sites. Office of Emergency and Remedial Response, Washington, DC.

Feyen, L., and S.M. Gorelick. 2005. Framework to evaluate the worth of hydraulic conductivity data for optimal groundwater resources management in ecologically sensitive areas. Water Resour. Res. 41:W03019. doi:10.1029/2003WR002901.

Finsterle, S. 1999. iTOUGH2 sample problems. Rep. LBNL-40042. Lawrence Berkeley National Laboratory, Berkeley, CA.

Finsterle, S. 2004. Multiphase inverse modeling: Review and iTOUGH2 applications. Vadose Zone J. 3:747–762.

Finsterle, S. 2007. iTOUGH2 command reference, Rep. LBNL-40041. Lawrence Berkeley National Laboratory, Berkeley, CA.

Finsterle, S., and K. Pruess. 1995. Solving the estimation-identification problem in two-phase flow modeling. Water Resour. Res. 31:913–924. doi:10.1029/94WR03038.

Finsterle, S., and Y. Zhang. 2011. Solving iTOUGH2 simulation and optimization problems using the PEST protocol. Environ. Model. Softw. 26:959–968. doi:10.1016/j.envsoft.2011.02.008.

Gallagher, M., and J. Doherty. 2007. Parameter estimation and uncertainty analysis for a watershed model. Environ. Model. Softw. 22:1000–1020. doi:10.1016/j.envsoft.2006.06.007.

Gelman, A., and D.B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–511. doi:10.1214/ss/1177011136.

Gimmi, T., M. Schneebeli, H. Fluhler, H. Wydler, and T. Baer. 1997. Field-scale water transport in unsaturated crystalline rock. Water Resour. Res. 33:589–598. doi:10.1029/96WR03974.

Good, I.J. 1952. Rational decisions. J. R. Stat. Soc., B 14:107–114.

Hill, M.C., and C.R. Tiedeman. 2007. Effective calibration of ground water models, with analysis of data, sensitivities, predictions, and uncertainty. John Wiley and Sons, New York.

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

Huisman, J.A., J. Rings, J.A. Vrugt, J. Sorg, and H. Vereecken. 2010. Hydraulic properties of a model dike from coupled Bayesian and multi-criteria hydrogeophysical inversion. J. Hydrol. 380:62–73. doi:10.1016/j.jhydrol.2009.10.023.

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

Laloy, E., D. Fasbender, and C.L. Bielders. 2010b. Parameter optimization and uncertainty analysis for plot-scale continuous modeling of runoff using a formal Bayesian approach. J. Hydrol. 380(1–2):82–93. doi:10.1016/j.jhydrol.2009.10.025.

Laloy, E., and J.A. Vrugt. 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{(ZS)}$ and high-performance computing. Water Resour. Res. 48:W01526. doi:10.1029/2011WR010608.

Laloy, E., M. Weynants, C.L. Bielders, M. Vanclooster, and M. Javaux. 2010a. How efficient are one-dimensional models to reproduce the hydrodynamic behavior of structured soils subjected to multi-step outflow experiments? J. Hydrol. 393:37–52. doi:10.1016/j.jhydrol.2010.02.017.

Liu, S.Y., and T.-C.J. Yeh. 2004. An integrative approach for monitoring water movement in the vadose zone. Vadose Zone J. 3:681–692.

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

Lu, D., M. Ye, and S.P. Neuman. 2012. Dependence of Bayesian model selection criteria and Fisher information matrix on sample size. Math. Geosci. 43(8):971–993. doi:10.1007/s11004-011-9359-0.

Luckner, L., M.Th. van Genuchten, and D. Nielsen. 1989. A consistent set of parametric models for the two-phase flow of immiscible fluids in the subsurface. Water Resour. Res. 25:2187–2193. doi:10.1029/WR025i010p02187.

Marshall, L., D. Nott, and A. Sharma. 2004. A comparative study of Markov Chain Monte Carlo methods for conceptual rainfall-runoff modeling. Water Resour. Res. 40:W02501. doi:10.1029/2003WR002378.

McClave, J.T., and T. Sincich. 2000. Statistics. 8th ed. Prentice Hall, Englewood Cliffs, NJ.

Meyer, P.D., M.L. Rockhold, and G.W. Gee. 1997. Uncertainty analysis of infiltration and subsurface flow and transport for SDMP sites. NUREG/CR-6565, PNNL-11705. U.S. Nucl. Regul. Commiss., Office of Nucl. Regul. Res., Washington, D.C.

Minasny, B., and D.J. Field. 2005. Estimating soil hydraulic properties and their uncertainty: The use of stochastic simulation in the inverse modeling of the evaporation method. Geoderma 126:277–290. doi:10.1016/j.geoderma.2004.09.015.

Morris, M.D. 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 21:239–245.

Neuman, S.P., L. Xue, M. Ye, and D. Lu. 2011. Bayesian analysis of data-worth considering model and parameter uncertainty. Adv. Water Resour. 36:75–85. doi:10.1016/j.advwatres.2011.02.007.

Nowak, W., F.P.J. de Barros, and Y. Rubin. 2010. Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain. Water Resour. Res. 46:W03535. doi:10.1029/2009WR008312.

Pan, F., M. Ye, J.T. Zhu, Y.S. Wu, B.X. Hu, and Z.B. Yu. 2009a. Numerical evaluation of uncertainty in water retention parameters and effect on predictive uncertainty. Vadose Zone J. 8:158–166. doi:10.2136/vzj2008.0092.

Pan, F., M. Ye, J. Zhu, Y.S. Wu, B. Hu, and Z. Yu. 2009b. Incorporating layer- and local-scale heterogeneities in numerical simulation of unsaturated flow and tracer transport. J. Contam. Hydrol. 103(3–4):194–205. doi:10.1016/j.jconhyd.2008.10.012.

Pan, F., J.T. Zhu, M. Ye, Y.A. Pachepsky, and Y.S. Wu. 2011. Sensitivity analysis of unsaturated flow and contaminant transport with correlated parameters. J. Hydrol. 397:238–249. doi:10.1016/j.jhydrol.2010.11.045.

Poeter, E., M. Hill, E. Banta, S. Mehl, and S. Christensen. 2005. UCODE2005 and six other computer codes for universal sensitivity analysis, calibration, and uncertainty evaluation. U.S. Geological Survey, Reston, VA.

Pruess, K., C. Oldenburg, and G. Moridis. 1999. TOUGH2 user's guide, version 2.0, Rep. LBNL-43134, Lawrence Berkeley Natl. Lab., Berkeley, CA.

Schaap, M.G., F.J. Leij, and M.Th. van Genuchten. 2001. Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. 251:163–176. doi:10.1016/S0022-1694(01)00466-8.

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.

Seber, G.A.G., and C.J. Wild. 2003. Nonlinear regression. John Wiley and Sons, New York.

Shi, X., M. Ye, G. Curtis, et al. 2012. Assessment of parametric uncertainty for surface complexation models of uranium reactive transport modeling. Water Resour. Res.

ter Braak, C.J.F. 2006. A Markov Chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces. Stat Comput. 16:239–249.

van Genuchten, M.Th. 1980. A closed form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 44:892–898. doi:10.2136/sssaj1980.03615995004400050002x.

Vecchia, A.V., and R.L. Cooley. 1987. Simultaneous confidence and prediction intervals for nonlinear-regression models with application to a groundwater-flow model. Water Resour. Res. 23:1237–1250. doi:10.1029/WR023i007p01237.

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. doi:10.1111/1467-9876.00082

Vrugt, J.A., and W. Bouten. 2002. Validity of first-order approximations to describe parameter uncertainty in soil hydrologic models. Soil Sci. Soc. Am. J. 66:1740–1751. doi:10.2136/sssaj2002.1740

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

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

Vrugt, J.A., C.J.F. ter Braak, C.G.H. Diks, B.A. Robinson, J.M. Hyman, and D. Higdon. 2009c. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int. J. Nonlinear Sci. Numer. Simul. 10:273–290. doi:10.1515/IJNSNS.2009.10.3.273.

Vrugt, J.A., C.J.F. ter Braak, H.V. Gupta, and B.A. Robinson. 2009a. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stochastic Environ. Res. Risk Assess. 23(7):1011–1026. doi:10.1007/s00477-008-0274-y.

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

Wohling, T., and J.A. Vrugt. 2011. Multiresponse multilayer vadose zone model calibration using Markov chain Monte Carlo simulation and field water retention data. Water Resour. Res. 47:W04510. doi:10.1029/2010WR009265.

Yang, J., P. Reichert, K.C. Abbaspour, J. Xia, and H. Yang. 2008. Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. J. Hydrol. 358:1–23.

Ye, M., and R. Khaleel. 2008. A Markov chain model for characterizing medium heterogeneity and sediment layering structure. Water Resour. Res. 44:W09427. doi:10.1029/2008WR006924.

Ye, M., R. Khaleel, M.G. Schaap, and J. Zhu. 2007b. Simulation of field injection experiments in heterogeneous unsaturated media using cokriging and artificial neural network. Water Resour. Res. 43:W07413, doi:10.1029/2006WR005030.

Ye, M., P.D. Meyer, and S.P. Neuman. 2008. On model selection criteria in multimodel analysis. Water Resour. Res. 44:W03428. doi:10.1029/2008WR006803.

Ye, M., S.P. Neuman, 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., F. Pan, Y.S. Wu, B. Hu, C. Shirley, and Z. Yu. 2007a. Assessment of radionuclide transport uncertainty in the unsaturated zone at Yucca Mountain. Adv. Water Resour. 30:118–134.

Yeh, T.C.J., and J. Simunek. 2002. Stochastic fusion of information for characterizing and monitoring the vadose zone. Vadose Zone J. 1:207–221.