Impacts of prior parameter distributions on Bayesian evaluation of groundwater model complexity

Saeideh Samani a,b, Ming Ye b,c,d,*, Fan Zhang e, Yong-zhen Pei c, Guo-ping Tang f, Ahmed Elshall b, Asghar A. Moghaddam a

a Department of Earth Sciences, University of Tabriz, Tabriz 5166616471, Iran
b Department of Scientific Computing, Florida State University, Tallahassee, FL 32306, USA
c School of Computer Science and Software Engineering, Tianjin Polytechnic University, Tianjin 300387, China
d Department of Earth, Ocean and Atmospheric Science, Florida State University, Tallahassee, FL 32306, USA
e Key Laboratory of Tibetan Environment Changes and Land Surface Processes, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China
f Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Received 14 September 2017; accepted 10 January 2018
Available online 22 June 2018

Abstract

This study used the marginal likelihood and Bayesian posterior model probability for evaluation of model complexity in order to avoid using over-complex models for numerical simulations. It focused on investigation of the impacts of prior parameter distributions (involved in calculating the marginal likelihood) on the evaluation of model complexity. We argue that prior parameter distributions should define the parameter space in which numerical simulations are made. New perspectives on the prior parameter distribution and posterior model probability were demonstrated in an example of groundwater solute transport modeling with four models, each simulating four column experiments. The models had different levels of complexity in terms of their model structures and numbers of calibrated parameters. The posterior model probability was evaluated for four cases with different prior parameter distributions. While the distributions substantially impacted model ranking, the model ranking in each case was reasonable for the specific circumstances in which numerical simulations were made. For evaluation of model complexity, it is thus necessary to determine the parameter spaces for modeling, which can be done by conducting numerical simulation and using engineering judgment based on understanding of the system being studied.

Keywords: Marginal likelihood; Posterior model probability; Advection-dispersion equation; Mobile-immobile model; Groundwater model

1. Introduction

While modelers have tended to avoid using over-complex models in groundwater and other fields of numerical modeling, identifying the appropriate level of model complexity is always challenging, as it involves many factors, such as complexity of the system, current understanding of system behaviors, quality and quantity of system observations, objectives of modeling projects, cost of model development, and mathematical/statistical analysis needed for identification (Clement, 2011; Hunt and Zheng, 1999; Simmons and Hunt, 2012). Model complexity can be explored in various ways by investigating the relation between model inputs and outputs (Jakeman and Hornberger, 1993; Hill, 2006; Hunt et al., 2007; Gómez-Hernández, 2006), evaluating variability of model outputs (Young et al., 1996; Arkesteijn and Pande, 2013), and/or examining model predictive performance (Brooks and

---

This work was supported by the U.S. Department of Energy Early Career Research Program Award (Grant No. DE-SC0008272) and U.S. National Science Foundation (Grant No. 1552329).

* Corresponding author.
E-mail address: mye@fsu.edu (Ming Ye).
Peer review under responsibility of Hohai University.

https://doi.org/10.1016/j.wse.2018.06.001

1674-2370/© 2018 Hohai University. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
A common practice in groundwater modeling is to use model selection criteria, e.g., the Akaike information criterion (AIC) and Bayesian information criterion (BIC), to select the best model from a set of models of different levels of complexity (Dai et al., 2012; Elshall and Tsai, 2014; Engelhardt et al., 2014; Massoudieh et al., 2013; Ye et al., 2008a, 2016); the selected model is considered to have the appropriate level of model complexity. All the model selection criteria favor models with high goodness-of-fit to data, and disfavor the models with unnecessary complexity by including penalty terms to create a balance between the model’s data-fitting ability and its complexity. Many of the penalty terms are a function of the number of model parameters. However, evaluating model complexity by only examining the number of calibrated parameters may inadequately quantify model complexity, because model complexity may be affected by other factors, such as parameter regions and model structures.

This study explored the ways in which the prior parameter distribution affects identification of model complexity. The prior parameter distribution contains more information than parameter numbers, as the distribution defines a parameter region in parameter space. As explained below, the size of the prior parameter region is important to evaluating model complexity, though this has not received enough attention from researchers. In this study, we used the prior parameter distribution to evaluate the posterior model probability, and used the posterior model probability as a measure of model complexity. Among a set of models, the model with an appropriate level of model complexity is expected to have the highest model probability. For a set of models $M_1, M_2, \ldots, M_K$ with $K$ alternative models and a dataset $D$, the posterior probability for model $M_k$ ($k = 1, 2, \ldots, K$) is given by Bayes’ theorem (Hoeting et al., 1999):

$$ p(M_k | D) = \frac{p(D | M_k) p(M_k)}{\sum_{k=1}^{K} p(D | M_k) p(M_k)} $$

(1)

$$ p(D | M_k) = \int p(D | \theta_k, M_k) p(\theta_k | M_k) d\theta_k $$

(2)

where $p(D | M_k)$ is the marginal likelihood (also known as the integrated likelihood or Bayesian evidence) of model $M_k$; $\theta_k$ is the vector of parameters associated with model $M_k$; $p(\theta_k | M_k)$ is the prior probability density of $\theta_k$ in model $M_k$; $p(D | \theta_k, M_k)$ is the joint likelihood of model $M_k$ and its parameters $\theta_k$; and $p(M_k)$ is the prior probability of $M_k$ as a correct model. In this study, the prior model probability was set as $1/K$, implying that all models were equally likely based on prior information before the dataset $D$ was used for evaluating the posterior model probability. Other ways of estimating the prior model probability are available (e.g., Elshall and Tsai, 2014; Ye et al., 2005, 2008b), but were not considered in this study. For the equal prior model probability, the posterior model probability was determined solely by the marginal likelihood. Therefore, the marginal likelihood was used to evaluate model complexity (e.g., Marshall et al., 2005; Schöninger et al., 2015; Schoups et al., 2008; Schoups and Vrugt, 2011).

A simple example illustrates that the marginal likelihood, $p(D | M_k)$, can be used as a measure of model complexity. For evaluating $p(D | M_k)$, the $p(D | \theta_k, M_k)$ term in Eq. (2) is the probability that model $M_k$ and its parameter $\theta_k$ can simulate data $D$, and a complex model has a higher potential than a simple model to achieve higher $p(D | \theta_k, M_k)$ values at the optimum values of $\theta_k$. However, the complex model is penalized by the prior probability density, $p(\theta_k | M_k)$, for two reasons. First, the parameter region defined by the prior parameter distribution, generally speaking, is larger for a complex model than for a simple model, because a complex model always has more parameters than a simple model. In addition, within the parameter region, while the joint likelihood is high at the point of optimum parameter values, it may be low at other points. This is illustrated with two models, $M_0$ and $M_1$, as an example. It is assumed that $M_0$ has one parameter that follows the uniform distribution, $U[0, 1]$, and that $M_1$ has two independent parameters that both follow $U[0, 1]$. Model $M_1$ is more complex than $M_0$ in terms of parameter numbers. It is further assumed that $p(D | \theta_k, M_0)$ is a constant $L_0$ over the parameter range of $[0.25, 0.75]$ and zero elsewhere, which leads to the marginal likelihood of $p(D | M_0) = L_0/2$. Similarly, it is assumed that $p(D | \theta_k, M_1)$ is $L_1$ over the parameter range of $[0.25, 0.75] \times [0.25, 0.75]$ and zero elsewhere, which leads to the marginal likelihood of $p(D | M_1) = L_1/4$. The complex model, $M_1$, is favored by the marginal likelihood only when $L_1 > 2L_0$, i.e., the likelihood that model $M_1$ reproduces the data is twice as large as the likelihood that model $M_0$ does. Otherwise, the complexity of model $M_1$ is penalized by the prior parameter distribution. This example can be extended with more models. For example, if model $M_2$ has three independent parameters (assuming that its prior parameter distributions and joint likelihood are similar to those of model $M_1$ but with an additional dimension), its marginal likelihood is $L_2/8$. This complex model is favored over $M_0$ and $M_1$ only when $M_2$ can simulate the data four times as accurately as model $M_1$. However, it should be noted that real-world problems are always more complicated, with irregular shapes of likelihood surfaces. Therefore, analytical expressions of the marginal likelihood are unavailable, and numerical evaluation of the marginal likelihood is necessary.

This study aimed at investigating the ways in which prior parameter distributions impact model complexity, a subject that has not been reported upon in groundwater study. The impacts can be observed using the process described above, by changing the prior parameter distribution of model $M_0$ from $U[0, 1]$ to $U[0, 2]$, and that of model $M_1$ from $U[0, 1] \times U[0, 1]$ to $U[0, 2] \times U[0, 2]$. As a result, $p(D | M_0) = L_0/4$ and $p(D | M_1) = L_1/16$. The condition in which $M_1$ is favored by the marginal likelihood is then $L_1 > 4L_0$, in contrast to the condition of $L_1 > 2L_0$ above. This change is not surprising, because enlarged prior parameter distributions indicate that simulations will be made in a larger parameter space that is believed to be reasonable by modelers. A complex
model is justified only when the complex model gives an overall better prediction than a simple model over the entire parameter space. The superiority of the marginal likelihood for evaluating model complexity lies in the fact that it takes into account prior parameter distributions in characterizing the parameters, as well as in model predictions.

In comparison with the marginal likelihood, model selection criteria do not consider prior parameter distributions. For example, three widely used criteria, \( AIC, AICC, \) and \( BIC \), are defined for model \( M_k \) as follows:

\[
AIC_k = -2 \ln \left( p(D | \hat{\theta}_k, M_k) \right) + 2N_k
\]

\[
AICC_k = -2 \ln \left( p(D | \hat{\theta}_k, M_k) \right) + 2N_k \frac{2N_k(N_k + 1)}{N - N_k - 1}
\]

\[
BIC_k = -2 \ln \left( p(D | \hat{\theta}_k, M_k) \right) + N_k \ln N
\]

where \( \hat{\theta}_k \) is the maximum likelihood estimate of \( \theta_k \), \( N_k \) is the dimension of \( \theta_k \), and \( N \) is the dimension of \( D \). None of these equations includes parameter distributions. The example above shows the difference between \( p(D | M_k) \) and model selection criteria for evaluating model complexity. Taking \( AIC \) as an example, its values for the models \( M_0 \) and \( M_1 \) discussed above are \( AIC_0 = -2 \ln L_0 + 2 \) and \( AIC_1 = -2 \ln L_1 + 4 \), respectively. The requirement for the complex model, \( M_1 \), to be favored by \( AIC \) is that \( L_1 > e^{L_0} \) regardless of prior parameter distributions. This indicates that, as long as the requirement is satisfied, the complex model, \( M_1 \), is favored over the simple model, \( M_0 \), regardless of the circumstances (with respect to model parameters) in which model predictions will be made. This does not appear to be reasonable for evaluating model complexity.

\( KIC \) is the only criterion that considers prior parameter distributions. \( KIC \) for model \( M_k \) is defined as follows (Kashyap, 1982):

\[
KIC_k = -2 \ln p(D | \hat{\theta}_k, M_k) - 2 \ln p(\hat{\theta}_k | M_k) + N_k \ln N + N_k \ln N/m
\]

where \( p(D | \hat{\theta}_k, M_k) \) is the observed Fisher information matrix that addresses estimation uncertainty of \( \hat{\theta}_k \) (Ye et al., 2010). However, the literature on \( KIC \) has not discussed how to use prior parameter distributions in the context of evaluating model complexity. In addition, it should be noted that \( BIC \) and \( KIC \) are two different variants of the Laplace approximation of \( -2 \ln p(D | M_k) \) (Ye et al., 2008a, 2010). The Laplace approximation becomes unnecessary when the marginal likelihood is evaluated using Monte Carlo (MC) approaches. The MC approaches take into account the prior parameter distribution. Using the simple MC approach as an example, parameter samples are generated from the prior parameter distribution, and the arithmetic mean estimator (AME) of the marginal likelihood function, \( p_{AME}(D | M_k) \), is evaluated as follows:

\[
p_{AME}(D | M_k) = \frac{\sum_{i=1}^{m} p(D | \theta_{prior,i}, M_k)}{m}
\]

where \( \theta_{prior,i} \) is the \( i \)th parameter sample generated from the prior distribution of parameter \( \theta_k \) of model \( M_k \), and \( m \) is the number of samples. More advanced MC approaches for evaluating the marginal likelihood can be referred to in Schöning et al. (2014), Liu et al. (2016), and Zeng et al. (2016a, 2016b). Ye et al. (2010) showed that \( KIC \) considers parameters in Gaussian space, whose center is at \( \hat{\theta}_k \), and its size is determined by the inverse of the Fisher information matrix. Both \( KIC \) and the simple MC method were used in this study to calculate the posterior model probability.

This study has developed a new perspective on the use of prior parameter distribution and posterior model probability for evaluating model complexity. We demonstrate below that prior parameter distributions significantly affect the calculation of model probabilities and thus the evaluation of model complexity. The use of non-informative priors (e.g., a uniform distribution with a wide parameter range), the standard practice in Bayesian parameter estimation, is not appropriate for evaluating model complexity, because such priors do not reflect the circumstances of intended model use. On the other hand, it is also inappropriate to evaluate model complexity at a fixed parameter value (e.g., the calibrated parameter value), because it implies that one knows the exact circumstances of intended model use (in terms of parameters) without any uncertainty, which is rarely the case in practice. Therefore, to evaluate model complexity using the posterior model probability, it is necessary to investigate the appropriate prior parameter distributions compatible with the intended model use. This new perspective on the prior parameter distribution is demonstrated below using four models developed to simulate column experiments of contaminant transport.

2. Research methods

This section describes the column experiments and the rankings of the four models in four cases with different prior parameter distributions. The column experiments used in this study were conducted by Anamosa et al. (1990) to understand how to describe macropore flow and immobile water regions in solute transport modeling. Two soil columns, 71.6 cm and 68.9 cm in length, respectively, were filled with soils taken from the Western Province of Cameroon. The soils were clayey-skeletal, oxidic isohyperthermic, and typic Gibs-siorthox. The breakthrough curves of tritiated water were measured for six experiments with different flow velocities, and the data of four experiments were used in this study. The measurements corresponding to the velocities of 111 cm/d, 36.7 cm/d, 2.71 cm/d, and 2.69 cm/d, respectively, were digitized by Tang et al. (2009), and used as the dataset \( D \) in this study. Two experiments, with flow velocities of 111 cm/d and 2.71 cm/d,
respectively, were conducted in column I, with a length of 71.6 cm, and the other two experiments, with flow velocities of 36.7 cm/d and 2.69 cm/d, respectively, were conducted in column II, with a length of 68.9 cm. For the convenience of discussion, the four experiments, with the flow velocities of 111 cm/d, 2.71 cm/d, 36.7 cm/d, and 2.69 cm/d are denoted as Q1, Q2, Q3, and Q4, respectively, in the analysis below. More details about the experiments can be found in Anamosa et al. (1990).

2.1. Four alternative models

The advection-dispersion equation (ADE) and mobile-immobile (MIM) model were used in this study to simulate the four breakthrough curves. Based on Tang et al. (2009), the dimensionless ADE model is

\[ \frac{\partial C}{\partial T} = \frac{1}{Pe} \frac{\partial^2 C}{\partial Z^2} - \frac{\partial C}{\partial Z} \]  

where \( R \) is the retardation factor, with \( R = 1 + \rho_b K_d / \theta; \) \( C \) is the dimensionless (normalized) concentration, with \( C = c/c_0; \) \( Z \) is the dimensionless distance, with \( Z = z/L; \) \( T \) is the dimensionless time, with \( T = vt/L; \) and \( Pe \) is the Peclet number. The variables used to define the dimensionless quantities are as follows: \( \rho_b \) is the bulk density (g/cm³); \( K_d \) is the absorption coefficient (mL/g); \( \theta \) is the volumetric water content (cm³/cm³); \( c \) and \( c_0 \) are the solute concentrations in the column and at the column inlet boundary during a pulse injection (Bq/mL), respectively; \( z \) is the distance from the inlet boundary (cm) downward along the column; \( L \) is the column length (cm); \( v \) is the average pore water velocity (cm/d); \( t \) is time (d); and \( \lambda \) is the dispersivity (cm). The initial condition is \( C(0,0) = 0 \). Appropriate inlet and outlet boundary conditions were described by Parker and van Genuchten (1984) and Toride et al. (1995).

The dimensionless MIM model is

\[ \beta R \frac{\partial C_1}{\partial T} = \frac{1}{Pe} \frac{\partial^2 C_1}{\partial Z^2} - \omega (C_1 - C_2) \]  

\[ (1 - \beta) R \frac{\partial C_2}{\partial T} = \omega (C_1 - C_2) \]  

where \( C_1 \) and \( C_2 \) are the dimensionless (normalized) concentrations in the mobile and immobile regions, respectively; \( \beta \) is the dimensionless fraction of solute in the mobile region, with \( \beta = (\theta_m + f \rho_b K_d) / (\theta + \rho_b K_d) \), where \( \theta_m \) is the mobile pore water content, and \( f \) is the fraction of the adsorption sites under equilibrium of solute in mobile pore water; and \( \omega \) is the dimensionless mass transfer coefficient, with \( \omega = \alpha L/\theta v \), where \( \alpha \) is the mass transfer coefficient between mobile and immobile pore water. Based on Tang et al. (2009), the retardation factor, \( R \), was fixed at 1.05, so that it would be consistent with the original analysis of Anamosa et al. (1990). Because the MIM model is non-equilibrium and considers mobile and immobile regions, its model structure is more complex than that of the equilibrium ADE model.

This study considered the following four models (two variants of ADE and two variants of MIM): (1) ADE1 with one calibrated parameter, i.e., the dispersivity; (2) ADE2 with two calibrated parameters, i.e., the dispersivity and volumetric water content; (3) MIM1 with three calibrated parameters, i.e., the dispersivity, dimensionless fraction of solute in the mobile region, and dimensionless mass transfer coefficient; and (4) MIM2 with four calibrated parameters, i.e., the dispersivity, dimensionless fraction of solute in the mobile region, dimensionless mass transfer coefficient, and duration of pulse injection.

For models MIM1 and MIM2, the volumetric water content was not explicitly calibrated, because the models separated the pore space into mobile and immobile zones. The volumetric water content was implicitly calibrated through the calibration of parameters \( \beta \) and \( \omega \). Given the model structures and numbers of calibrated parameters, intuitively speaking, the order of model complexity, from the simplest to the most complex, was ADE1, ADE2, MIM1, and MIM2. With respect to model complexity, the following three questions arise:

1. Will the increasing level of model complexity be justified by the breakthrough curve data?
2. Will the structurally complex MIM models outperform the structurally simple ADE models in simulating the data?
3. Will a model with more calibrated parameters outperform its counterpart (e.g., ADE1 vs. ADE2, and MIM1 vs. MIM2) in simulating the data?

The three questions were explored in this study in a statistical way based on the posterior probabilities of the four models.

2.2. Four cases of model ranking

The four models were ranked for the following four cases: Case 1 was based on the root mean square error (RMSE) of the calibrated models; Case 2 was based on the posterior model probability evaluated using \( KIC \); and cases 3 and 4 were based on the posterior model probability evaluated using the MC method via Eq. (7) with the two sets of different parameter ranges listed in Table 1. It should be noted that the parameter ranges for cases 3 and 4 were chosen only for demonstrating the impacts of prior parameter distributions on the calculated posterior model probability.

The four cases were designed to examine to what extent the prior parameter distributions affect model ranking. Case 1 represents a special case in which prior information is not used; Case 2 does not consider prior information but considers uncertainty in parameter estimates; and cases 3 and 4 consider prior information in a true Bayesian manner. The model rankings for the four cases shown below demonstrate the impacts of the prior information on the evaluation of model complexity. The results suggest the importance of choosing proper prior information for Bayesian evaluation of model complexity.
The four cases correspond to three ways of evaluating the marginal likelihood for model ranking. For Case 1, by virtue of the mean integral value theorem, we have
\[
p(D|M_k) = \int p(D|\theta_k, M_k) p(\theta_k|M_k) d\theta_k = p(D|\xi_k, M_k) \int p(\theta_k|M_k) d\theta_k = p(D|\xi_k, M_k) \tag{11}
\]
where \(\xi_k\) is an unknown value of \(\theta_k\). Replacing \(\xi_k\) with \(\hat{\theta}_k\) gives \(p(D|M_k) \approx p(D|\hat{\theta}_k)\). If \(p(D|\theta_k, M_k)\) follows the Gaussian distribution, \(RMSE = -2\ln p(D|\hat{\theta}_k, M_k)\) (Hill and Tiedeman, 2007). For Case 2, Ye et al. (2008a) showed that \(KIC = -2\ln p(D|M_k)\) and proposed
\[
p(D|M_k) \approx p(D|\hat{\theta}_k, M_k) \exp \left[ -\frac{1}{2} (\theta_k - \hat{\theta}_k)^T \mathbf{F}_k (\hat{\theta}_k, M_k) (\theta_k - \hat{\theta}_k) \right] d\theta_k \tag{12}
\]
This equation indicates that \(KIC\) is not evaluated at \(\hat{\theta}_k\), but over a Gaussian parameter space defined by the integral on the right-hand side of Eq. (12). In other words, the Gaussian space is centered at \(\hat{\theta}_k\), and its size is determined by the inverse of the Fisher information matrix (Ye et al., 2010; Lu et al., 2011). The Gaussian space is a sub-space of the prior parameter space, and this is the difference between Case 2 and cases 3 and 4. Therefore, the four cases can be viewed as model ranking based on gradually expanded parameter space, starting from the point \(\hat{\theta}_k\), extending to the Gaussian sub-space, and finally proceeding to the entire parameter space defined by the prior parameter distributions.

The four cases correspond to three ways of evaluating model complexity. In Case 1, the evaluation was made using the calibrated parameter values but ignoring the parametric uncertainty. This implies that one knows exactly (with respect to parameters) how the models will be used for numerical modeling, which is rarely the case in practice as parametric uncertainty always exists. Case 2 considered the parametric uncertainty in the Gaussian parameter space centered at the calibrated parameter values, and was more realistic than Case 1. Cases 3 and 4 considered the parametric uncertainty in different spaces with large parameter ranges defined by the prior parameter distributions. If the space is large, it implies that one expects to use the models with a wide range of parameter values for numerical simulation.

### 3. Results

This section presents the model rankings for the four cases. The model rankings are based on the posterior model probability.

#### 3.1. Model ranking based on RMSE (Case 1)

Table 2 lists the calibrated parameter values and corresponding estimation errors for the four experiments obtained using the software CXTFIT/Excel developed by Tang et al. (2010). Fig. 1 plots the measured and simulated breakthrough curves using the calibrated parameters, where the effluent pore volume is a measure of experimental time, i.e., the time of water flushing in the pore space of a sand-filled column. While all four models can satisfactorily simulate the breakthrough curves (Fig. 1), the \(RMSE\) values indicate that, except for experiment Q4 conducted in column II, the model ranking is consistent with the level of model complexity discussed above. For example, the most complex model (MIM2 with four calibrated parameters) and the simplest model (ADE1 with one calibrated parameter) are ranked as the best and the worst, respectively. The structurally complex MIM models outperform the structurally simple ADE models. For the models with the same structure, calibrating more parameters improves model fitting, although the improvement is relatively smaller for the MIM models than for the ADE models. These findings are not surprising, as complex models have greater potential than simple models to achieve a better model fitting. However, this does not mean that a complex model gives better simulations of the data when numerical simulations are made with larger parameter ranges.

#### 3.2. Model ranking based on KIC (Case 2)

Table 3 lists the posterior model probability evaluated using KIC. The \(p(D|\theta_k, M_k)\) term took the Gaussian distribution form with the mean being 0 and the standard deviation being 1. When evaluating the model selection criteria, the time series analysis method described in Lu et al. (2013) was used to determine the covariance matrix needed to evaluate the \(p(D|\theta_k, M_k)\) term so that residual temporal correlation could be taken into account for model ranking. The analysis shows that the correlation is weak, mainly due to the satisfactory model fitting to the data (Fig. 1).

### Table 2
Calibrated parameter values and their ranges for cases 3 and 4.

<table>
<thead>
<tr>
<th>Model</th>
<th>(\lambda) (cm)</th>
<th>Parameter range</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(\theta) (cm(^3)/cm(^3))</td>
<td>(\beta)</td>
<td>(\omega)</td>
</tr>
<tr>
<td>ADE1</td>
<td>(1, 200)</td>
<td>(0.1, 0.99)</td>
<td>(0.1, 10)</td>
<td>(0.1, 0.99)</td>
</tr>
<tr>
<td>ADE2</td>
<td>(1, 200)</td>
<td>(0.1, 0.99)</td>
<td>(0.1, 10)</td>
<td>(0.1, 0.99)</td>
</tr>
<tr>
<td>MIM1</td>
<td>(1, 200)</td>
<td>(0.1, 0.99)</td>
<td>(0.1, 10)</td>
<td>(0.1, 0.99)</td>
</tr>
<tr>
<td>MIM2</td>
<td>(1, 200)</td>
<td>(0.1, 0.99)</td>
<td>(0.1, 10)</td>
<td>(0.1, 0.99)</td>
</tr>
</tbody>
</table>

Note: * means that the value of pore volume is used to reflect the magnitude of the duration of pulse injection.
Table 3 shows that, for the experiments with the high velocities of 111 cm/d in Q1 and 36.7 cm/d in Q3, the ADE models are considered implausible by $KIC$, because their $KIC$-based posterior probabilities are zero. This agrees with the conclusion drawn by Anamosa et al. (1990) that the equilibrium ADE models cannot be used to simulate the non-equilibrium experiments. The reasons given by Anamosa et al. (1990) are that, when the velocity is high, the residence time of tritiated water in the soil columns is short, and the immobile-water region cannot reach physical equilibrium with the mobile effluent. In line with this, low velocity allows physical equilibrium between the mobile and immobile regions due to dispersion. This may explain why the two ADE models receive a total of 25% probability in Q4, with the lower velocity of 2.69 cm/d. However, the ADE models receive only 0.1% probability in Q2, with the lower velocity of 2.71 cm/d. This may be attributed to the difference between the two soil columns. Anamosa et al. (1990) pointed out that column I has slightly less bulk density and porosity than column II. As a summary, the experimental data justify the structural complexity of the MIM models.

The $KIC$-based posterior probabilities of models MIM1 and MIM2 indicate that, for all the experiments, model MIM1 is substantially more plausible than model MIM2. In particular, the posterior probability of model MIM2 is less than 5% for experiments Q1, Q2, and Q4, suggesting that the complexity of model MIM2 is not justified by the data. This may be because that the pulse duration was accurately measured by Anamosa et al. (1990). Considering that the calibrated pulse duration values listed in Table 2 are almost identical to the measured values reported in Anamosa et al. (1990), it is unnecessary to calibrate the pulse duration in model MIM2.

### 3.3. Model ranking based on prior parameter distributions (cases 3 and 4)

We assume that the calibrated parameters are independent and that each parameter follows a uniform distribution.

![Fig. 1. Measured and calibrated breakthrough curves of four models for four experiments.](image-url)
Table 1 lists two different ranges of the uniform distributions. The ranges are relatively large, in order to include the calibrated values listed in Table 2. On the other hand, the large ranges correspond to circumstances in which one does not know the parameter values of models for numerical simulations. In cases 3 and 4, the marginal likelihood function was evaluated using the simple MC method via Eq. (7), and three million parameter samples were used for the calculation. The results are denoted as \( p_{\text{AME}} \) and listed in Table 3. The model ranking in Case 3 is entirely different from that in Case 1 but consistent with that in Case 2, in that model MIM1 is the best for all four experiments. Comparison of the model rankings in cases 3 and 4 indicates that changing the prior parameter ranges changes the posterior model probability. While model MIM1 is ranked as the best model in Case 3 consistently for all four experiments, this is not the case in Case 4. For example, model MIM1 is ranked as the best model in experiments Q1 and Q2, whereas model ADE1 is ranked as the best model in the other two experiments in Case 4. The implication of this change is discussed below.

### 4. Discussion

The Gaussian parameter distributions involved in Case 2 and the uniform parameter distributions involved in Cases 3 and 4 were examined. Fig. 2 plots the probability density function (PDF) of dispersivity \( (\lambda) \) estimated in Case 2 as an example, calibrated for all four models. Since the uniform parameter distributions in Case 4 are the same as those in Case 3, they are not plotted in Fig. 2. While the prior parameter distribution, U(1, 200), is the same for the four experiments, the Gaussian distributions are different for the four experiments, as the distributions are obtained from model calibration. As expected, the Gaussian distributions are significantly narrower than the prior distribution. This is especially the case for models ADE1 and ADE2 in experiments Q2 and Q4 (Fig. 2(b) and (d)), which is consistent with the small parameter estimation errors listed in Table 2. The Gaussian distributions in Case 2 and prior distributions in Case 3 are used for the discussion below. It is noted that, based on the Gaussian distributions, negative \( \lambda \) values may result, due to the use of the inverse of the Fisher information matrix as an approximation of the covariance matrix of the Gaussian distribution. While the negative values are physically unreasonable, using the Gaussian distributions should not affect the discussion below.

#### 4.1. Analysis of different model rankings in four cases

In the context of evaluating model complexity, the statistical meaning of model ranking is different in the four cases. In Case 1, the RMSE-based model ranking indicates that the most complex model is the best model, because the parametric uncertainty is not a concern in this case, as indicated by Eq. (11). In other words, if the parameter values of models used for numerical simulations are known exactly, the more complex model provides better simulations of the data. In Case 2, the narrow parameter ranges of the Gaussian distributions imply that the model parameters used for numerical simulation are subject to a low degree of uncertainty. Since the marginal likelihood in Case 2 is estimated by integrating the joint likelihood with respect to the Gaussian parameter distribution (Eq. (12)), the joint likelihood is important to the estimation of the marginal likelihood due to the small parameter ranges. Since models MIM1 and MIM2 provide better simulations of the data than models ADE1 and ADE2 do, the two former models have greater marginal likelihood than the latter two models (except for Q4), and are thus favored according to the marginal likelihood and the posterior model probability. The reasons that MIM1 is favored over MIM2 are discussed later. In cases 3 and 4, the wide parameter ranges of the prior distributions of models mean that the parameters used for numerical simulations are largely unknown, and the large variability in parameter values leads to a large variability in numerical simulations. This affects the posterior
model probability, which leads to model ranking different from that in Cases 1 and 2. Therefore, as indicated by Eq. (7), the prior parameter distribution is critical to the evaluation of the posterior model probability and subsequently to the evaluation of model complexity. Taking model MIM1 as an example, comparing the parameter ranges of the model in Cases 3 and 4 (Table 1) shows that the only difference between the two cases is that the range of parameter $\omega$ is smaller in Case 3 than in Case 4. As a result, the posterior probability of model MIM1 is larger in Case 3 than in Case 4. If the parameter range is further reduced, the posterior probability of model MIM1 will further increase, as shown in Case 2.

### 4.2. Comparison of model MIM1 with MIM2

Fig. 3 plots the relation between simulated normalized concentrations and the parameter values sampled from the prior parameter distributions for models MIM1 and MIM2. The simulations correspond to the maximum concentrations (marked by the dashed lines in the figure) in experiment Q2. Each subplot in Fig. 3 was produced by first simulating the normalized maximum concentration using the parameters sampled from the uniform parameter distributions listed in Table 1, and then plotting the simulated concentrations with the corresponding samples of a model parameter. Fig. 3(a) and (d) show the relation between simulated concentration and dispersivity. An individual sample value of dispersivity corresponds to a range of simulated concentration, because the other two model parameters ($\beta$ and $\omega$) also vary during the MC simulation. Fig. 3 shows that increasing model complexity in MIM2 (by adding the parameter of pulse duration) does not increase the model’s capability to simulate the data. Fig. 3(a) and (d) show that the data cannot be simulated with either model when the dispersivity is greater than 20 cm, because the dashed lines in red (used to represent the data) are not included in the ranges of the simulated concentration (the area in blue or black) for the dispersivity greater than 20 cm. Fig. 3(b) and (c) show that the data are already captured by model MIM1 (as the dashed red lines are located within the blue areas) without including the pulse duration as a calibrated parameter. In other words, in comparison of MIM1 and MIM2, the more complex model does not have a better predictive capability for simulating the observed concentration. Therefore, the complexity of model MIM2 is not justified, which is confirmed by the observation that the posterior probability of model MIM1 is substantially larger than that of model MIM2 (Table 3). This may explain why the decrease of $RMSE$ (shown in Table 2) is negligible after the pulse duration is calibrated (except for experiment Q3), considering that MIM1 can capture the data without calibrating the pulse duration.

### 4.3. Analysis of reasonability of model ADE1 in Case 4

Table 3 shows that, while model MIM1 is ranked as the best model in Case 3 for all the experiments, this model is ranked as the best model in Case 4 for experiments Q1 and Q2, whereas model ADE1 is ranked as the best model in Case 4 for experiments Q3 and Q4, which were conducted in column II. This
The question of whether model ADE1 can be more plausible than model MIM1 was investigated by examining the variability of model outputs and the ways in which the models simulate the calibrated data discussed at the beginning of Section 2. Fig. 4 plots the PDFs of simulated normalized concentrations using the parameters sampled from the prior distributions in Case 4. The PDFs were obtained using the kernel density function of MATLAB. The simulations correspond to the maximum concentration of the four experiments, and the observed maximum concentrations are marked by the dashed lines in the figure. The PDFs show dispersivity and biasness of numerical simulations, which jointly determine the performance of the models. For experiments Q1 and Q2, as shown in Fig. 4(a) and (b), model ADE1 has the smallest biasness and dispersivity among the four models. Despite of the smallest biasness, due to the smallest dispersivity, model ADE1 simulates the observed concentration with a small probability, because the observed concentration is located at the very end of the tails of the PDF curves. This is not the case for experiments Q3 and Q4 in that the dispersivity of ADE1 becomes larger. For experiment Q3, Fig. 4(c) shows that ADE1 has the smallest biasness and reasonably large dispersivity. Therefore, among the four models, ADE1 simulates the observed concentration with the highest probability. This is also observed in Fig. 4(d) for experiment Q4, except that ADE1 simulates the observed concentration with the second highest probability. This however does not mean that ADE1 is the second most plausible model, as the model ranking listed in Table 3 is not based on the simulations of the maximum observed concentration (shown in Fig. 4), but on the simulations of all observed concentrations. The corresponding implication is that, when the circumstances (with respect to model parameters) for numerical simulations are largely unknown, it may be more conservative to use a simple model than to use a complex model.

4.4. Determination of proper prior parameter distributions

As the prior parameter distribution significantly affects the evaluation of model complexity, it is necessary to determine the prior parameter distribution to be proper to the Bayesian evaluation of model complexity. While it was beyond the scope of this study to develop a detailed procedure for choosing proper prior parameter distributions, we provide a number of guidelines in the following steps:

(1) Determining the purpose of evaluating the complexity of alternative models: Since a model that is over-complex for one purpose may be under-complex for another purpose, it is critical to determine the purpose of evaluating model complexity. The purpose may be defined by multiple factors. For example, one may focus on model structure and the number of model parameters.

(2) Selecting the state variables that serve the purpose of evaluating model complexity: The state variables will be used for the sensitivity analysis discussed below, and the observations of the state variables will be used for evaluating the posterior probability of alternative models.

(3) Conducting a sensitivity analysis to understand how the state variables are affected by model parameters: Special attention should be paid to the parameters that are influential to the state variables when determining the prior parameter distributions. A formal sensitivity analysis for multiple models can be conducted using the methods of Dai and Ye (2015) and Dai et al. (2017a, 2017b).

(4) Developing preliminary prior parameter distributions based on literature review: If prior information is scarce, uniform distributions with large parameter ranges can be used.

(5) Re-examining the purpose of evaluating the complexity of alternative models to determine whether it is possible to...
reduce the parameter ranges obtained through literature review: A solid understanding of the inputs and outputs of the alternative models is needed in this step and the next step.

(6) Conducting numerical simulations to further refine the parameter ranges by examining model outputs from the following aspects:

(a) Determining whether the model outputs are physically unreasonable (e.g., negative solute concentration) or implausible (e.g., solute concentration exceeding certain thresholds) given the purpose of evaluating model complexity determined in Step (1): If such model outputs are identified, it is necessary to link the outputs with model parameters. If the problem is caused by wide parameter ranges, the parameter ranges can be reduced to remove the physically unreasonable or implausible outputs.

(b) Determining whether the modes of the model outputs are unexpected based on the physical understanding of the alternative models: If prior parameter distributions and the alternative models are reasonable, the modes of model outputs should be expected (e.g., not too large or too small in comparison with expected values of the state variables). Otherwise, one should re-examine the alternative models and the prior parameter distributions.

(c) Determining whether the joint likelihood surfaces are meaningful given the prior parameter distributions: Constructing the likelihood surface requires data regarding the state variables selected in Step (2), and the likelihood surface should span the prior parameter space in a reasonable way. If the likelihood surface only occupies a small portion of the prior parameter space, the prior parameter distributions are too wide. If the likelihood surface is truncated by the prior parameter space, the prior parameter distributions are too narrow.

If Step (3) of the sensitivity analysis is conducted, the model outputs of the sensitivity analysis can be used in Step (6) to save computational cost. The key to choosing proper prior parameter distributions is that the distributions should serve the purpose of evaluating the complexity of alternative models. Qualitatively speaking, if the purpose is well defined and the parameter values used for evaluating model complexity are well known, narrow parameter ranges should be used. Otherwise, wide parameter ranges should be used. More research on the detailed procedure for choosing proper prior parameter distributions is warranted in future studies.

5. Conclusions

This study used the marginal likelihood and Bayesian posterior model probability for statistical evaluation of model complexity in order to avoid using over-complex models for numerical simulations. The contribution of this study was the investigation of impacts of prior parameter distributions...
(involved in calculating the marginal likelihood) on the evaluation of model complexity. We have shown that the parameter distributions define the parameter space in which numerical simulations are conducted. New perspectives on the prior parameter distribution and marginal likelihood have been demonstrated in the numerical simulation, in which models ADE1, ADE2, MIM1, and MIM2 were used to simulate four column experiments of contaminant transport. The four models had different levels of complexity in terms of their model structures and numbers of calibrated parameters. The numerical modeling with four different cases of prior parameter distributions shows that the prior parameter distributions substantially impact model ranking and identification of model complexity. The most complex model (MIM2) is ranked as the best in Case 1, which considers only a point (the calibrated parameter values) in the parameter space. The second most complex model (MIM1) is ranked as the best in Case 2, when considering parameters in a Gaussian sub-space in the parameter space. This is also the case for Case 3, which considers larger parameter ranges. However, when the parameter ranges change in Case 4, model ADE1 receives higher model ranking. The model ranking in each case is reasonable for the specific circumstances (with respect to model parameters) in which numerical simulations are made. Therefore, before evaluating model complexity, it is necessary to determine the parameter space for modeling, which can be done by conducting numerical simulation and using expert judgment based on understanding of the system being studied.

Acknowledgements

The first author would like to acknowledge Florida State University, where she conducted the present research as a visiting student.

References


Parker, J.C., van Genuchten, M.T., 1984. Determining Transport Parameters from Laboratory and Field Tracer Experiments, Virginia Agricultural Experiment Station Bulletin 84-3. Virginia Agricultural Experiment Station, Blacksburg.


