AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2014WR016718

Key Points:

- Thermodynamic integration is more accurate and consistent than several other numerical methods
- Thermodynamic integration is mathematically general and straightforward to implement
- Power coefficients of thermodynamic integration can be selected empirically but objectively

Correspondence to:
M. Ye,
mye@fsu.edu

Citation:
Liu, P., A. S. Elshall, M. Ye, P. Beerli, X. Zeng, D. Lu, and Y. Tao (2016), Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods, *Water Resour. Res.*, *52*, doi:10.1002/2014WR016718.

Received 20 NOV 2014
Accepted 13 JAN 2016
Accepted article online 18 JAN 2016

Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods

Peigui Liu[1,2], Ahmed S. Elshall[2], Ming Ye[2], Peter Beerli[2], Xiankui Zeng[3], Dan Lu[4], and Yuezan Tao[1]

[1]School of Civil Engineering, Hefei University of Technology, Hefei, China, [2]Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA, [3]School of Earth Sciences and Engineering, Nanjing University, Nanjing, China, [4]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

Abstract Evaluating marginal likelihood is the most critical and computationally expensive task, when conducting Bayesian model averaging to quantify parametric and model uncertainties. The evaluation is commonly done by using Laplace approximations to evaluate semianalytical expressions of the marginal likelihood or by using Monte Carlo (MC) methods to evaluate arithmetic or harmonic mean of a joint likelihood function. This study introduces a new MC method, i.e., thermodynamic integration, which has not been attempted in environmental modeling. Instead of using samples only from prior parameter space (as in arithmetic mean evaluation) or posterior parameter space (as in harmonic mean evaluation), the thermodynamic integration method uses samples generated gradually from the prior to posterior parameter space. This is done through a path sampling that conducts Markov chain Monte Carlo simulation with different power coefficient values applied to the joint likelihood function. The thermodynamic integration method is evaluated using three analytical functions by comparing the method with two variants of the Laplace approximation method and three MC methods, including the nested sampling method that is recently introduced into environmental modeling. The thermodynamic integration method outperforms the other methods in terms of their accuracy, convergence, and consistency. The thermodynamic integration method is also applied to a synthetic case of groundwater modeling with four alternative models. The application shows that model probabilities obtained using the thermodynamic integration method improves predictive performance of Bayesian model averaging. The thermodynamic integration method is mathematically rigorous, and its MC implementation is computationally general for a wide range of environmental problems.

1. Introduction

Multimodel analysis has been used widely in environmental modeling for quantification of model uncertainty, which arises when multiple conceptualizations and mathematical descriptions are considered to be plausible for simulating an environmental system. By evaluating multiple models simultaneously, multimodel analysis not only addresses model uncertainty but also provides a quantitative framework to quantify model uncertainty. This is done by aggregating model predictions and associated uncertainty from the competing models in a model averaging process. Among various methods of multimodel analysis, this paper is focused on the Bayesian Model Averaging (BMA) method. Consider a set of models, $\mathbf{M}=(M_1,\ldots,M_K)$ and denote Δ as a quantity to be predicted by model M_k based on available data $\mathbf{D}$. The weighted average estimate of the probability density function of Δ is

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

where $p(M_k|\mathbf{D})$ is the posterior probability of model M_k. The posterior model probabilities add up to one, and can be viewed as model averaging weight. The posterior probability for model M_k is given by Bayes' rule [*Hoeting et al.*, 1999],

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

where $p(M_k)$ is prior probability and

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

is marginal likelihood of model M_k, $\boldsymbol{\theta}_k$ is the vector of parameters associated with model M_k, $p(\boldsymbol{\theta}_k|M_k)$ is the prior density of $\boldsymbol{\theta}_k$ under model M_k, $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$ is the joint likelihood of model M_k and its parameters $\boldsymbol{\theta}_k$. The marginal likelihood function in equation (3) is one of the most critical variables in BMA, and evaluating it numerically is the focus of this paper.

The marginal likelihood, also called integrated likelihood or Bayesian evidence, measures overall model fit, i.e., to what extent that the data, **D**, can be simulated by model M_k. The measure is not based on a single parameter sample, but based on the model fit averaged over the entire parameter space, which corresponds to the integration of $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$ with respect to the prior parameter density. The marginal likelihood is important to model selection. In the context of Bayesian factor [*Kass and Raftery*, 1995], models with higher values of marginal likelihood are considered to be more plausible. This concept was used by *Marshall et al.* [2005] for hydrologic model selection. The marginal likelihood is also used in many other areas of environmental modeling. For example, it was used by *Schoups et al.* [2008] and *Schoups and Vrugt* [2011] to evaluate the appropriate level of model complexity. When the level of model complexity (in terms of model parameter and structure) increases but the marginal likelihood does not improve, the increased level of complexity is not justified. Similarly, the marginal likelihood has been used for inverse modeling [*Elsheikh et al.*, 2013] and data assimilation [*Vrugt et al.*, 2013]. It is therefore necessary to accurately and efficiently evaluate the marginal likelihood.

Analytical expressions of the marginal likelihood are only available under limited situations, e.g., with linear models and Gaussian likelihood and prior [*Schöniger et al.*, 2014]. Semianalytical expressions can be derived by using the Laplace approximation method [*Kass and Raftery*, 1995; *Friel and Wyse*, 2012], and the expressions are given in section 2. The expressions are semianalytical, because inverse modeling is needed to obtain estimates of parameter, $\boldsymbol{\theta}_k$, using either maximum likelihood (ML) or maximum a posterior (MAP) method. While the Laplace approximation methods have been widely used in groundwater modeling [*Neuman*, 2003; *Ye et al.*, 2004, 2008, 2010a,b; *Lu et al.*, 2013; *Tsai and Elshall*, 2013; *Elshall and Tsai*, 2014; *Zhang et al.*, 2014], it is subjective to truncation error, and linearization of nonlinear models is always needed during the inverse modeling. A comprehensive evaluation of the Laplace approximation methods can be found in *Schöniger et al.* [2014].

Approximating the marginal likelihood using Monte Carlo (MC) methods has become popular [*Kass and Raftery*, 1995; *Han and Carlin*, 2001; *von Toussaint*, 2011; *Friel and Wyse*, 2012; *Vrugt et al.*, 2013]. The most popular MC approximation used in groundwater modeling is to simply take the arithmetic mean of $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$ evaluated for parameter samples obtained by various methods, e.g., generalized likelihood uncertainty estimation methods [*Rojas et al.*, 2008], method of anchored distributions [*Rubin et al.*, 2010], Markov chain Monte Carlo (MCMC) methods [*Lu et al.*, 2011], maximum likelihood methods [*Neuman et al.*, 2012; *Lu et al.*, 2012b], and ensemble Kalman filter methods [*Xue and Zhang*, 2014]. However, it is well known that calculating arithmetic mean suffers from slow convergence. In addition, the arithmetic mean always underestimates the marginal likelihood, if the parameter samples are generated from a prior parameter space that is high dimensional and has wide range. The reason for the underestimation is that, the joint likelihood, $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$, is small for most samples generated from the prior space. A solution to this problem is to use the harmonic mean evaluated by using MCMC parameter samples from the posterior parameter space [*Kass and Raftery*, 1995]. It however has been found that the harmonic mean may overestimate the marginal likelihood, because the MCMC samples have high joint likelihood. In addition, the harmonic mean is unstable, because its evaluation may be dominated by parameter samples with small values of the joint likelihood. To resolve this issue raised by using only prior or posterior samples, *Newton and Raftery* [1994] developed the stabilized harmonic mean method that uses importance sampling to generate parameter samples from a mixture of prior and posterior spaces. The mixture however is subjective and difficult to be determined; the coefficient used for the mixture always needs to be calibrated. A similar method was developed by *Chib and Jeliazkov* [2001] and used by *Marshall et al*. [2005] for hydrologic modeling. This method uses both MCMC samples and samples generated from the MCMC proposal distributions; the proposal samples are not limited to the prior and posterior spaces. The effort of developing more advanced methods to stabilize harmonic mean is on-going [*Raftery et al*., 2007].

This study is focused on two MC approaches that do not use harmonic mean, and they are the thermodynamic integration method [*Gelman and Meng*, 1998; *Neal*, 2000; *Lartillot and Philippe*, 2006; *Friel and Pettitt*, 2008] and the nested sampling method [*Skilling*, 2004, 2006]. Previous studies have shown that the two approaches are more suitable than other MC approaches for evaluating the marginal likelihood. *Beerli and Palczewski* [2010] showed that thermodynamic integration method gives considerably better estimates than the stabilized harmonic mean does. *Schöniger et al*. [2014] showed that nested sampling method outperforms the Laplace approximation methods. The thermodynamic integration and nested sampling methods have a common characteristic that is to convert the multivariate integral of equation (3) into a one-dimensional integral. For thermodynamic integration, the one-dimensional integral is with respect to a power coefficient applied to the joint likelihood; for nest sampling, the one-dimensional integral is with respect to an element of prior mass (i.e., integration of prior within a region). More details of the conversion are given in section 2. While nested sampling has been recently used for groundwater modeling [*Elsheikh et al*., 2013; *Schöniger et al*., 2014], to the best of our knowledge, thermodynamic integration has not been attempted, except the work of *Schoups and Vrugt* [2011]. One objective of this study is to introduce the thermodynamic integration method into groundwater modeling for model uncertainty quantification.

Another objective of this study is to compare the commonly used numerical methods, including Laplace approximation, arithmetic mean, harmonic mean, nested sampling, and thermodynamic integration, in terms their accuracy, convergence, and consistency. There have been only few studies that compare these methods [*Schöniger et al*., 2014; *Friel and Wyse*, 2012]. Since these numerical methods are widely used in environmental modeling, the comparison can provide direct insights for various research areas of environmental modeling. There have been other MC methods for evaluating the marginal likelihood, such as those using Savage-Dickey density ratio [*Morey et al*., 2011], Lebesgue integration theory [*Weinberg*, 2012], and stepping-stone method [*Xie et al*., 2011]. However, including these methods into the numerical comparison is beyond the scope of this study. A special attention is paid in this study to compare thermodynamic integration with nested sampling, which has not been reported in the literature. During the numerical comparison, it is realized that the Metropolis-Hasting MCMC techniques used in *Elsheikh et al*. [2013] for implementing nested sampling are standard and do not have advanced features of the DREAM code [*Vrugt et al*., 2009; *Laloy and Vrugt*, 2012] used to implement thermodynamic integration. Three modifications are made in this study to improve the sampling efficiency, such as using block and component-wise updating and randomized step-size reduction factor to generate proposal samples. More details of the modifications are given in section 2.

The numerical comparison is conducted using the following three analytical functions: (1) a linear function with two parameter, (2) a nonlinear function with two parameters whose distributions are multimodal, and (3) a nonlinear function with 10 parameters and sharp likelihood surface. Three evaluation criteria are used in the numerical comparison, i.e., accuracy, convergence, consistency, and computational cost of the numerical methods are also considered in this study. While the thermodynamic integration method may not be the best one with respect to the individual criteria and for a single function, it overall outperforms the other methods, especially for the third analytical function that is more representative for environmental modeling. The thermodynamic integration method is also applied to a synthetic case of groundwater modeling. The synthetic case considers four alternative models postulated for different conceptualizations of a confining layer, a commonly encountered conceptual model uncertainty in groundwater modeling [*Lemke and Cypher*, 2010]. The four models have a relatively large number of parameters, ranging from 12 to 21, which is common in groundwater modeling. The groundwater modeling shows that using the model probabilities given by thermodynamic integration improves BMA predictive performance (measured for simulating streamflow change due to pumping) compared to using the model probabilities given by Laplace approximation, arithmetic mean, and harmonic mean.

2. Methodologies

This section starts with a brief description of the Laplace approximation method, followed by the definitions of arithmetic mean estimator (AME), harmonic mean estimator (HME), thermodynamic integration estimator

(TIE), and nested sampling estimator (NSE) of the marginal likelihood. Since these methods are applied to individual models, the subscript k of model M_k drops hereinafter for the convenience of mathematical notation.

2.1. Laplace Approximations

There are two variants of the Laplace approximation method, depending on how the Taylor series expansion is conducted for the integrand of equation (3) [*Kass and Raftery*, 1995; *Schöniger et al.*, 2014]. The first one rewrites equation (3) as

$$p(\mathbf{D}|M)=\int \exp\left[\ln p(\mathbf{D}|\boldsymbol{\theta},M)p(\boldsymbol{\theta}|M)\right]d\boldsymbol{\theta}, \tag{4}$$

and then expands $\ln\left(p(\mathbf{D}|\boldsymbol{\theta},M)p(\boldsymbol{\theta}|M)\right)$ in a Taylor series around the parameter estimate, $\tilde{\boldsymbol{\theta}}$, that maximizes $\ln\left(p(\mathbf{D}|\boldsymbol{\theta},M)p(\boldsymbol{\theta}|M)\right)$. After truncating all terms of the Taylor series that have the orders of $\tilde{\boldsymbol{\theta}}$ higher than two, the quadratic Taylor series expansion is

$$\ln p(\mathbf{D}|\boldsymbol{\theta},M)\approx \ln p(\mathbf{D}|\tilde{\boldsymbol{\theta}},M)+\ln p\left(\tilde{\boldsymbol{\theta}}\middle|M\right)-\frac{1}{2}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})^T\mathbf{F}(\mathbf{D}|\tilde{\boldsymbol{\theta}},M)(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}), \tag{5}$$

where $\mathbf{F}$ is the Fisher information matrix and its element is defined as

$$F_{ij}(\mathbf{D}|\boldsymbol{\theta},M)=\frac{\partial^2\ln\left(p(\mathbf{D}|\boldsymbol{\theta},M)p(\boldsymbol{\theta}|M)\right)}{\partial\theta_i\partial\theta_j}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \tag{6}$$

Substituting equation (5) into equation (4) leads to the semianalytical expression of the marginal likelihood,

$$p(\mathbf{D}|M)\approx 2\pi^{\frac{N}{2}}\left|\mathbf{F}(\mathbf{D}|\tilde{\boldsymbol{\theta}},M)\right|^{-\frac{1}{2}}p(\mathbf{D}|\tilde{\boldsymbol{\theta}},M)p\left(\tilde{\boldsymbol{\theta}}\middle|M\right), \tag{7}$$

where N is the dimension of model parameter, $\boldsymbol{\theta}$. Equation (7) is referred to as Laplace-MAP, because the expansion is for the maximum a posterior (MAP) estimate, $\tilde{\boldsymbol{\theta}}$, of model parameters [*Carrera and Neuman*, 1986].

The other variant of the Laplace approximation is similar but expands first $\ln p(\mathbf{D}|\boldsymbol{\theta},M)$ and then $\ln p(\boldsymbol{\theta}|M)$ in a Taylor series around the parameter estimate, $\hat{\boldsymbol{\theta}}$, that maximizes $\ln p(\mathbf{D}|\boldsymbol{\theta},M)$. The corresponding semianalytical expression of the marginal likelihood is [*Kass and Raftery*, 1995]

$$p(\mathbf{D}|M)\approx 2\pi^{\frac{N}{2}}\left|\mathbf{F}(\mathbf{D}|\hat{\boldsymbol{\theta}},M)\right|^{-\frac{1}{2}}p(\mathbf{D}|\hat{\boldsymbol{\theta}},M)p\left(\hat{\boldsymbol{\theta}}\middle|M\right), \tag{8}$$

where the elements of the Fisher information matrix are defined as $F_{ij}(\mathbf{D}|\boldsymbol{\theta},M)=\frac{\partial^2\ln p(\mathbf{D}|\boldsymbol{\theta},M)}{\partial\theta_i\partial\theta_j}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$. Equation (8) is referred to as Laplace-MLE, because the expansion is for the maximum likelihood (ML) estimate, $\hat{\boldsymbol{\theta}}$, of model parameters. For a linear model, Laplace-MAP is accurate because the terms in the Taylor series with order higher than two are zero. This is not the case for Laplace-MLE, because $\hat{\boldsymbol{\theta}}$ may not maximize $\ln p(\boldsymbol{\theta}|M)$. In addition, the difference in the Fisher information matrix between Laplace-MAP and Laplace-MLE may also impact the accuracy of Laplace-MLE. Theoretically speaking, Laplace-MAP is more accurate than Laplace-MLE, when the prior is informative relative to the likelihood. A detailed comparison of the two estimators can be found in *Schöniger et al.* [2014].

2.2. Arithmetic Mean Estimator (AME) and Harmonic Mean Estimator (HME)

Following *Kass and Raftery* [1995], AME is defined as

$$\hat{p}_{AME}(\mathbf{D}|M)=\frac{1}{m}\sum_{i=1}^{m}p(\mathbf{D}|\boldsymbol{\theta}_{prior}^{(i)},M), \tag{9}$$

where m is the number of parameter samples $\boldsymbol{\theta}_{prior}^{(i)}$ from the prior distribution, $p(\boldsymbol{\theta}|M)$. While AME is conceptually straightforward, it is computationally inefficient, especially for problems with a large number of parameters. In addition, if the posterior parameter distribution is significantly narrower than the prior parameter distribution, the values of joint likelihood, $p(\mathbf{D}|\boldsymbol{\theta},M)$, are small for the prior samples, $\boldsymbol{\theta}_{prior}^{(i)}$. This may lead to large variance and/or underestimation of the marginal likelihood, unless a large number of prior samples are used.

HME is defined as [*Kass and Raftery*, 1995]

$$\hat{p}_{HME}(\mathbf{D}|M)=\left\{\frac{1}{m}\sum_{i=1}^{m}p(\mathbf{D}|\boldsymbol{\theta}_{posterior}^{(i)},M)^{-1}\right\}^{-1}, \tag{10}$$

where $\boldsymbol{\theta}_{posterior}^{(i)}$ is a parameter sample from its posterior distribution, and can be obtained using MCMC methods. Equation (10) can be derived either from the angle of importance sampling [*Kass and Raftery*, 1995] or based on the identity,

$$\frac{1}{p(\mathbf{D}|M)}=\int\frac{1}{p(\mathbf{D}|\boldsymbol{\theta},M)}p(\boldsymbol{\theta}|\mathbf{D},M)d\boldsymbol{\theta}, \tag{11}$$

which is derived by substituting the equation of posterior parameter distribution,

$$p(\boldsymbol{\theta}|\mathbf{D},M)=\frac{p(\mathbf{D}|\boldsymbol{\theta},M)p(\boldsymbol{\theta}|M)}{p(\mathbf{D}|M)}, \tag{12}$$

into equation (11). Equation (11) differs from equation (3) in that the former is the mean of $p(\mathbf{D}|\boldsymbol{\theta},M)^{-1}$ with respect to the posterior but the latter is the mean of $p(\mathbf{D}|\boldsymbol{\theta},M)$ with respect to the prior. While HME is computationally more efficient than AME, it may overestimate the marginal likelihood, because the values of $p(\mathbf{D}|\boldsymbol{\theta}_{posterior}^{(i)},M)$ are large for the posterior samples, especially when the posterior is significantly narrower than the prior. In addition, HME suffers from instability, because a posterior sample with a small value of $p(\mathbf{D}|\boldsymbol{\theta}_{posterior}^{(i)},M)$ can dominates the calculation. This problem cannot be resolved by using a large number of posterior samples, as shown in the numerical examples below.

2.3. Thermodynamic Integration Estimator (TIE)

To resolve the problems above about AME and HME, a new sampling method is needed, and the key idea is to avoid sampling solely in the prior or posterior parameter space. This can be achieved by TIE, which is also known as path sampling [*Gelman and Meng*, 1998; *Neal*, 2000]. TIE is based on the power posterior defined for any $0\le\beta\le 1$ as

$$q_\beta(\boldsymbol{\theta})=p(\mathbf{D}|\boldsymbol{\theta},M)^{\beta}p(\boldsymbol{\theta}|M), \tag{13}$$

which is a continuous and differentiable path in the space of unnormalized densities, with $q_0(\boldsymbol{\theta})=p(\boldsymbol{\theta}|M)$ and $q_1(\boldsymbol{\theta})=p(\mathbf{D}|\boldsymbol{\theta},M)p(\boldsymbol{\theta}|M)$. Note that $\boldsymbol{\theta}$ is implicitly a function of β. For $\beta=0$ and $\beta=1$, TIE samples the prior and posterior parameter space, respectively. For $0<\beta<1$, the likelihood surface is smoothed out to explore the posterior parameter space specific to β. The TIE derivation is not straightforward, because it requires defining two interim variables that do not have physical or statistical meaning but only facilitate the derivation. Following *Lartillot and Philippe* [2006], define the intermediate variable, p_β and Z_β, as

$$p_\beta(\boldsymbol{\theta})=\frac{1}{Z_\beta}q_\beta(\boldsymbol{\theta}), \tag{14}$$

and

$$Z_\beta=\int q_\beta(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{15}$$

Taking the derivative of $\ln Z_\beta$ with respect to β leads to

$$\begin{aligned}\frac{\partial\ln Z_\beta}{\partial\beta}&=\frac{1}{Z_\beta}\frac{\partial Z_\beta}{\partial\beta}=\frac{1}{Z_\beta}\int\frac{\partial q_\beta(\boldsymbol{\theta})}{\partial\beta}d\boldsymbol{\theta}\\&=\int\frac{1}{q_\beta(\boldsymbol{\theta})}\frac{\partial q_\beta(\boldsymbol{\theta})}{\partial\beta}\frac{q_\beta(\boldsymbol{\theta})}{Z_\beta}d\boldsymbol{\theta}=\int\frac{\partial\ln q_\beta(\boldsymbol{\theta})}{\partial\beta}p_\beta(\boldsymbol{\theta})d\boldsymbol{\theta}\\&=E_\theta\left[\frac{\partial\ln q_\beta(\boldsymbol{\theta})}{\partial\beta}\right],\end{aligned} \tag{16}$$

where E_θ denotes the expectation with respect to $p_\beta(\boldsymbol{\theta})$. Integrating (16) with respect to β over [0, 1] yields

$$\ln Z_1 - \ln Z_0 = \int_0^1 E_\theta\left[\frac{\partial \ln q_\beta(\boldsymbol{\theta})}{\partial \beta}\right] d\beta. \tag{17}$$

Considering that

$$\begin{aligned} Z_0 &= \int q_0(\boldsymbol{\theta})d\boldsymbol{\theta} = \int p(\boldsymbol{\theta}|M)d\boldsymbol{\theta} = 1, \\ Z_1 &= \int q_1(\boldsymbol{\theta})d\boldsymbol{\theta} = \int p(\mathbf{D}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta} = p(\mathbf{D}|M), \end{aligned} \tag{18}$$

equation (17) becomes

$$\ln p(\mathbf{D}|M) = \int_0^1 E_\theta\left[\frac{\partial \ln q_\beta(\boldsymbol{\theta})}{\partial \beta}\right] d\beta, \tag{19}$$

and it leads to

$$p(\mathbf{D}|M) = \exp\left[\int_0^1 E_\theta\left[\frac{\partial \ln q_\beta(\boldsymbol{\theta})}{\partial \beta}\right] d\beta\right]. \tag{20}$$

This is the key to TIE that converts the multivariate integral of equation (3) into the univariate integral with respect to the scalar, β.

The next step is to derive the analytical expression of $\frac{\partial \ln q_\beta(\boldsymbol{\theta})}{\partial \beta}$. Considering that $\frac{\partial \ln q_\beta(\boldsymbol{\theta})}{\partial \beta} = \frac{1}{q_\beta(\boldsymbol{\theta})}\frac{\partial q_\beta(\boldsymbol{\theta})}{\partial \beta}$ and $\frac{\partial q_\beta(\boldsymbol{\theta})}{\partial \beta} = p(\mathbf{D}|\boldsymbol{\theta}, M)^\beta \ln p(\mathbf{D}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)$ by virtue of equation (13), we have

$$\begin{aligned} \frac{\partial \ln q_\beta(\boldsymbol{\theta})}{\partial \beta} &= \frac{1}{q_\beta(\boldsymbol{\theta})}\frac{\partial q_\beta(\boldsymbol{\theta})}{\partial \beta} = \frac{1}{q_\beta(\boldsymbol{\theta})} p(\mathbf{D}|\boldsymbol{\theta}, M)^\beta \ln p(\mathbf{D}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M) \\ &= \frac{1}{q_\beta(\boldsymbol{\theta})} q_\beta(\boldsymbol{\theta}) \ln p(\mathbf{D}|\boldsymbol{\theta}, M) = \ln p(\mathbf{D}|\boldsymbol{\theta}, M). \end{aligned} \tag{21}$$

Substituting (21) into (20) gives

$$p(\mathbf{D}|M) = \exp\left[\int_0^1 E_\theta[\ln p(\mathbf{D}|\boldsymbol{\theta}, M)]d\beta\right]. \tag{22}$$

At this step, the two interim variables drop off from the evaluation. Equation (22) is a one-dimensional integral, and can be easily estimated using various quadrature rules. Using the simple composite trapezoidal rule with discrete power coefficients, $0 = \beta_1 < \beta_2 \ldots < \beta_k < \ldots \beta_n = 1$, TIE is obtained as

$$\hat{p}_{TIE}(\mathbf{D}|M) = \exp\left[\sum_{k=2}^{n} (\beta_k - \beta_{k-1}) \frac{y_k + y_{k-1}}{2}\right], \tag{23}$$

where y_k corresponding to power coefficient value β_k is the average of log likelihood, $\ln p(\mathbf{D}|\boldsymbol{\theta}^{(i)}, M)$, i.e.,

$$y_k = \frac{1}{m}\sum_{i=1}^{m} \ln p_{\beta_k}(\mathbf{D}|\boldsymbol{\theta}^{(i)}, M). \tag{24}$$

The above procedure is based on the annealing-melting integration. Another TIE derivation based on model-switch integration is referred to *Lartillot and Philippe* [2006]. It should be noted that TIE outperforms AME and HME only when the posterior is significantly narrower than the prior. In an extreme case that the posterior is the same as the prior (e.g., when data does not update the prior), AME, HME, and TIE will be the same.

Below is a procedure of implementing TIE using MCMC:

Step 1: Determine the discrete power coefficient values, β_k, which usually start with equal intervals for $0 \leq \beta \leq 1$.

Step 2: Run MCMC with different β_k values in parallel to obtain the corresponding parameter samples.

Step 3: Calculate y_k corresponding to each β_k using equation (24).

Step 4: Compute TIE using equation (23).

Step 5: If it is necessary to add more power coefficient values, go to Step 1. Otherwise, stop.

As shown in section 3, it is straightforward to determine the number and values of β_k in an empirical manner by increasing the number of β_k gradually. TIE becomes more accurate when the number of β_k increases, and converges when the number is large enough.

2.4. Nested Sampling Estimator (NSE)

Following *Skilling* [2006], NSE evaluates the marginal likelihood via

$$p(\mathbf{D}|M)=\int p(\mathbf{D}|\boldsymbol{\theta},M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}=\int p(\mathbf{D}|\boldsymbol{\theta},M)dX=\int L(\boldsymbol{\theta}|\mathbf{D},M)dX, \tag{25}$$

where $L(\boldsymbol{\theta}|\mathbf{D},M)=p(\mathbf{D}|\boldsymbol{\theta},M)$ denotes the joint likelihood function, and $dX = p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$ is an element of prior mass. Define a cumulative prior mass

$$X(\lambda)=\int\limits_{L(\boldsymbol{\theta}|\mathbf{D},M)>\lambda} p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}, \tag{26}$$

for the parameter values that satisfy the condition, $L(\boldsymbol{\theta}|\mathbf{D},M)>\lambda$. Note that $0\le X\le 1$. As the restriction on the joint likelihood become tighter as λ increases, this condition allows exploring a "hard-edged likelihood-constrained domain" (called by *Skilling* [2006]) evolving to higher likelihood values. Denote the inverse function of $X(\lambda)$ as $p(\mathbf{D}|X,M)$ that satisfies $p(\mathbf{D}|X(\lambda),M)=\lambda$. Substituting the inverse function into equation (25) leads to a one-dimensional integral,

$$\mathrm{N}\hat{p}_{NSE}(\mathbf{D}|M)=\int_0^1 p(\mathbf{D}|X,M)dX=\int_0^1 L(X|\mathbf{D},M)dX. \tag{27}$$

This is the key to NES that converts the multivariate integral of equation (3) into this univariate integral with respect to X, which is equivalent to mapping the multidimensional parameter space into the one-dimensional space of X. Unlike TIE that samples from the prior to the posterior parameter space by exploring a "likelihood weighted space" called by *Skilling* [2006], NSE samples within the hard constraint, λ, and thus explores the hard-edged likelihood-constrained domain. This is the reason that NSE is in general computationally more efficient than TIE. However, finding the hard-edged likelihood-constrained domain may be difficult, especially for high-dimensional problems with wide parameter range and peaked joint likelihood. Such a problem is presented in section 3, for which NSE implemented using the commonly used Metropolis-Hasting algorithm [e.g., *Elsheikh et al.*, 2013; *Schöniger et al.*, 2014] fails to find the hard-edged likelihood-constrained domain.

The one-dimensional integral in equation (27) can be evaluated using any quadrature rule,

$$\hat{p}_{NSE}(\mathbf{D}|M)=\sum_{i=1}^{I} w_i L(X_i|\mathbf{D},M)=\sum_{i=1}^{I} w_i L_i, \tag{28}$$

where I is the number of discrete X_i, and w_i is a weight corresponding to X_i. *Skilling* [2006] denoted a right-to-left sequence of X_i points as $0<X_I<\cdots<X_2<X_1<1$, and suggested using the weight of $w_i=X_{i-1}-X_i$ with $X_0=1$. The X_i and L_i values are unknown, and they are determined iteratively in the following procedure [*Skilling*, 2006] based on equation (26):

Step 1: Construct an active set of size N_{as} with parameter samples generated from the prior distribution, calculate for each sample its corresponding joint likelihood, and determine the smallest joint likelihood denoted as L_{worst}, which is λ in equation (26).

Step 2: Start a loop to evaluate equation (28) in two substeps.

Step 2.1: For the ith iteration, denote L_{worst} as L_i, and then compute $X_i=\exp(-i/N_{as})$. Subsequently, evaluate weight, $w_i=X_{i-1}-X_i$, and then calculate w_iL_i in equation (28).

Step 2.2: Replace the parameter sample of L_{worst} in the active set with a new sample generated from the prior. The new sample should satisfy the hard constraint, $L(\boldsymbol{\theta}|\mathbf{D},M)>L_{worst}$. The sample of L_{worst} is removed from the active set.

Step 3: Repeat Step 2 until reaching two user-specified termination criteria, whichever is reached first. The two criteria are the maximum number of iterations (e.g., $I=25\times N_{as}$ used in this study) and the desired accuracy of $\hat{p}_{NSE}$ (e.g., $\sum_{i=1}^{I} w_iL_i-\sum_{i=1}^{I-1} w_iL_i\le 10^{-5}$ used in this study).

Although the procedure is straightforward, from the theoretical point of view, how to evaluate X_i is still an open question. *Skilling* [2006] acknowledged the uncertainty in the current way of evaluating X_i in Step 2.1.

The existing practice of implementing NSE also has operation limitations, in particular for generating a new sample in Step 2.2 to replace the parameter sample of L_{worst} in the current active set. Because of the hard constraint, $L(\boldsymbol{\theta}|\mathbf{D}, M) > L_{worst}$, the random sample generation becomes computationally expensive when L_{worst} becomes large as the algorithm proceeds. In other words, a large number of samples generated from the prior are rejected, when the active set evolves to contain samples with high likelihood. To resolve this problem, *Elsheikh et al*. [2013] provided a Metropolis-Hasting algorithm that starts a short Markov chain from a sample picked from the active set, except the one with the lowest likelihood. A similar algorithm was used in *Schöniger et al*. [2014].

This study makes three modifications on the Metropolis-Hasting algorithm described by *Elsheikh et al*. [2013] on the aspect of generating proposal samples. The modifications however are minor, as they do not substantially change the way of generating proposal samples. Given that MT-DREMA$_{(ZS)}$ [*Laloy and Vrugt*, 2012] is a more robust algorithm than the Metropolis-Hasting algorithm, it would be more ideal to use the DREAM code for implementing NES, which however is beyond the scope of this study.

The first modification is to use the prior probability ratio, not the likelihood ratio, for the acceptance ratio, as was done in *Schöniger et al*. [2014]. The reason for this change is to ensure that sufficient X values, especially small X values, are used in the nested sampling to avoid overestimation of the marginal likelihood (equation (28)). When the likelihood ratio is used, it is observed that the likelihood evolves quickly to a high value. The fast evolution may cause premature termination of the iteration in Step 2, because no more higher likelihood can be found. Its consequence is insufficient construction of the one-dimensional integral with respect to X and overestimation of the marginal likelihood. While using the prior probability ratio (e.g., sampling from the prior) decreases the sampling efficiency, NES is still the most computationally efficient algorithm in comparison with other algorithms.

The second modification is related to proposing a new sample. In *Elsheikh et al*. [2013], a block-wise updating for the proposal distribution was implemented via

$$\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{old} + \delta\omega, \tag{29}$$

where $\boldsymbol{\theta}_{new}$ is proposed sample, $\boldsymbol{\theta}_{old}$ is evolved sample, δ is a user-specified step size, and ω is a random number following the standard normal distribution, i.e., $\omega \sim N(0,1)$. While using this block-wise update can be computationally more efficient, it may result in a high rejection rate, especially for high-dimensional problems. At a cost of computational efficiency, this problem can be alleviated by using the component-wise scheme to update each dimension independently while fixing other components. To gain computational efficiency without a high rejection rate, both block and component-wise update are used in this study but in a random manner. Before selecting the update scheme, a random number is drawn from the uniform distribution, $U[0,1]$. If the random number is larger than 0.5, the component-wise update is used; otherwise, the block-wise update is used.

The third modification is to avoid premature termination, which can occur during the MCMC run, when all proposed MCMC samples are rejected (because their likelihood values are smaller than L_{worst}) but the X space has not been sufficiently explored. To avoid the premature termination, the MCMC simulation will start over but with a smaller step size estimated via

$$\delta_{new} = \alpha R \delta_{old}, \tag{30}$$

where α is a user-specified step-size reduction factor (less than one), and $R \sim \mathrm{U}[0,1]$ is a randomized step-size reduction factor. While α is fixed for all the realizations of the new MCMC run, a random R value is used in each realization. This helps find a new sample that satisfies the hard constrain on the joint likelihood defined in equation (26).

3. Numerical Examples

The marginal likelihood is evaluated using Laplace-MAP, Laplace-MLE, AME, HME, TIE, and NSE for: (1) a linear analytical function with two parameters, (2) a nonlinear analytical function with two parameters whose distributions are multimodal, and (3) a nonlinear analytical function with 10 parameters whose joint likelihood is highly peaked. The Laplace-MAP and Laplace-MLE methods are implemented using the

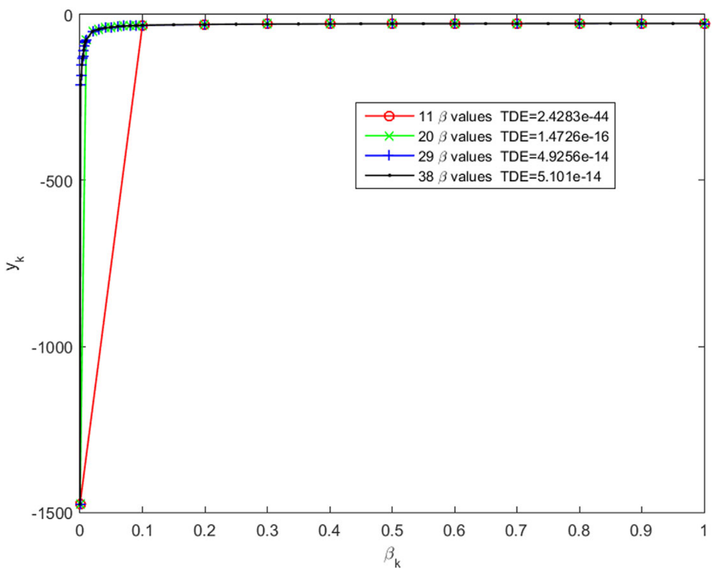

Figure 1. Variation of y_k (equation (24)) with β_k (power coefficient in equation (13)) for evaluating the thermodynamic integration estimator (TIE). The initial number of βk is 11 (red), 9 more values (green) are added between 0 and 0.1, 9 more values (blue) are added between 0 and 0.01, and the final number is 38 (black).

commercial software, Mathematica; AME is evaluated using simple Monte Carlo simulation that generates random parameter samples from prior parameter distributions directly; the MCMC simulation needed by HME and TIE is conducted using MT-DREMA$_{(ZS)}$ [*Laloy and Vrugt*, 2012]. The MCMC simulation needed by NSE is based on the Metropolis-Hasting algorithm used by *Elsheikh et al.* [2013] with the three modifications described in section 2.4. The selection of discrete power coefficient values for TIE and tuning of NES (e.g., the size of active set and use of randomized step-size reduction factor) is demonstrated for the linear analytical function as an example.

For each of the three analytical function, accuracy of the numerical methods is evaluated by comparing their results with reference values; the details of determining the reference values are given below. With the reference values, the convergence of AME, HME, TIE, and NES is also compared by conducting tens of millions of MC runs. Computational cost of each method needed for reaching convergence is discussed. Consistency of AME, HME, TIE, and NES is evaluated by examining variability of their results of 10 repeated runs.

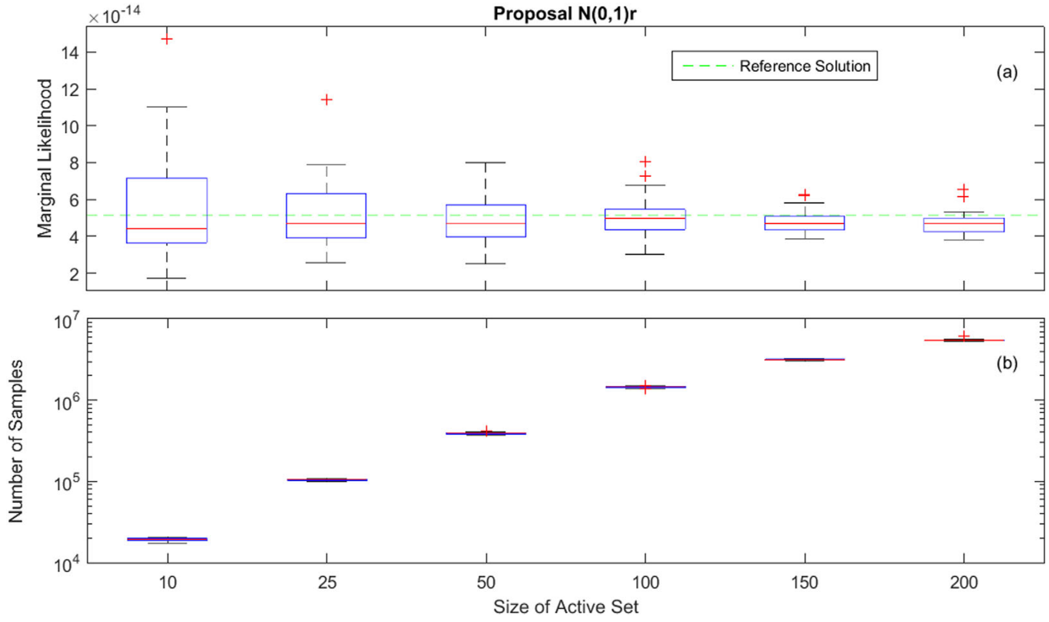

Figure 2. (a) Estimated marginal likelihood, and (b) number of model executions for different sizes of active set used in the nested sampling for the linear function. The proposal distribution based on $\omega \sim N(0,1)r$, where ω is the random number used in equation (29), and r indicates that the randomized step-size reduction factor, R, of equation (30) is used.

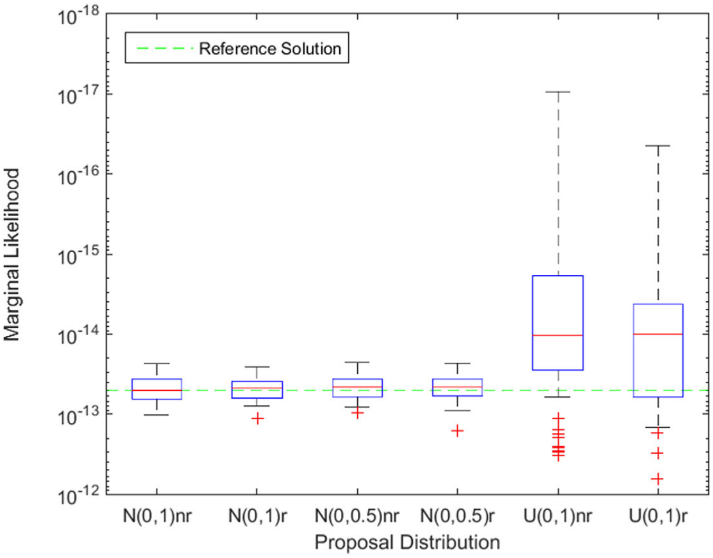

Figure 3. Impacts of distributions (N(0,1), N(0,0.5), and U(0,1)) of ω (in equation (29)) and randomized step-size reduction factor, *R* (in equation (30)), on accuracy and consistency of NSE estimates for the linear function. For the *x* axis label, when *R* is fixed at one, it is denoted as *nr*, when random *R* is used, it is denoted as *r*.

Laplace-MLE, AME, HME, and TIE are used for a synthetic case of groundwater modeling to evaluate the numerical methods in the context of groundwater modeling. Four alternative groundwater flow models are considered in the synthetic case corresponding to four different descriptions of a confining layer. For the four models, the number of random parameters range from 12 to 21, which are moderately high but typical for groundwater modeling. Due to the lack of reference values for the model probabilities, a predictive performance is conducted to investigate practical advantages of TIE. The MCMC simulation needed by HME and TIE is conducted using the MCMC function of UCODE_2014 software recently developed by *Lu et al*. [2014] based on the DREAM code of *Vrugt et al*. [2008, 2009].

3.1. Linear Analytical Function

The linear test function is

$$y=ax+m+\varepsilon, \tag{31}$$

where the true parameter values are $a=2$ and $m=3$. Twenty samples of y are first generated with $\mathbf{x}=\{1,2,\ldots,20\}$, and subsequently corrupted using one realization of white noise, ε, with mean zero and variance $\sigma^2=1$. The joint likelihood function, $p(\mathbf{D}|\boldsymbol{\theta},M)$, is Gaussian; with the conjugate priors of $a\sim N(2,1)$ and $m\sim N(3,1)$, the analytical solution of the marginal likelihood is evaluated using Mathematica. It however should be noted that analytical solutions are only available for special cases such as linear models and Gaussian likelihood functions and priors [e.g., *Schöniger et al*., 2014]. Before discussing the results of estimating the marginal likelihood, we provide the detailed procedure of selecting the discrete power coefficient values of TIE and the procedure for tuning NES. Because the procedure is similar for all the three cases, such details are not provided for the other three cases.

3.1.1. Selecting TIE Discrete Power Coefficient Values

Since there has been no theoretical method for selecting the discrete power coefficient values, this is done in an empirical but straightforward manner in this study. TIE accuracy heavily depends on the location and number of the discrete power coefficients. For example, using the trapezoidal integration given in equation (23) with 11 equally distributed β values, $\beta=\{0,0.1,0.2,\ldots,1\}$, the TIE estimate is 2.43×10^{-44}, which is an immense underestimation in comparison with the reference value of 5.1194×10^{-14} (the details of determining the reference value are given in section 3.1.3). On the other

Table 1. Numerical Estimates and Their Relative Errors for Calculating Marginal Likelihood of a Linear Function and a Nonlinear Function, Each Having Two Parameters[a]

	Linear Function		Nonlinear Function	
Method	Estimate	Relative Error	Estimate	Relative Error
Reference	5.1194×10^{-14}		1.1667×10^{-14}	
Laplace-MAP	5.1194×10^{-14}	0%	1.0478×10^{-14}	−10.19%
Laplace-MLE	5.6485×10^{-14}	10.34%	1.0599×10^{-14}	−9.15%
AME	5.1114×10^{-14}	−0.16%	1.1669×10^{-14}	0.02%
HME	3.7138×10^{-13}	625.43%	2.1076×10^{-24}	−100%
TIE	5.1010×10^{-14}	−0.36%	1.1682×10^{-14}	0.13%
NSE	5.0804×10^{-14}	−0.76%	1.1684×10^{-14}	0.15%

[a] The reference value is analytical for the linear function but numerical for the nonlinear function.

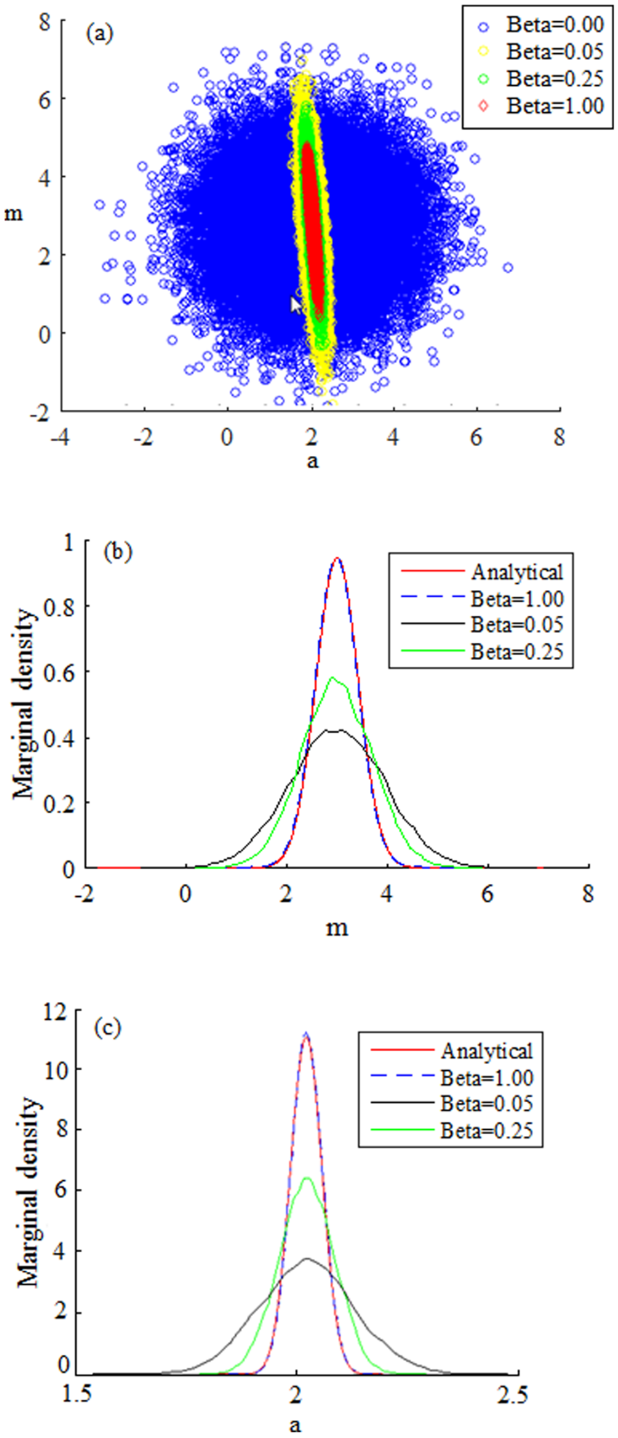

Figure 4. (a) TIE-related MCMC samples of *a* and *m* for different β values, and (b and c) marginal density functions of *m* and *a* based on analytical solution and TIE-related MCMC samples for the linear function.

hand, using only five power coefficients, $\beta=\{0, 0.001, 0.01, 0.1, 1\}$, the TIE estimate is 2.88×10^{-15}, only 1 order of magnitude smaller than the reference value. Therefore, the key to obtaining accurate TIE is to determine the β values at the important locations shown in the example below.

Our practice of determining the β values is to start with evenly distributed β_k values and then examine the variation of y_k (used in equation (23)) with the β_k values. Extra β_k values are added to the locations where y_k changes dramatically with β_k. An illustration is given in Figure 1. The figure shows that, for the 11 uniform distributed β_k (red line), y_k changes dramatically between 0 and 0.1. Therefore, nine values (green symbols) are added to this interval with the smallest one being 0.01, which improves TIE from 2.43×10^{-44} to 1.47×10^{-16}. Since y_k still changes dramatically between 0 and 0.01 (Figure 1), nine values (blue symbols) are added to this interval with the smallest one being 0.001, which improves TIE from 1.47×10^{-16} to 4.93×10^{-14}, only 3.79% smaller than the reference value. Adding nine extra β_k values between 0.1 and 1 only slightly improves the TIE accuracy from 4.93×10^{-14} to 5.10×10^{-14}. Since this result is only 0.36% less than the reference value, no more β_k values are added. As demonstrated in this procedure, while the β_k values are unknown, they can be selected in a systematic way to obtain accurate TIE. The rule of thumb is to add more β_k values near $\beta=0$, where the shape of y_k always change dramatically. Although using more β_k values increases computational cost, given that each β_k is independent, evaluating y_k for each β_k can be done in parallel to improve computational efficiency.

3.1.2. Tuning NSE

A total of four variables are tuned to obtain efficient and accurate NSE estimates. The first two are deterministic parameters that do not vary when evaluating NSE, and they are the size, N_{as}, of the active set (used in Step 1 described in section 2.4) and the number, N_{MCMC}, of MCMC samples (used in Step 2.2 described in section 2.4). The number of MCMC samples needs to be carefully selected for balancing efficiency and premature termination discussed in section 2.4. Based on trials and errors, $N_{MCMC}=10\times N_{as}$ is determined. This is the number of model executions in each MCMC simulation for determining X_i. N_{as} is the most important parameter for NSE, because it determines

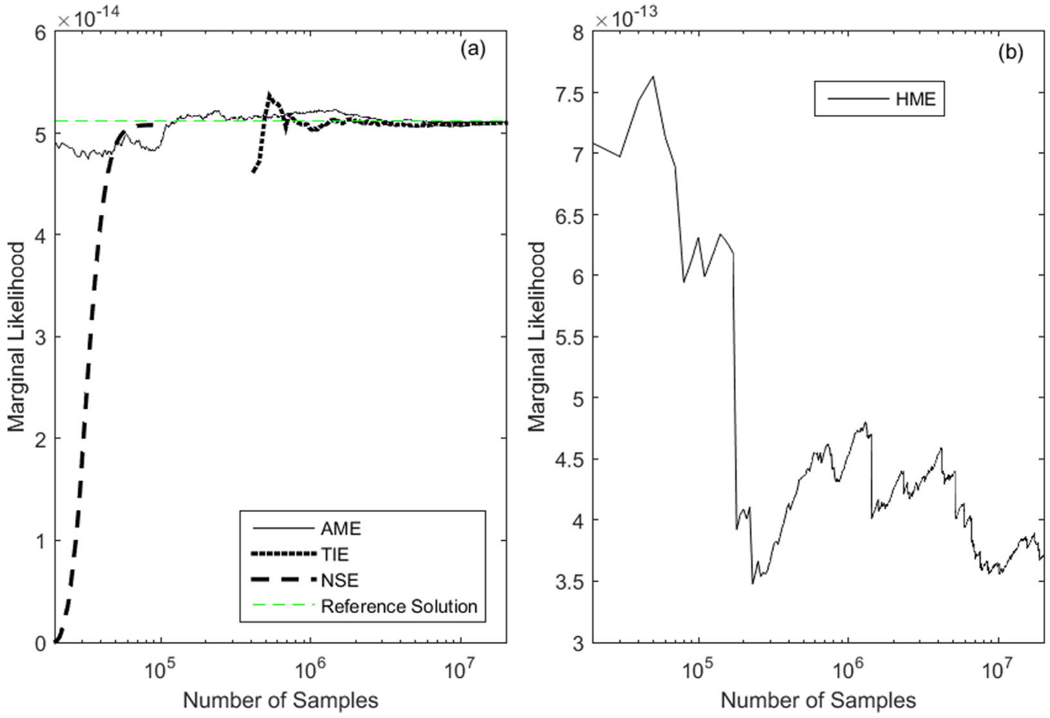

Figure 5. Convergence of (a) AME, TIE, and NSE and (b) HME for the linear function.

the number of X_i used in NSE and thus NSE accuracy. As shown in Figure 2, a total six values of N_{as} are considered; for each N_{as}, 50 repeated runs are conducted. Figure 2a plots the boxplots of the 50 NSE estimates for each N_{as}. The figure shows that, when N_{as} increases from 10 to 200, NSE estimate becomes more accurate and more consistent. This is reasonable, because more samples in the active set allow a better chance to explore the parameter space for constructing the X space. However, increasing N_{as} results in a dramatic increase in computational cost. Figure 3b shows that, when N_{as} increases from 10 to 200, the number of samples (i.e., number of model executions) increases 3 orders of magnitude, from 10,000 to 10 million. To balance between computational cost and accuracy, $N_{as} = 25$ is used.

This study also tunes two other parameters that are random during the NSE evaluation. The first one is the random number, ω, used in equation (29); in addition to the normal distribution, $N(0,1)$, two other distributions, $N(0,0.5)$ and U(0,1), are also used. The other parameter is the randomized step-size reduction factor, R, used in equation (30). It is either fixed at $R = 1$ or sampled randomly from U[0,1]. These variations are used to determine whether smaller update of the evolving samples increase accuracy and

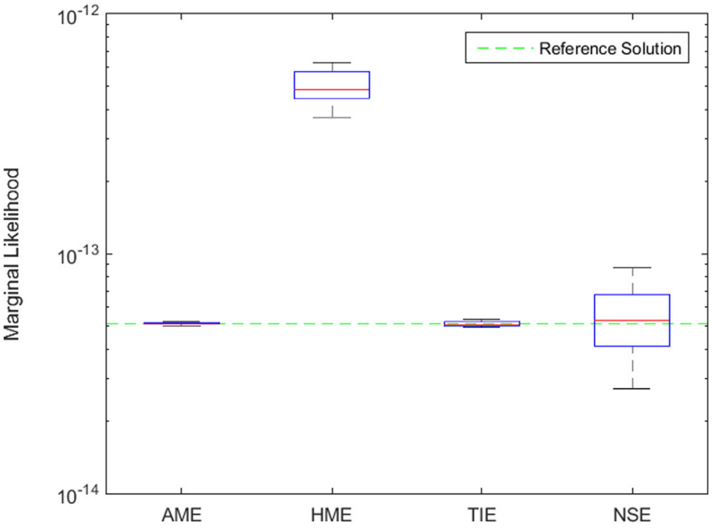

Figure 6. Accuracy and consistency of AME, HME, TIE, and NSE for the linear function based on 10 repeated runs.

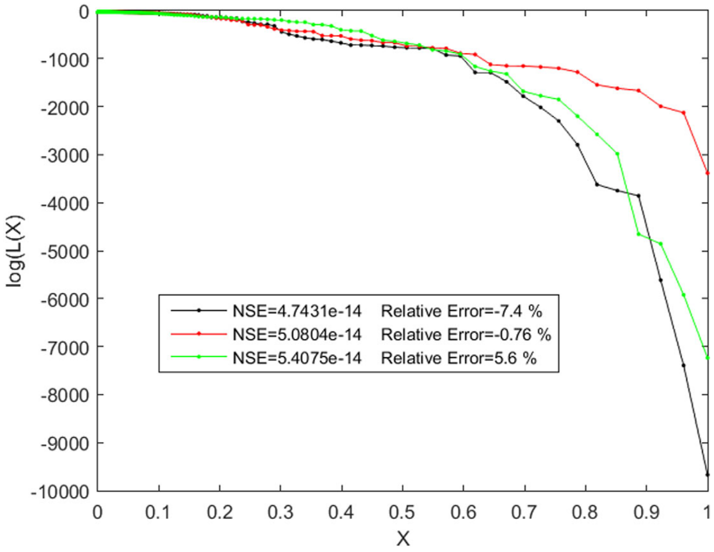

Figure 7. Variation of log likelihood (Log(L(X)) for three repeated NSE runs for the linear function.

efficiency of NSE, because smaller update increases the chance of finding next X_i before NSE iteration is terminated. Based on 50 repeated runs, Figure 3 shows the boxplots of the NSE estimates for the three distributions with either fixed or random *R*. The first two boxplots in the figure indicate that, for $N(0,1)$, using either fixed or random *R* does not affect the accuracy and consistency. The computational cost of the two cases are similar; the number of model executions for $N(0,1)_{nr}$ (fixed *R*) and $N(0,1)_r$ (random *R*) are $1.05\times10^5\pm2.75\times10^3$ and $1.04\times10^5\pm2.64\times10^3$, respectively. The same conclusion is also drawn for the distribution of *N*(0,0.5). For the distribution of U(0,1), poor NSE estimates can be obtained, and this distribution should not be used.

3.1.3. Accuracy of Numerical Solutions

Table 1 lists the numerical solutions of the marginal likelihood given by Laplace-MAP, Laplace-MLE, AME, HME, TIE, and NSE. For the linear model and Gaussian likelihood function and priors, the analytical solution of the marginal likelihood is of 5.1194×10^{-14}, calculated using the commercial software, Mathematica. It is labeled as the reference value in Table 1. The solutions of Laplace-MAP and Laplace-ML are also obtained using Mathematica. For each of the four MC methods, 10 repeated runs are conducted, and the best results (closest to the reference value) are listed in Table 1. The number of samples for each run is given in section 3.1.4 where convergence of the MC methods is discussed.

Table 1 shows that the Laplace-MAP solution is exactly the same as the reference solution. This is expected for the linear function and Gaussian likelihood and priors, as explained in section 2. Laplace-MLE overestimates the marginal likelihood by about 10.34%, which is caused by the determinant of the Fisher information matrix. For the MAP parameter estimates ($\tilde{a}=2.0255$, $\tilde{m}=2.9978$), the determinant is 0.007859; for the ML parameter estimates ($\hat{a}=2.0256$, $\hat{m}=2.9969$), the determinant becomes 0.008671. The Fisher information matrix is not always the reason for inaccurate Laplace-MLE results. In the third example with the ten-dimensional analytical function, the prior probability is the major factor affecting accuracy of Laplace-MLE, when an informative prior is used. The AME solution and the reference solution are almost identical, with the relative error being -0.16%, because a large number of prior samples are used for this two-dimensional problem. TIE and NSE also give accurate estimates with the relative error being -0.36% and -0.76%, respectively.

HME substantially overestimates the marginal likelihood by about 1 order of magnitude. This is not caused by inefficiency of MCMC sampling conducted for evaluating HME, because Figures 4b and 4c show that the marginal posterior parameter distributions ($\beta=1$) are visually identical to their analytical counterparts evaluated using Mathematica. The reason that HME overestimates the marginal likelihood is that the method uses only posterior samples that only occupy a small portion of the prior parameter space. This is shown in Figure 4a that plots the MCMC parameters for different β values. The figure shows that the posterior space ($\beta=1$) is significantly narrower than the prior space ($\beta=0$). In fact, the parameter samples gradually spread from the posterior out to the prior space. This is also observed in the marginal density plots of the two parameters in Figures 4b and 4c; the marginal posterior density functions ($\beta=1$) are significantly narrower than those of the other two β values of 0.05 and 0.25. Therefore, if the posterior samples ($\beta=1$) are used for evaluating HME, HME overestimates the marginal likelihood. This also explains why TIE is accurate,

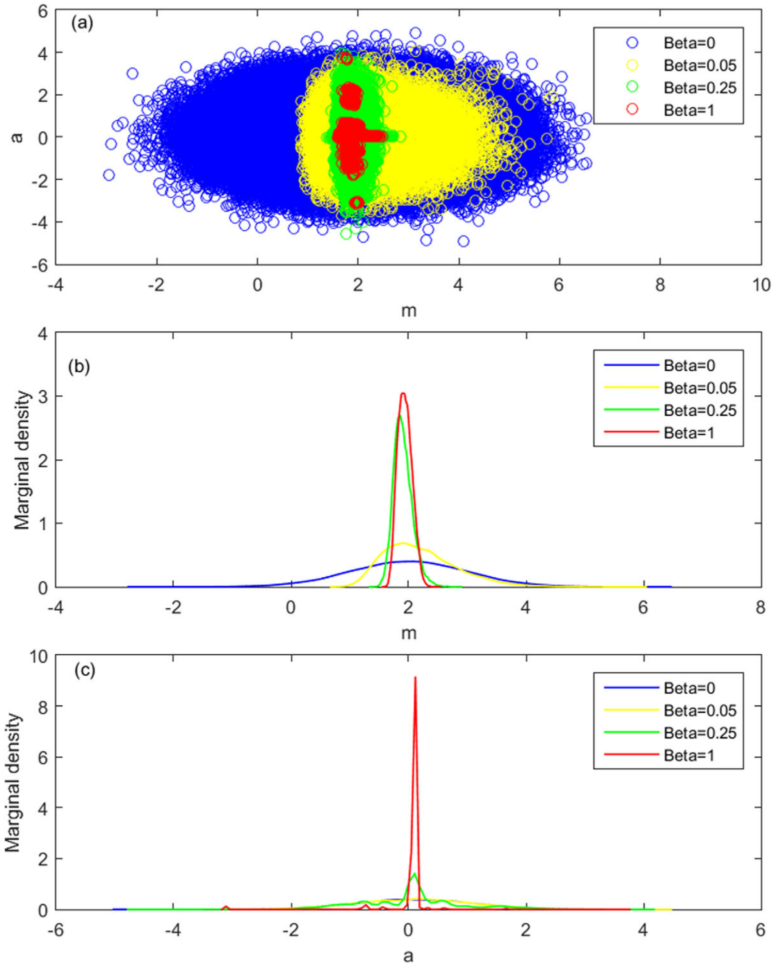

Figure 8. (a) TIE-related MCMC samples of *a* and *m*, and (b and c) marginal density functions of *m* and *a* for different β values for the two-dimensional nonlinear function.

because more parameter samples (from $\beta = 0$ to $\beta = 1$) with low values of the joint likelihood are used for calculating the marginal likelihood. It should be noted that HME is inaccurate only when the posterior distributions are narrower than the prior distributions. In an extreme case that the prior and posterior distributions are identical, HME and TIE should yield the same result.

3.1.4. Convergence and Consistency of MC Solutions

Figure 5 plots the convergence of the marginal likelihood estimates given by AME, HME, TIE, and NSE. As mentioned above, for each of the four estimators, 10 repeated MC runs are conducted, and the best run with the estimates closest to the reference value is used for the convergence analysis. In each run, 20 million samples are used for AME; 20 millions of parameter samples are used for HME, and the number of burn-in sample is 100,000. For each of the 38 β_k values of TIE, 500,000 samples are used, and the number of burn-in sample is 10,000. Therefore, the total function evaluation is 19 million. Because of the burn-in samples, the convergence profile of TIE starts from the sample size of 380,000. For NSE, the number of function evaluation is about 100,000.

Figure 5a shows that NSE has the fastest convergence rate, stabilizing with less than 100,000 samples. Although AME reaches the reference value after about 200,000 samples, AME increases again after about 400,000 samples and then converges after about 2 million samples. In this sense, TIE converges faster than AME, because TIE converges after about 1 million samples. Figure 5b shows that HME diverges. The sudden drops of HME are caused by the parameter samples that have small values of joint likelihood but are accepted with probability in the Metropolis sense. Because HME uses the posterior sample, its value is consistently larger than the reference value. Figure 5 indicate that HME is the worst estimator and that NSE is the most computationally efficient one.

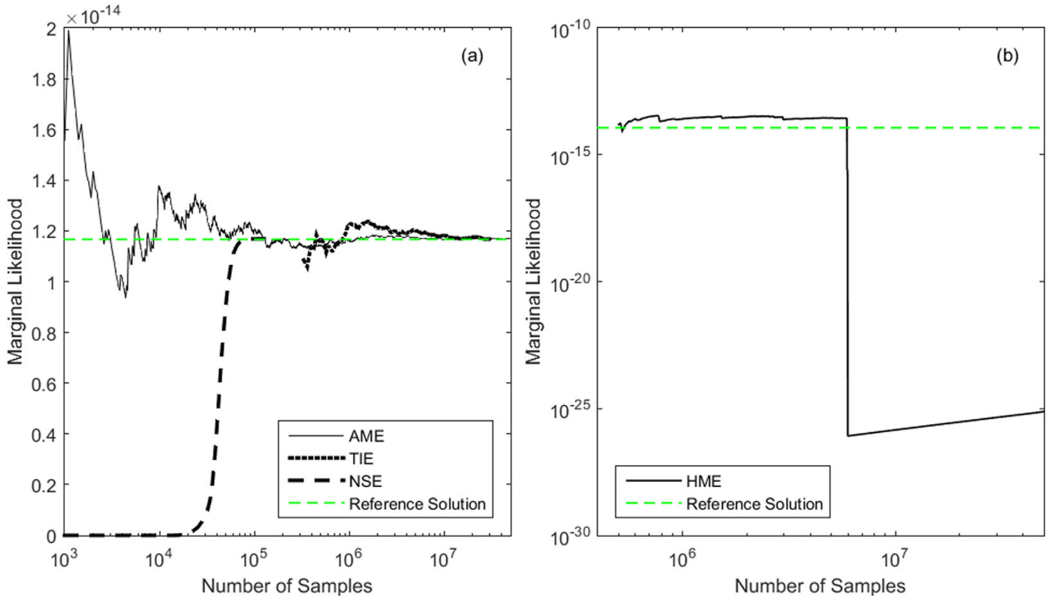

Figure 9. Convergence of (a) AME, TIE, and NSE and (b) HME for the two-dimensional nonlinear function.

Figure 6 shows the boxplots of the estimates of AME, HME, TIE, and NSE based on the 10 repeated model runs mentioned above. The figure shows that, while NSE is computationally more efficient than TIE, TIE is significantly more consistent than NSE in that the results of TIE have a significantly smaller variability. The NSE variability is not surprising, because, whenever a new active set is used and new proposal samples are drawn, the likelihood values used to evaluate equation (28) change. This is illustrated in Figure 7 for three repeated NSE runs; log likelihood values are plotted for demonstration. The figure shows that the variation starts after $X=0.2$ and becomes substantial after $X=0.6$. It is thus concluded that, although NSE is computationally more efficient than TIE, NSE may not provide reliable estimate of the marginal likelihood. This problem becomes worse in the third case with a wide prior range, because NSE cannot find parameter samples with high likelihood for accurate estimate of the marginal likelihood.

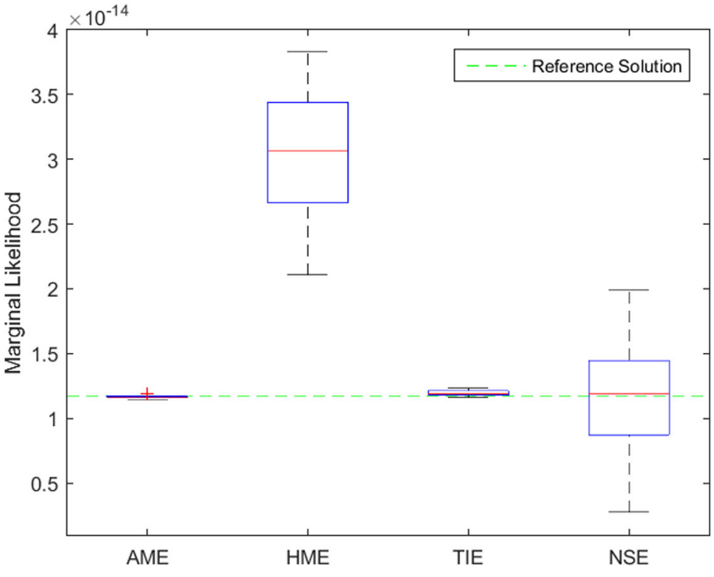

Figure 10. Accuracy and consistency of AME, HME, TIE, and NSE for the two-dimensional nonlinear function based on 10 repeated runs.

3.2. Two-Dimensional Nonlinear Analytical Function

Consider a two-dimensional nonlinear function,

$$y = x/a + \sin(amx) + \varepsilon, \qquad (32)$$

where the true parameter values are $a=2$ and $m=0.1$. Twenty samples of y are generated with $x=\{1, 2, \ldots 20\}$, and subsequently corrupted using one realization of white noise, ε, with mean zero and variance, $\sigma^2=1$. The posterior parameter distribution of parameter, a, is multimodal due to the sine function and the multiplication of a and m. This is shown in Figure 8 that plots the posterior parameter samples and marginal distributions of the two

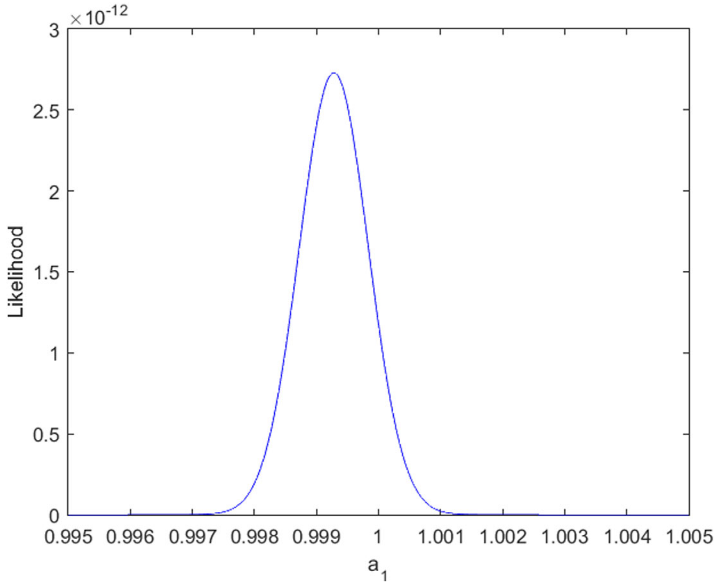

Figure 11. Joint likelihood for parameter a_1 of the ten-dimensional nonlinear function.

parameters obtained from the MCMC simulation with different β values; those of $\beta = 1$ are the posterior samples and distributions. Figure 8c shows that there are multiple modes on the distributions of *a*; the mode near the true values are significantly larger than other modes. Because the multimodal distribution, more realizations are needed for the MCMC simulations.

3.2.1. Accuracy of Numerical Solutions

Table 1 lists the numerical solutions of the marginal likelihood given by Laplace-MAP, Laplace-MLE, AME, HME, TIE, and NSE. Since analytical solution of the marginal likelihood is unavailable for this nonlinear function, a numerical reference value is obtained using AME with 1 billion samples, which givens an estimate of 1.1667×10^{-14} with a standard deviation of 1.65×10^{-17}. The solutions of Laplace-MAP and Laplace-ML are obtained using Mathematica. For the MC solutions, 10 repeated runs are conducted, and the best results (closest to the reference value) are listed in Table 1. In each run, about 50 million model executions are conducted for AME, HME, and TIE to ensure convergence. For TIE, by following the procedure described in section 3.1, 30 discrete β values are determined. For NSE, by following the procedure described in section 3.2, the size of active set is chosen to be 25, and the distribution of ω is set as *N* $(0, 1)$ without using the randomized step-size reduction factor.

Table 1 shows that Laplace-MAP and Laplace-MLE underestimate the marginal likelihood by −10.19% and −9.15%, respectively. This may be attributed to the multimodal nature of the parameter distributions, recalling that Laplace-MAP and Laplace-ML only evaluate the marginal likelihood around the MAP and ML parameter estimates, as explained in section 2.1. The underestimation is relatively small, because the modes other than the MAP and ML are small, as shown in Figure 8c. For the MC solutions, similar to those of the linear function, AME gives the most accurate estimate with the relative error of 0.02%; the relative errors of TIE and NSE are slightly larger than 0.1%. HME still give the worst result, because of the divergence problem that is shown in Figure 9.

3.2.2. Convergence and Consistency of MC Solutions

Figure 9 plots the convergence of the marginal likelihood estimates given by AME, HME, TIE, and NSE. For each of the four estimators, 10 repeated MC runs are conducted, and the best run with the estimates closest to the reference value is used for the plot and the convergence analysis. Figure 9a shows that, for the two-dimensional nonlinear function, NSE is the most computationally efficient method, converging to the

Table 2. Numerical Estimates and Their Relative Errors For Calculating Marginal Likelihood of a Nonlinear Function With 10 Parameters[a]

	Narrow Prior ($\sigma = 0.01$)			Wide Prior ($\sigma = 1$)		
Method	Mean	Relative Error	Std	Mean	Relative Error	Std
Reference	4.8999×10^{-14}			5.6537×10^{-14}		
Laplace-MAP	4.9077×10^{-14}	0.16%		7.4233×10^{-15}	−86.87%	
Laplace-MLE	6.244×10^{-126}	−100		3.4371×10^{-13}	507.92%	
AME	4.9094×10^{-14}	−0.07%	3.03×10^{-17}	9.2513×10^{-22}	−100	3.51×10^{-21}
HME	6.2256×10^{-19}	−100%	9.86×10^{-19}	4.1430×10^{-27}	−100	5.86×10^{-27}
TIE	4.8838×10^{-14}	−0.38%	1.32×10^{-15}	6.1139×10^{-14}	8.14%	3.26×10^{-15}
NSE	5.9961×10^{-14}	22.37%	6.80×10^{-14}	1.3545×10^{-105}		4.06×10^{-105}

[a] A narrow and a wide priors are used for the evaluation. The reference values are numerical for the two priors.

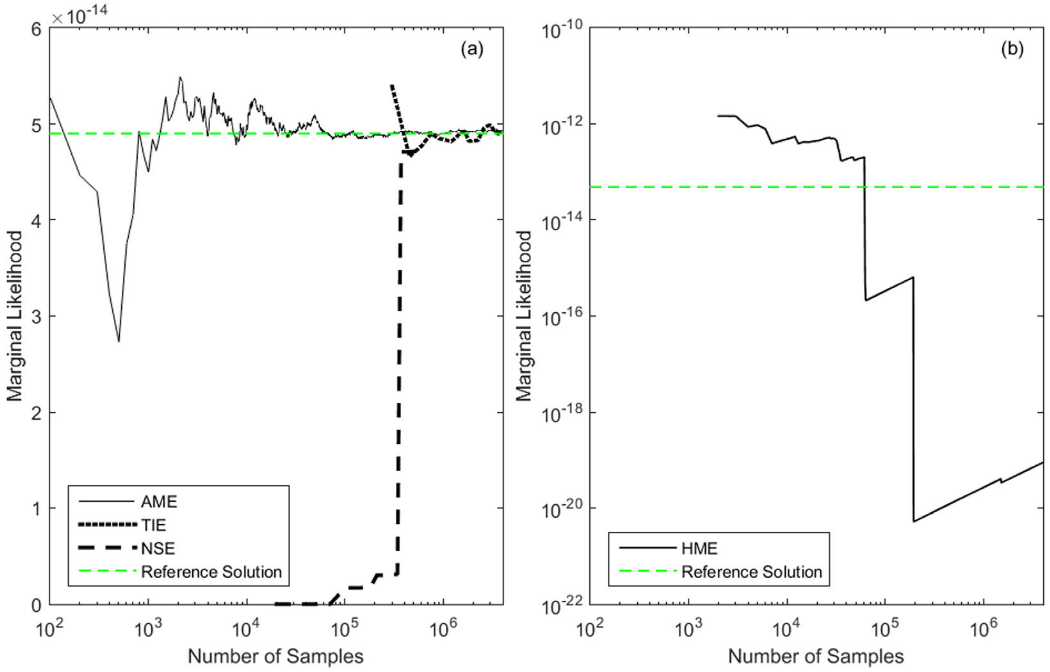

Figure 12. Convergence of (a) AME, TIE, and NSE and (b) HME for the ten-dimensional nonlinear function.

reference value with less than 100,000 samples. AME converges with less than 1 million samples. TIE is more computationally demanding, and needs 10 million samples to converge. However, the computational cost of TIE depends on the required accuracy. Although 10 million samples are needed to reach the relative error of 0.13% (Table 1), only 500,000 samples are needed to reach the relative error of 5%. Figure 9b shows that HME diverges after about 5 million samples, as it drops dramatically for about 10 orders of magnitude. The drop is attributed to a sample with an extremely small joint likelihood. After the sudden drop, HME does not return to the best estimate even after 50 millions of samples. The HME value listed in Table 1 is the best estimate before the sudden drop occurs.

Figure 10 shows the boxplots of the estimates of AME, HME, TIE, and NSE based on the 10 repeated model runs mentioned above. While NSE is computationally more efficient than TIE and accurate on average, TIE is significantly more consistent than NSE in that the results of TIE have a significantly smaller range. As explained for the linear function, the inconsistency is inherent to NSE due to the high uncertainty in selecting the likelihood values for computing the marginal likelihood. Given that NSE is accurate on average and computationally more efficient, one may conduct a large number of repeated runs to obtain reliable NSE estimate of the marginal likelihood.

3.3. Ten-Dimensional Nonlinear Analytical Function

The two cases above appear to suggest that AME is the best MC approaches, because it is conceptually straightforward, provides accurate results, and converges fast. For the same reasons, NSE is the second best MC approach. These conclusions are incorrect for the ten-dimensional, nonlinear function defined as,

$$\begin{aligned} y &= \sum_{i=1}^{n-1} (a_i - 1)^2 + \frac{1}{2}\left(x^2 - a_{i+1}\right)^2 + \varepsilon \\ &= (a_1 - 1)^2 + \frac{1}{2}\left(\mathbf{x}^2 - a_2\right)^2 + \cdots + (a_{n-1} - 1)^2 + \frac{1}{2}\left(\mathbf{x}^2 - a_n\right)^2 + \varepsilon, \end{aligned} \tag{33}$$

where $n = 10$ is the number of parameter, $\mathbf{a} = \{a_1, a_2, \ldots, a_{10}\}$. The true value of each parameter is taken as one. Twenty samples of y are generated with $\mathbf{x} = \{1, 2, \ldots, 20\}$ and corrupted using one realization of white noise with mean zero and variance, $\sigma^2 = 1$. The likelihood of this ten-dimensional analytical function has a sharp peak that decreases to zero quickly within a narrow region in the parameter space. This is illustrated in Figure 11 for parameter a_1, with the other nine parameters fixed at their true values. The parameter range with nonzero likelihood is smaller than 0.004. The narrow range, as explained below, makes it difficult to obtain accurate estimates of the marginal likelihood, especially when the prior is wide. Using again the Gaussian prior for each parameter, two situations with small and large variance are considered. For the narrow prior of $N(1, 0.01)$, the numerical approximations (except Laplace-ML and HME) give satisfactory results; for the wide prior of $N(1, 1)$, only TIE gives satisfactory results.

3.3.1. Narrow Prior of N(1,0.01)

Table 2 lists the numerical solutions of the marginal likelihood given by Laplace-MAP, Laplace-MLE, AME, HME, TIE, and NSE. Since analytical solution of the marginal likelihood is unavailable for this nonlinear function, a numerical reference value is obtained using AME with 1 billion samples, which givens an estimate of 4.8999×10^{-14} with the standard deviation of 3.08×10^{-17}. The solutions of Laplace-MAP and Laplace-ML are obtained using Mathematica. For TIE, by following the procedure described in section 3.1, 12 discrete β values are determined. For NSE, by following the procedure described in section 3.2, the size of active set is chosen to be 25, and the distribution of ω is adjusted to $N(0, 0.05)$ for the narrow prior; the step-size reduction factor is fixed during the simulation. For the four MC approaches (AME, HME, TIE, and NSE), 10 repeated MC runs are conducted, and the best results (close to the reference value) are listed in Table 2.

Table 2 shows that, while Laplace-MAP is accurate, Laplace-MLE substantially underestimates the marginal likelihood. The underestimation is attributed to the small value (6.244×10^{-126}) of $p(\hat{\boldsymbol{\theta}}|M)$ (i.e., evaluation of the prior at the maximum likelihood parameter estimate used in equation (8)) due to the narrow prior. Taking parameter a_1 as an example, its maximum likelihood estimate is 1.069, and the prior evaluated at this value is close zero (Figure 11). This problem does not occur for Lapalce-MAP, whose $p(\tilde{\boldsymbol{\theta}}|M)$ value is 1.018×10^{16}. The reason is that using the narrow prior substantially increases the accuracy of the maximum a posterior parameter estimates. For example, the estimate of a_1 is 1.00045, close to the true value of one. It should be noted that the substantial impact of prior on estimating the marginal likelihood is not common in practice, because narrow priors are seldom used and uniform prior is used more often than Gaussian prior.

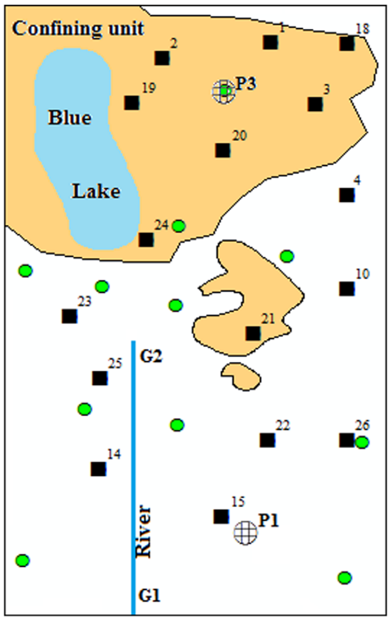

Figure 13. Modeling domain of the true model with the confining unit in yellow, Blue Lake, and Straight River. Measurements of hydraulic conductivity and hydraulic head are available at the locations marked by black squares. Only head observations are available at the locations marked by the green circles.

For the MC solutions, similar to those of the previous two cases, AME gives the most accurate estimate with the relative error of 0.07%. This is not surprising, because of millions of parameter samples are generated for the narrow prior space. TIE is more accurate than NSE, as the relative errors of TIE and NSE are 0.38% and 22.37%, respectively. The accuracy of NSE is substantially affected by the prior range and shape of the likelihood function. For the wide prior, NSE gives unacceptable estimate of the marginal likelihood, and more details are given below. HME still give the worst result, because of the divergence problem shown in Figure 12.

Figure 12 plots the convergence of the marginal likelihood estimates given by AME, HME, TIE, and NSE. For each of the four estimators, 10 repeated MC runs are conducted, and the best run with the estimates (closest to the reference value) is used for the plot and the convergence analysis. Because of the narrow prior, the MC simulations converge faster than the previous two cases, and only 4 million samples are needed for each MC method. For the narrow prior, AME converges after about 600,000 samples. NSE needs about 1 million samples mainly because the peaked likelihood, and TIE needs about 3

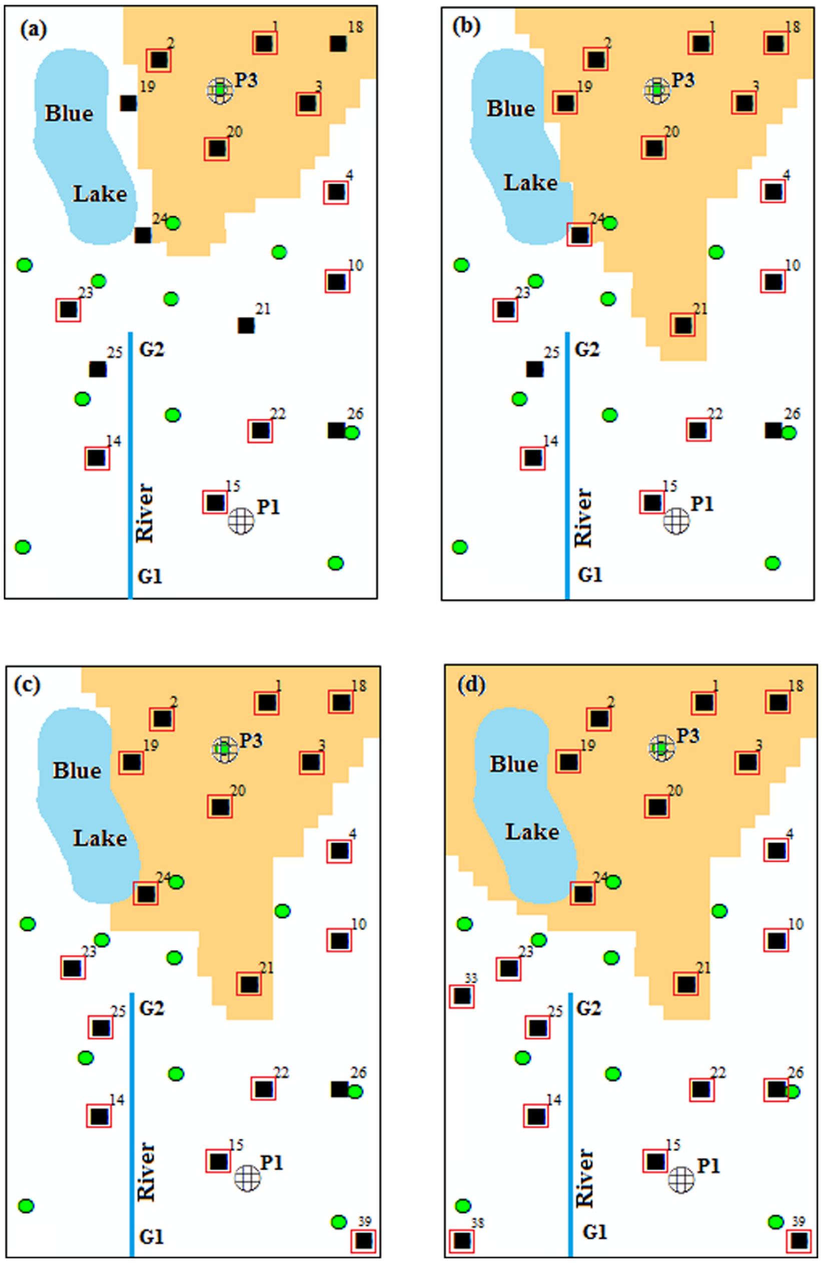

Figure 14. Four alternative groundwater models with different geometric configurations of the confining layer highlighted in yellow. The models also have different number of calibrated hydraulic conductivity, and their locations are highlighted with red squares.

million samples. The convergence behavior of HME is similar to that of the two-dimensional nonlinear case shown in Figure 9, dropping suddenly when a sample with a dramatically small likelihood is used for evaluating HME.

For examining consistency of the four MC approaches, Table 2 lists the standard deviations of the MC estimates based on the 10 repeated runs based on 4 million. The boxplots used for the previous two cases are not used, because the mean and standard deviation vary by several orders of magnitude. The standard deviations show again that NSE estimates have the largest variability, indicating inconsistency of the NSE results. In the numerical case with wide prior shown below, NSE fails to estimate the marginal likelihood.

3.3.2. Wide Prior of N(1,1)

Table 2 lists the numerical solutions of the marginal likelihood given by Laplace-MAP, Laplace-MLE, AME, HME, TIE, and NSE for the wide prior. Since AME does not converge with 1 billion samples for the wide

Table 3. Values of AME, HME, TIE, and Their Corresponding Model Averaging Weight (%) of Four Alternative Models

	A	B	C	D
Laplace-MAP	7.8816×10^{-32}	5.2905×10^{-26}	8.3172×10^{-25}	1.7860×10^{-23}
Laplace-MAP-based probabilities	0.00%	0.29%	4.43%	95.28%
AME	5.3421×10^{-47}	2.0904×10^{-45}	6.3129×10^{-42}	7.3611×10^{-40}
AME-based probabilities	0.00%	0.00%	0.85%	99.15%
HME	6.3991×10^{-24}	4.8562×10^{-20}	8.9377×10^{-18}	3.4615×10^{-16}
HME-based probabilities	0.00%	0.01%	2.52%	97.47%
TIE	1.0420×10^{-42}	1.5356×10^{-39}	6.3992×10^{-39}	4.3887×10^{-38}
TIE-based probabilities	0.00%	2.96%	12.35%	84.69%

range, the numerical reference value of 5.6537×10^{-14} is obtained using TIE with 21 beta values and about 100 million samples. Laplace-MAP and Laplace-MLE are evaluated using Mathematica. For each of the four MC approaches, 10 repeated runs are conducted, and each uses about 10 million samples. The best results (closest to the reference) are listed in Table 2. Laplace-MAP and Laplace-MLE are evaluated using Mathematica.

Table 2 shows that the result of Laplace-MAP is about 1 order of magnitude smaller than the reference value, and that of Laplace-MLE is about 1 order of magnitude larger. Despite of the errors, these Laplace results are still significantly more accurate than those of AME, HME, and NSE. The inaccuracy of AME is attributed to its slow convergence, and the inaccuracy of HME is attributed to its erroneous convergence whenever samples of small likelihood are used for evaluating HME.

Table 2 shows that TIE outperforms NSE. TIE gives the best result that is only 8.14% larger than the reference value. The NSE result is enormously small. This however is not surprising when considering the sampling strategy used for NSE in this study and the peak likelihood with a small region of the parameter space. When parameter samples are drawn from the prior distribution for an active set, for the wide prior, it is highly likely that the samples have small likelihood. When MCMC simulation described in section 2.4 is conducted, due to the highly peaked likelihood function, the chance of finding new samples with larger likelihood is low. As a result, small likelihood is used in equation (28) for evaluating NSE. This problem cannot resolved by tuning the Metropolis-Hasting algorithm, such as using larger MCMC samples, several different proposal distributions used in equation (29), and different initial step size and step-size reduction factor used in equation (30). The only solution to this problem seems to use advanced MCMC algorithms for NSE, such as MT-DREMA$_{(ZS)}$ [*Laloy and Vrugt*, 2012], which however is beyond the scope of this study.

3.4. Alternative Groundwater Models

The exercises above for the analytical functions are extended to a synthetic case of groundwater modeling with a "true" model and four alternative models. The true model is the same as the synthetic model of *Lu et al*. [2012a], and it is used to generate data for model calibration and predictive analysis. For the confining layer of the true model shown in Figure 13, four geometric configurations are proposed, and they lead to four alternative models (A–D) shown in Figure 14. Model D is the closest to the true model, and Model A is the worst because it does not include the south portion of the confining layer. Except the geometric configuration, other model components of the four models are the same as those of the INT model described in *Lu et al*. [2012a]. The model calibration data include 54 observations of hydraulic head from 27 wells shown in Figure 13 (two head from each well in layers 1 and 3), one lake stage observation of the Blue Lake, and two

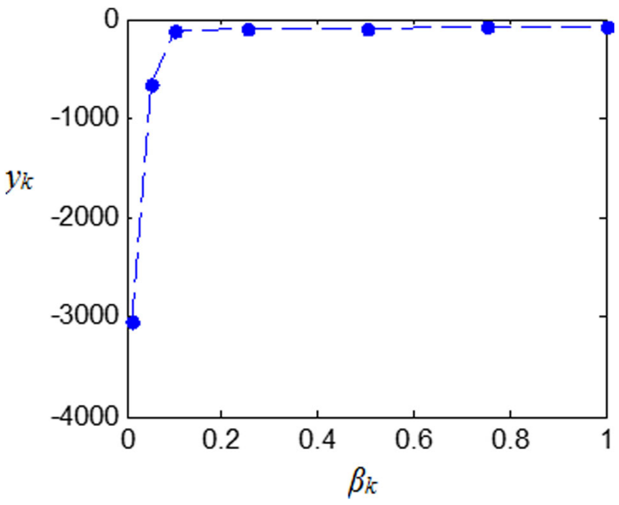

Figure 15. Variation of y_k (equation (24)) with β_k (power coefficient in equation (13)) for model D of the groundwater example.

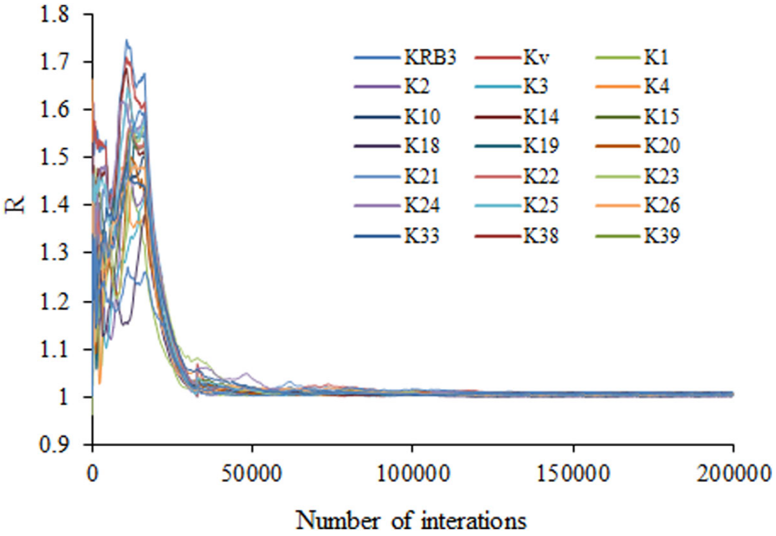

Figure 16. Variation of *R* statistics with number of MCMC samples for $\beta = 1$. The 200,000 samples ensure the MCMC convergence.

observations of streamflow gain at locations G1 and G2 along the Straight River shown in Figure 13. The calibrated parameters include leakance of the confining layer, conductance of the riverbed, and hydraulic conductivities. The number of calibrated parameters is 12, 16, 18, and 21 for models A–D, respectively. These number of model parameters are moderately high in groundwater modeling. The locations of the calibrated hydraulic conductivity are shown in Figure 14 for the four models. Prior information is used for all the parameters as described in *Lu et al*. [2012a].

Table 3 lists the marginal likelihood evaluated using AME, HME, and TIE. Due to the high computational cost of solving the groundwater models (relative to that for the analytical functions), obtaining reference values by running AME with billion samples is computationally unaffordable. Instead, 1.4 million samples are used for AME, and this also is the number of samples used for TIE. For each model, TIE uses seven β values, and Figure 15 plots the variation of y_k with β_k for Model D as an example; the relation between y_k and β_k is similar to this for the other three models. Although adding more β_k values may improve the TIE calculation, it is not attempted due to the computational cost. For each β_k value, 200,000 MCMC simulations are conducted. This large number of simulation ensures convergence of MCMC, as shown by the *R* statistics plotted in Figure 16 for $\beta = 1$ as an example. The number of samples used for calculating HME is also 200,000. NSE is not evaluated for this example, because of the computational cost.

Table 3 shows that the AME values are smaller than the TIE values, which is reasonable given that the dimensions of the groundwater models are moderately high. The HME values are consistently larger than the TIE values, as observed in the previous cases. The TIE results are believed to be more accurate than the AME and HME results, because posterior parameter distributions are narrower than the prior parameter distributions for most of model parameters. Figure 17 plots the marginal prior and posterior distributions of three parameters of Model D. The posterior distributions correspond to four β values. The three plots of in Figure 16 shows three different situations that the posterior distributions are significantly, moderately, and slightly smaller than the prior distributions. In these situations, underestimation of AME and overestimation of HME is unavoidable, and TIE should yield more accurate results than AME and HME.

Table 3 also lists the model probability calculated using the marginal likelihood and the uniform prior model probability of 25% for each model. The results of AME and HME are dramatically different from those of TIE. For AME and HME, Model D receives nearly 100% model probability. However, the TIE-based probability of model D is reduced to 84.69%. Correspondingly, the probability of model C increases from 2.52% (given by HME) to 12.35%, when TIE is used. The TIE-based model probabilities appear to be more reasonable, given the similarity in the configuration of the confining layer.

Since reference values of the marginal likelihood and model probability are unavailable, a predictive analysis is conducted to evaluate the effect of AME, HME, and TIE-based model probabilities on BMA

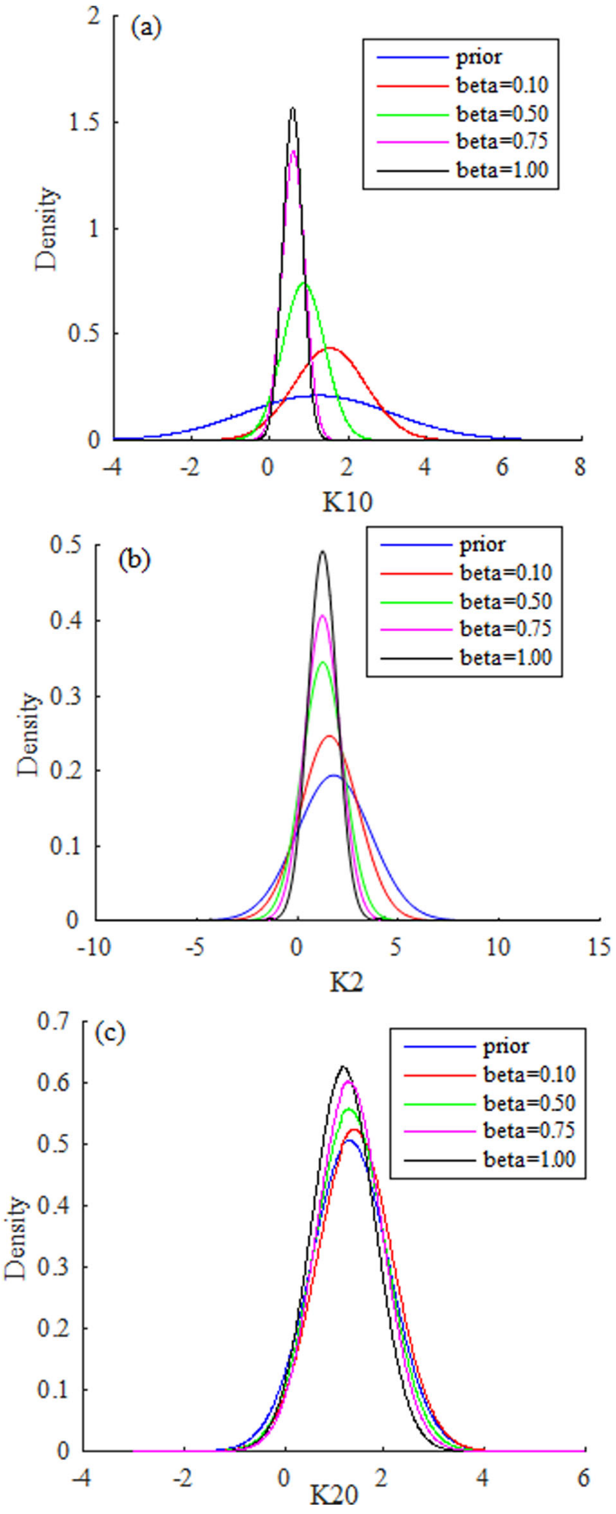

Figure 17. Marginal density functions of three parameters of Model D obtained using MCMC simulations with different β values. The posterior parameter distributions are (a) significantly, (b) moderately, and (c) slightly smaller than the prior parameter distributions.

model prediction. In the predictive analysis, the model prediction is the decrease of streamflow at gauge site G2 due to pumpage at wells P1 and P3 (locations shown in Figure 13). Figure 18 plots the 95% credible intervals of models A–D estimated from the MCMC simulation with $\beta=1$ (the asymmetric credible intervals indicates that the predictions are non-Gaussian). The figure shows that the mean prediction of model C is more accurate than that of model D, although the credible intervals of the two models are similar. Therefore, using the TIE-based model probability should yield better BMA predictive performance than using the AME and HME-based model probability, considering that TIE-based probability of Model C is 12.35%. This is confirmed by the logscore of BMA defined as [*Ye et al*., 2004]

$$-\ln p(\Delta^*|\mathbf{D})=-\ln\left(\sum_{k=1}^{K}p(\Delta^*|M_k,\mathbf{D})p(M_k|\mathbf{D})\right), \tag{34}$$

where Δ^* is the value of streamflow change (due to the pumping) simulated by the true model. The logscore of an individual model is defined as $-\ln p(\Delta^*|M_k,\mathbf{D})$. A smaller logscore means a larger probability of predicting the true value, and thus indicates better predictive performance. The TIE-based logscore is -3.65, smaller than the HME-based logscore of -3.62, indicating that TIE improves BMA predictive performance. It should be noted that, although the improvement is marginal, it is only for one prediction. More significant improvement is expected for a large number of predictions.

4. Conclusions and Discussion

This study evaluates several numerical methods for evaluating the marginal likelihood. The methods can be classified into two categories: Laplace approximation and Monte Carlo (MC) approximation. The Laplace approximation method has two variants: Laplace-MAP using maximum a posterior estimates of model parameters, and Laplace-MLE using maximum likelihood estimates of model parameters. The MC approximation method includes four estimators: arithmetic mean estimator (AME), harmonic mean estimator (HME), thermodynamic integration

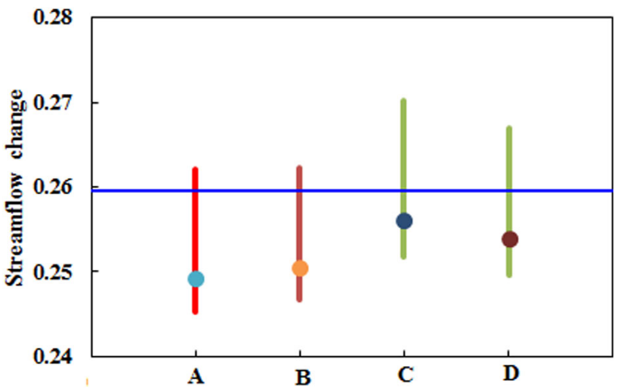

Figure 18. Median and 95% credible intervals of predicted streamflow change at gauge site G2 for models A–D. The horizontal blue line represents the true value of the prediction.

estimator (TIE), and nested sampling estimator (NSE). While Laplace-AME, Laplace-ML, AME, and HME have been used widely in environmental modeling, NSE has not been used until recently. It is the first time that TIE is used for quantifying model uncertainty in groundwater modeling. Three analytical functions are used to evaluate these numerical methods in terms of their accuracy, convergence, computational cost, and consistency. For each analytical function, the reference value of the marginal likelihood is obtained either analytically or numerically using a large number of MC simulations. For the evaluation using groundwater models, a predictive analysis is used to investigate the impacts of Laplace-MAP, AME, HME, and TIE on the calculation of model probability and BMA predictive performance.

This study leads to the following major conclusions:

1. The marginal likelihood estimates given by Laplace-AME and Laplace-MLE are overall comparable with those obtained using the MC methods, despite that the truncation error is inherent to the Laplace approximation method.
2. Laplace-AME gives more accurate results than Laplace-MLE, when informative priors are used, because the priors improve accuracy of parameter estimation. If it is difficult to obtain maximum a posterior parameter estimates, noninformative should be used to reduce the prior impacts on the Laplace-MLE evaluation.
3. Given that Laplace-AME and Laplace-MLE can be evaluated with low-computational cost, model probability should be calculated using the Laplace approximation method, before the computationally expensive MC method is attempted to obtain more accurate estimates.
4. For the MC-based numerical methods, the key to obtain an accurate estimate of the marginal likelihood is to use parameter samples that can sufficiently cover the entire parameter space. The reason that AME underestimates the marginal likelihood is that it uses prior samples that may not have enough samples from the posterior parameter space where the joint likelihood is high. The reason that HME overestimates the marginal likelihood is that it uses samples from the posterior parameter space where the joint likelihood is high. AME can be used to estimate the marginal likelihood, when a large number of samples are generated from the prior space that is low dimensional and/or has a narrow range. It is recommended not to use HME under any circumstances, because it diverges when samples with small joint likelihood are used for its evaluation. This problem cannot be resolve even a large number of samples are used for the evaluation.
5. TIE uses samples that are systematically generated from the prior to the posterior parameter space by conducting a path sampling with a number of discrete power coefficient values. TIE is mathematically rigorous, and its implementation is straightforward. The implementation is also general in that it can use any MCMC simulation for evaluating the y_k term used in equations (23) and (24). Although the discrete power coefficient values are unknown before starting the MCMC simulation, they can be determined in an empirical but objectively manner with negligible ambiguity. Due to the repeated MCMC simulations for a number of power coefficient values, TIE is computationally expensive. The computational burden may be alleviated, if the MCMC simulations for different power coefficient values are conducted in parallel.
6. Although NSE does not explicitly sample from the prior to the posterior parameter space, the procedure of constructing the X space and finding the corresponding $L(X_i|\mathbf{D}, M)$ values is equivalent to searching from the prior to the posterior parameter space. However, NSE cannot guarantees that samples are generated systematically from the prior to the posterior space. This makes NSE theoretically inferior to TIE, especially when the joint likelihood is peaked. How to construct the X space remains a challenging issue for NSE.

7. Because of the randomness in constructing the X space, NSE is less consistent than TIE in that NES estimates of repeated runs have larger variability than TIE estimates of repeated runs. However, since the NSE results are accurate on average, a large number of repeated runs may be conducted for NSE to alleviate the problem of inconsistency.

More research is warranted for evaluating TIE and NSE. One focus of the future research is to improve the sampling algorithms used for NSE. The currently used Metropolis-Hasting algorithm may yield unfavorable results for NSE, because the algorithm is inefficient to generate samples for high-dimensional parameter space and irregular joint likelihood (e.g., highly peaked and multimodal). NSE performance could be improved dramatically, if more advanced MCMC sampling methods were used. Another focus of the future research is to further evaluate TIE and NSE in groundwater modeling context, especially for reactive transport modeling whose likelihood surface is extremely irregular [*Shi et al*., 2014]. Since computational cost is always a barrier for extensive MC simulation in practice, the evaluation will have to rely on using surrogates of groundwater models. This is feasible, given that surrogate modeling has been widely used in uncertainty quantification of groundwater modeling [*Razavi et al*., 2012; *Laloy et al*., 2013; *Zhang et al*., 2013]. However, building accurate surrogates for highly nonlinear problems is another challenge that the community of groundwater modeling is facing, and interdisciplinary research with collaboration with the applied mathematics community is indispensable.

Acknowledgments
This work was supported in part by the Natural Science Foundation of China grants 51309071 and 51328902, U.S. DOE Early Career Award DE-SC0008272, and U.S. National Science Foundation grant DEB-1145999. The first author conducted part of the research when she was a visiting scholar at the Florida State University. We thank Mary Hill for her help on developing the alternative groundwater models. We comply with AGU's data policy, and the data and computer codes used to produce this paper are available on contact of the corresponding author at mye@fsu.edu.

References

Beerli, P., and M. Palczewski (2010), Unified framework to evaluate panmixia and migration direction among multiple sampling locations, *Genetics*, *185*, 313–326.

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.

Chib, S., and I. Jeliazkov (2001), Marginal likelihood from the Metropolis-Hastings output, *J. Am. Stat. Assoc.*, *96*(543), 270–281.

Elshall, A. S., and F. T.-C. Tsai (2014), Constructive epistemic modeling of groundwater flow with geological structure and boundary condition uncertainty under Bayesian paradigm, *J. Hydrol.*, *517*, 105–119, doi:10.1016/j.jhydrol.2014.05.027.

Elsheikh, A. H., M. F. Wheeler, and I. Hoteit (2013), Nested sampling algorithm for subsurface flow model selection, uncertainty quantification, and nonlinear calibration, *Water Resour. Res.*, *49*, 8383–8399, doi:10.1002/2012WR013406.

Friel, N., and A. N. Pettitt (2008), Marginal likelihood estimation via power posteriors, *J. R. Stat. Soc., Ser. B*, *70*, 589–607.

Friel, N., and J. Wyse (2012), Estimating the statistical evidence—A review, *Stat. Neer.*, *66*, 288–308, doi:10.1111/j.1467-9574.2011.00515.x.

Gelman, A., and X.-L. Meng (1998), Simulating normalizing constants: From importance sampling to bridge sampling to path sampling, *Stat. Sci.*, *13*, 163–185.

Han, C., and B. C. Carlin (2001), Markov chain Monte Carlo methods for computing Bayesian factors: A comparative review, *J. Am. Stat. Assoc.*, *96*(455), 1122–1132.

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

Kass, R. E., and A. E. Raftery (1995), Bayes factors, *J. Am. Stat. Assoc.*, *90*, 773–795.

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

Laloy, E., B. Rogiers, J. A. Vrugt, D. Mallants, and D. Jacques (2013), Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov Chain Monte Carlo simulation and polynomial chaos expansion, *Water Resour. Res.*, *49*, 2664–2682, doi:10.1002/wrcr.20226.

Lartillot, N., and H. Philippe (2006), Computing Bayes factors using thermodynamic integration, *Syst. Biol.*, *55*(2), 195–207.

Lemke, L. D., and J. A. Cypher (2010), Postaudit evaluation of conceptual model uncertainty for a glacial aquifer groundwater flow and contaminant transport model, *Hydrogeol. J.*, *18*, 945–958.

Lu, D., M. Ye, and S. P. Neuman (2011), Dependence of Bayesian model selection criteria and Fisher information matrix on sample size, *Math. Geosci.*, *43*(8), 971–993, doi:10.1007/s11004-011-9359-0.

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

Lu, D., M. Ye, S. P. Neuman, and L. Xue (2012b), Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs, *Adv. Water Resour.*, *35*, 69–82, doi:10.1016/j.advwatres.2011.10.007.

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

Lu, D., M. Ye, M. C. Hill, E. P. Poeter, and G. P. Curtis (2014), A computer program for uncertainty analysis integrating regression and Bayesian methods, *Environ. Modell. Software*, *60*, 45–56, doi:10.1016/j.envsoft.2014.06002.

Marshall, L., D. Nott, and A. Sharma (2005), Hydrological model selection: A Bayesian alternative, *Water Resour. Res.*, *41*, W10422, doi:10.1029/2004WR003719.

Morey, R. D., J. N. Rouder, M. S. Pratte, and P. L. Speckman (2011), Using MCMC chain outputs to efficiently estimate Bayesian factors, *J. Math. Psychol.*, *55*, 368–378.

Neal, R. M. (2000), Markov chain sampling methods for Dirichlet process mixture models, *J. Comput. Graphical Stat.*, *9*, 249–265.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, *Stochastic Environ. Res. Risk Assess.*, *17*, 291–305, doi:10.1007/s00477-003-0151-7.

Neuman, S. P., L. Xue, M. Ye, and D. Lu (2012), Bayesian analysis of data-worth considering model and parameter uncertainties, *Adv. Water Resour.*, *36*, 75–85, doi:10.1016/j.advwatres.2011.02.007.

Newton, M. A., and A. E. Raftery (1994), Approximate Bayesian inference with the weighted likelihood bootstrap, *J. R. Stat. Soc., Ser. B*, *56*(1), 3–48.

Raftery, A. E., M. A. Newton, J. M. Satagopan, and P. N. Krivitsky (2007), Estimating the marginal likelihood via posterior simulation using the harmonic mean identidy (with discussion), in *Bauesian Statistics*, vol. 8, edited by J. M. Bernardo et al., pp. 1–45, Oxford Univ. Press, Oxford, U. K.

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

Rojas, R., L. Feyen, and A. Dassargues (2008), Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resour. Res.*, *44*, W12418, doi:10.1029/2008WR006908.

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

Schöniger, A., T. Wohling, L. Samaniego, and W. Nowak (2014), Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence, *Water Resour. Res.*, *50*, 9484–9513, doi:10.1002/2014WR016062.

Schoups, G., and J. A. Vrugt (2011), Bayesian selection of hydrological models using sequential Monte Carlo sampling, Abstract H23D-1310 presented at 2011 Fall Meeting, AGU, San Francisco, Calif. [Available at http://faculty.sites.uci.edu/jasper/files/2012/10/poster_AGU2011.pdf.]

Schoups, G., N. C. van de Giesen, and H. H. G. Savenije (2008), Model complexity control for hydrologic prediction, *Water Resour. Res.*, *44*, W00B03, doi:10.1029/2008WR006836.

Shi, X., M. Ye, G. P. Curtis, G. L. Miller, P. D. Meyer, M. Kohler, S. Yabusaki, and J. Wu (2014), Assessment of parametric uncertainty for groundwater reactive transport modeling, *Water Resour. Res.*, *50*, 4416–4439, doi:10.1002/2013WR013755.

Skilling, J. (2004), Nested sampling, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conference Proceedings*, vol. 735, edited by R. Fischer, R. Preuss, and U. von Toussaint, American Institute of Physics, pp. 395–405, AIP Conference Proceedings, Melville, N. Y., doi:10.1063/1.1835238.

Skilling, J. (2006), Nested sampling for general Bayesian computation, *Bayesian Anal.*, *1*(4), 833–859, doi:10.1214/06-BA127.

Tsai, F. T.-C., and A. S. Elshall (2013), Hierarchical Bayesian model averaging for hydrostratigraphic modeling: Uncertainty segregation and comparative evaluation, *Water Resour. Res.*, *49*, 5520–5536, doi:10.1002/wrcr.20428.

Von Toussaint, U. (2011), Bayesian inference in physics, *Rev. Mod. Phys.*, *83*(3), 943–999.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon (2009), Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*(3), 271–288.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, and G. Schoups (2013), Hydrologic data assimilation using particle filter Markov chain Monte Carlo simulation: Theory, concepts and applications, *Adv. Water Resour.*, *51*, 457–478.

Weinberg, M. D. (2012), Computing the Bayes factor from a Markov Chain Monte Carlo simulation of the posterior distribution, *Bayesian Anal.*, *7*(3), 737–769.

Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen (2011), Improving marginal likelihood estimaton for Bayesian phylogenetic model selection, *Syst. Biol.*, *60*(2), 150–160.

Xue, L., and D. Zhang (2014), A multimodel data assimilation framework via the ensemble Kalman filter, *Water Resour. Res.*, *50*, 4197–4219, doi:10.1002/2013WR014525.

Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*, W05113, doi:10.1029/2003WR002557.

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

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

Ye, M., K. F. Pohlmann, J. B. Chapman, G. M. Pohll, and D. M. Reeves (2010b), A model-averaging method for assessing groundwater conceptual model uncertainty, *Ground Water*, *48*, 716–728, doi:10.1111/j.1745-6584.2009.00633.x.

Zhang, G., D. Lu, M. Ye, M. Gunzburger, and C. Webster (2013), An adaptive sparse-grid high-order stochastic collocation method for Bayesian inference in groundwater reactive transport modeling, *Water Resour. Res.*, *49*, 6871–6892, doi:10.1002/wrcr.20467.

Zhang, X., G.-Y. Niu, A. S. Elshall, M. Ye, G. A. Barron-Gafford, and M. Pavao-Zuckerman (2014), Assessing five evolving microbial enzyme models against field measurements from a semiarid savannah: What are the mechanisms of soil respiration pulses?, *Geophys. Res. Lett.*, *41*, 6428–6434, doi:10.1002/2014GL061399.