Stoch Environ Res Risk Assess (2010) 24:1003–1022
DOI 10.1007/s00477-010-0395-y

ORIGINAL PAPER

Parameter estimation in nonlinear environmental problems

Xiaoyi Liu · Michael A. Cardiff · Peter K. Kitanidis

Published online: 12 May 2010

Abstract Popular parameter estimation methods, including least squares, maximum likelihood, and maximum a posteriori (MAP), solve an optimization problem to obtain a central value (or best estimate) followed by an approximate evaluation of the spread (or covariance matrix). A different approach is the Monte Carlo (MC) method, and particularly Markov chain Monte Carlo (MCMC) methods, which allow sampling from the posterior distribution of the parameters. Though available for years, MC methods have only recently drawn wide attention as practical ways for solving challenging high-dimensional parameter estimation problems. They have a broader scope of applications than conventional methods and can be used to derive the full posterior pdf but can be computationally very intensive. This paper compares a number of different methods and presents improvements using as case study a nonlinear DNAPL source dissolution and solute transport model. This depth-integrated semi-analytical model approximates dissolution from the DNAPL source zone using nonlinear empirical equations with partially known parameters. It then calculates the DNAPL plume concentration in the aquifer by solving the advection-dispersion equation with a flux boundary. The comparison is among the classical MAP and some versions of computer-intensive Monte Carlo methods, including the Metropolis–Hastings (MH) method and the adaptive direction sampling (ADS) method.

X. Liu (✉) · M. A. Cardiff · P. K. Kitanidis
Department of Civil and Environmental Engineering,
Stanford University, Stanford, CA 94305, USA
e-mail: shawnliu@stanford.edu

Keywords Markov chain Monte Carlo · Bayesian method · Maximum a posteriori estimation · MCMC convergence

1 Introduction

Parameter estimation is a fundamental problem for researchers and practitioners who work with mathematical models in almost every field of endeavor. Every model has parameters that must be selected, and this problem is even more important when the model describes subsurface processes, where direct measurements are expensive and sometimes even impossible to obtain (Frind and Pinder 1973; Kitanidis and Vomvoris 1983; Butler et al. 1999; Yeh and Liu 2000; Liu et al. 2007). Parameters must be inferred from data, some of which may only be indirectly related to the parameters of interest.

Generally we can conceptualize the physical model in the form of functions as

$$\mathbf{y} = \mathbf{h}(\boldsymbol{\theta}), \tag{1}$$

where $\mathbf{y}$ is a vector of quantities that can be predicted from the model (output of the model), $\boldsymbol{\theta}$ is the vector of unknown parameters, $\mathbf{h}$ is a set of functions that map the parameter space to the output space. If quantities $\mathbf{y}$ are also measured, Eq. 1 can be used to estimate the values of $\boldsymbol{\theta}$. The representation of Eq. 1 is incomplete because, in practice, it is usually unreasonable to expect that model predictions should perfectly match observations even if the right parameters could be found. The measurement process causes error, i.e., the measured value is not exactly equal to the true value. Furthermore, the conceptualization process itself regularly leads to an inexact representation (or model error) of the physical processes, which is another reason why the model should not be expected to exactly match the observations. For practical purposes, it is common practice to incorporate all the uncertainty from the model and the measurement into a term ϵ that describes the total deviation of the measurement from model predictions given the "ideal" parameter set $\boldsymbol{\theta}$. Hence a more useful representation that we will be working with is

$$\mathbf{y} = \mathbf{h}(\boldsymbol{\theta}) + \epsilon. \tag{2}$$

A relevant issue is nonlinearity. A parameter estimation problem is called nonlinear when the transformation from the parameter set to the observations is nonlinear, i.e., when $\mathbf{h}$ is nonlinear. The topic of nonlinear estimation is important because most physical processes are nonlinear. Parameter estimation for such models, particularly the probabilistic quantification of uncertainty, is mathematically and computationally difficult to tackle in an exact way. The most common approach (Bard 1973; van den Bos 2007) in applications involves approximations of a best estimate and estimation error, usually through a series of linearization steps. Examples are the methods of nonlinear least squares (NLS), maximum likelihood (ML), and maximum a posteriori (MAP) estimation. In these methods, the best estimate is obtained by minimizing a "fitting criterion" followed by a linearized uncertainty analysis (Bard 1973). The criterion can be fitting to the data (NLS); a probability model of the data (ML); or the posterior distribution of the parameters (MAP). We will refer to such methods as classical methods. These methods have good asymptotic properties, i.e., when the number of observations tends to infinity, parameter estimates are unbiased and exhibit minimum variance with Gaussian distributions. However, in real-world cases, data are sparse. Although it is hoped that, in many practical cases, these methods should work quite adequately and give results that are reasonably close to the "correct solution", there has been a dearth of studies that verify this expectation.

Uncertainty analysis based on such classical estimation methods is valid when a number of implicit assumptions are met. They perform best when the distribution of errors is nearly symmetric and not very different from Gaussian. Such behavior is met when the confidence interval is sufficiently small so that the function can be approximated by a linear function over that range. Otherwise, a single best estimate, which may not even be close to the mean of the distribution, and an approximately evaluated covariance matrix are not adequate to represent the probability distribution of errors and thus may be unacceptable for use in probabilistic (i.e., risk based) assessment of management plans or strategies. Optimization under uncertainty may require generating a number of equi-probable sets of parameters, which are representative of the uncertainty in the parameters.

When classical methods are inadequate, we must resort to methods that are not dependent on the linearity of the model, among which Monte Carlo methods are most prominent. For example, Sahuquillo et al. (1992) developed the sequential self-calibration method to simulate the transmissivity field conditioned on piezometric data and later Gomez-Hernandez et al. (1997) provided theoretical basis for it. Gomez-Hernandez et al. (2001) applied this method in conductivity simulation of a fractured rock block and Franssen et al. (2003) extended it to a coupled groundwater flow and mass transport problem. Ramarao et al. (1995) developed the pilot point method for conditional simulation of transmissivity field and it was later applied in a fractured aquifer by Lavenue and de Marsily (2001). Kentel and Aral (2005). combined fuzzy set theory with Monte Carlo methods to include incomplete information and applied it in health risk analysis. A recent review of the application of Monte Carlo methods in the inverse modeling of groundwater flow can be found in Franssen et al. (2009).

An early Monte Carlo application can be traced back to the famous needle-throwing experiment (Buffon's Needle) conducted by Georges Buffon in 1777 (Dorrie 1965). Because the Monte Carlo method converges at a rather slow rate (proportional to the square root of the number of samples, according to the Central Limit Theorem), its potential use is highly dependent on an efficient sampling strategy to produce samples $\boldsymbol{\theta}$ that follow a certain distribution, such as the posterior distribution, which may be hard to work with and defined only within a multiplicative constant. The rejection method (von Neumann 1951) can be used for this purpose. However, application of this method is hindered by difficulties in finding the constant term in the acceptance ratio and in finding a proper proxy distribution that is not too different from the target distribution and that is easy to sample from. Almost two hundred years after the Buffon's Needle experiment, the Metropolis–Hastings algorithm was invented and generalized. This is a powerful and comprehensive approach that can be used to create a Markov chain that is composed of samples from any distribution. The theory underlying the Markov chain Monte Carlo (MCMC) method has been rigorously investigated and many algorithms have been developed to construct the Markov chain, including the Gibbs sampler (Geman and Geman 1984), which does directional sampling along the coordinates, and the hybrid MCMC sampler (Duane et al. 1987), which uses a series of deterministic iteration steps (or a surrogate distribution) to generate samples.

The nonlinear parameter estimation problems can be formulated in a Bayesian framework leading to an expression for the posterior probability density function (pdf) (Bard 1973; van den Bos 2007), which is known up to a multiplicative constant and is difficult to sample directly. We are primarily interested in the application of the MCMC method to generate a modest number of independent and equiprobable samples from the posterior probability density function in Bayesian inference. These samples are to be used in risk analysis and optimization under uncertainty.

There have been numerous applications of the MCMC method in science and engineering, but we will focus on reviewing applications in the water resources field. Kuczera and Parent (1998), Bates and Campbell (2001), Vrugt et al. (2003), Feyen et al. (2007), and Blasone et al. (2008) used the MCMC method to evaluate parameter uncertainty in hydrological models. Marshall et al. (2004) and Smith and Marshall (2008) performed comparison studies of the MCMC application in rainfall-runoff modeling. The MCMC method was also used to generate conditional realizations in parameter/function estimation (Michalak and Kitanidis 2003). Oliver et al. (1997), Michalak (2008) and Fu and Gomez-Hernandez (2009) applied MCMC methods in various groundwater applications. Vrugt et al. (2008) designed the differential evolution adaptive Metropolis (DREAM) algorithm especially for efficient sampling of the posterior distribution of hydrological models, and Vrugt et al. (2009) later compared this algorithm with the generalized likelihood uncertainty estimation (GLUE) method. The MCMC method thus is increasingly recognized as a promising approach to solving challenging parameter estimation problems in water resources areas.

Classical parameter estimation methods are "local" in nature because they involve optimization algorithms to determine a single estimate (e.g. the peak of the posterior distribution) followed by an approximate procedure to compute a covariance matrix that describes parameter estimation error. In contrast, the MCMC approach is global in the sense that it attempts to represent the probability distribution through a large number of samples. The MCMC method per se is theoretically sound and straightforward to program. However, the method may require many iterations and the generation and evaluation of many samples. Each sample requires at least the implementation of one run of the model and thus the total computational cost can be high.

As an important member in the family of MCMC samplers, Gibbs sampling has the advantage of eliminating the worry of acceptance rate in MCMC sampling because it searches in the axial directions and accepts all proposal samples acquired from axial line sampling. As Gilks et al. (1994) pointed out, one major drawback of the Gibbs sampler is that when the support domain of the target probability distribution comprises several disjoint subdomains, such as that shown in Fig. 1, it is impossible for the Gibbs sampler to sample the whole domain. Hence Gilks et al. (1994) proposed an adaptive direction sampling method in which the Snooker algorithm starts from multiple initial samples that compose the initial population. Then two samples are uniformly chosen without replacement from the population and a new sample is generated from an adjusted conditional distribution along the line determined by the two chosen samples. This line sampling strategy overcomes the drawback of Gibbs line sampling. Furthermore, as Liu et al. (2000) propose, this sampling strategy can be easily combined with a local optimization algorithm such as the Conjugate Gradient method. While Gilks et al. (1994) did not mention exactly how to sample from the adjusted distribution, Liu et al. (2000) developed a multiple-try method (MTM) for this purpose and named the new sampler the Conjugate Gradient Monte Carlo (CGMC) sampler. However, in a high-dimensional problem, the local mode search step is rather computationally intensive, and our test cases showed that it might not be worthwhile to perform a local mode search in high-dimensional applications. Hence in this paper, we will use the ADS sampling strategy combined with the multiple-try line sampling method.

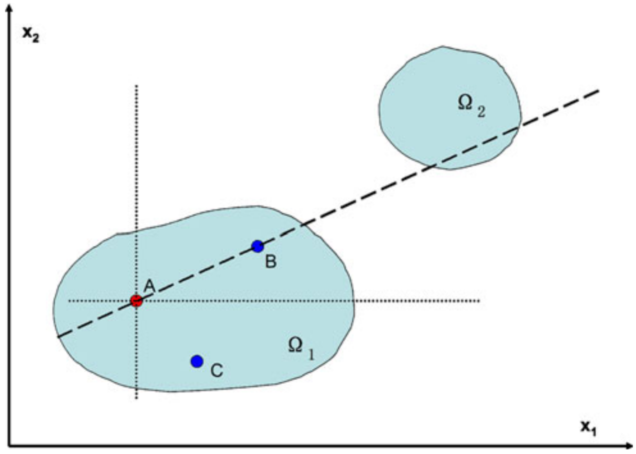


Fig. 1 The failure of Gibbs sampling and the success of ADS sampling. The *dotted lines* are the two possible Gibbs sampline directions, and the *dashed line* represents one of the possible sampling directions when there are three sequences

Since MCMC methods are computationally expensive, the quality of the samples is of paramount importance in MCMC sampling. However, this issue has not received enough attention in many applications of MCMC methods in the water resources area. We recommend a combination of two diagnostic methods in this paper. The first method comes from the perspective of the independence requirement of the samples for Monte Carlo simulation. Thus, we test the auto-correlation of samples from one MCMC chain as a function of lag distance, expecting that the auto-correlation coefficient stabilize around 0 after a relatively short lag distance. The second method diagnoses the convergence of the MCMC chains, using the Scale Reduction Factor (SRF) proposed by Gelman and Rubin (1992). The SRF compares the cross-chain variance and the within-chain variance, and it serves as an effective measure of the convergence of MCMC sampling. In this work, we also adopt the graphical approach suggested by Brooks and Gelman (1998) to visualize the SRF as a function of the number of samples.

In this paper, we generalize on several aspects in the application of the MCMC method to parameter estimation problems. In the process, we review all the necessary steps required in MCMC sampling, i.e. choice of the starting sample(s); choice of the candidate generating function, and the diagnosis of the samples. We also test a relatively new MCMC sampler, the ADS sampler in conjunction with a semi-analytical DNAPL dissolution and transport model. This methodology is applicable to the parameter estimation in environmental problems and associated risk analysis. We also compare a classical parameter estimation method (MAP) with the MCMC methods to test the applicability of the classical methods in environmental problems.

2 Bayesian probability model

In Bayesian theory, the posterior distribution of parameters $\boldsymbol{\theta}, \pi(\boldsymbol{\theta}|\mathbf{y})$, (i.e., the distribution conditional on the observations $\mathbf{y}$) is proportional to two terms: the prior distribution, which is the unconditional distribution of $\boldsymbol{\theta}, \pi_0(\boldsymbol{\theta})$, and the conditional distribution of the observations $\mathbf{y}$ given the parameters $\boldsymbol{\theta}, L(\mathbf{y}|\boldsymbol{\theta})$, which is also called the likelihood function. Then, using Bayes' theorem

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{L(\mathbf{y}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y})}, \tag{3}$$

where $f_{\mathbf{Y}}(\mathbf{y})$ is the unconditional distribution of the observations $\mathbf{y}$,

$$\pi(\boldsymbol{\theta}) \equiv \pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi_0(\boldsymbol{\theta}) \times L(\mathbf{y}|\boldsymbol{\theta}) \tag{4}$$

where $\pi(\boldsymbol{\theta})$ is the target distribution of $\boldsymbol{\theta}$ from which we want to generate samples. We drop the conditional notation for notational simplification, keeping in mind that $\pi(\boldsymbol{\theta})$ is the posterior probability density function of $\boldsymbol{\theta}$. When parameters $\boldsymbol{\theta}$ can only possibly have values in a certain domain, we call it the support domain and denote it as $\boldsymbol{\Theta}$.

Although many parameter estimation applications utilize only the likelihood, which means the exclusion of any prior/subjective information on the distribution of $\boldsymbol{\theta}$, we strongly recommend this term be included in the target distribution for the following reasons. (1) In all physical models, we always know something about the parameters we want to estimate even before we collect measurements. The Bayesian formulation and the prior distribution offer a convenient and systematic way to introduce this information. (2) Most physical parameters are meaningful only within certain bounds. For instance, contaminant mass can not be negative; and (3) Required relationships/constraints among various parameters as well as the information from earlier observations can be included in the prior distribution. Furthermore, without prior information, the unidentifiability problem often arises due to the infinitely large support domain of the parameters.

The next issue is what kind of prior distribution we should use. Our criteria for the prior distribution are as follows.

1. When there is information available indicating a parameter should follow a certain distribution, use that distribution for the parameter.
2. When there is no information about the distribution of the parameter but a typical value and its variance can be postulated, use normal distribution for the parameter. If the parameter can only have one sign, use a log-normal distribution for the parameter and formulate the problem in terms of the logarithm of the parameter instead.
3. When there is no information about the distribution but a physical range of the parameter is known, use a uniform distribution defined on the range. In this case, the prior information known about the parameter is rather limited. When a uniform distribution defined on $[-\infty, \infty]$ is used as a prior, which means no prior information at all, the posterior distribution is exactly the likelihood function.

Another assumption we often need to make is the form of the conditional distribution of the observation given all the parameters, i.e., the likelihood function $L(\mathbf{y}|\mathbf{x})$. With a model as in Eq. 2, we know that when $\mathbf{h}$ is linear and $\boldsymbol{\theta}$ is normal, $\mathbf{y}$ will also be normal. However, when $\mathbf{h}$ is nonlinear, the conditional distribution of $\mathbf{y}|\mathbf{x}$ can not be directly derived and can not be written in a closed form. Hence, according to the same rules mentioned above for the prior distribution, we assume that $\mathbf{y}|\mathbf{x}$ is normal and $L(\mathbf{y}|\mathbf{x})$ follows the pdf of a multivariate normal distribution with mean $\mathbf{h}(\boldsymbol{\theta})$ and covariance matrix $\mathbf{R}$ that is the same as that of the observation error ϵ.

In case we use the multivariate normal distribution as the prior distribution, we can rewrite the target pdf, Eq. 4 as

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}})^T \mathbf{Q}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}})\right) \times \exp\left(-\frac{1}{2}(\bar{\mathbf{y}}-\mathbf{h}(\boldsymbol{\theta}))^T \mathbf{R}^{-1}(\bar{\mathbf{y}}-\mathbf{h}(\boldsymbol{\theta}))\right) \tag{5}$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\mathbf{Q}$ are respectively the mean, and covariance matrix of the prior multivariate normal distribution; and $\bar{\mathbf{y}}$ and $\mathbf{R}$ are respectively the measurement data and covariance matrix of $\mathbf{y}$.

3 MAP parameter estimation with linearized uncertainty analysis

One of the most popular methods for parameter estimation is the maximum a posteriori (MAP) approach. The MAP method is related to the Maximum Likelihood (ML) method (for example, Kitanidis and Lane 1985; Carrera and Neuman 1986) in the sense that the target function of the ML method is a part of the target function of the MAP method. Nonetheless, the ideas behind these two methods are somewhat different, the former being a sampling-theory tool and the latter being a Bayesian approach. The logic behind the ML method is, given the conditional distribution of the observations and a sample (the observed values) from this distribution, under what kind of conditions (parameters) can the probability of the occurrence of the sample be maximized. While the logic behind the MAP method is, given the conditional distribution of the parameters, what is the mode of this distribution, or what are the most probable values for the parameters, or what is the best estimate of the parameter given all the information we have. Thus, for the MAP method, we try to maximize the posteriori probability density function $\pi(\boldsymbol{\theta})$ as in Eq. 5 by solving the following constrained optimization problem:

$$\max_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}) = \exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}}\right)^T\mathbf{Q}^{-1}\left(\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}}\right)\right) \times \exp\left(-\frac{1}{2}(\bar{\mathbf{y}}-\mathbf{h}(\boldsymbol{\theta}))^T\mathbf{R}^{-1}(\bar{\mathbf{y}}-\mathbf{h}(\boldsymbol{\theta}))\right) \tag{6}$$

or equivalently,

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \left(\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}}\right)^T\mathbf{Q}^{-1}\left(\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}}\right) + (\bar{\mathbf{y}}-\mathbf{h}(\boldsymbol{\theta}))^T\mathbf{R}^{-1}(\bar{\mathbf{y}}-\mathbf{h}(\boldsymbol{\theta})) \tag{7}$$

subject to

$\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Given nonlinearity in $\mathbf{h}$, the covariance matrix of the distribution in Eq. 5 can rarely be easily derived. Here we use a first order approximation of functions $\mathbf{h}$ and the posterior covariance matrix can be written as

$$\widehat{\mathbf{V}} = \left(\widehat{\mathbf{H}}^T\mathbf{R}^{-1}\widehat{\mathbf{H}} + \mathbf{Q}^{-1}\right)^{-1} \tag{8}$$

where $\widehat{\mathbf{V}}$ is the MAP covariance matrix estimate, and $\widehat{\mathbf{H}}$ is the sensitivity matrix evaluated at the solution of Eq. 7 ($\hat{\boldsymbol{\theta}}$): $\widehat{\mathbf{H}}_{i,j} = \frac{\partial h_i}{\partial \theta_j}|_{\hat{\boldsymbol{\theta}}}$ where h_i is the ith component in $\mathbf{h}$ and θ_j is the jth component in $\boldsymbol{\theta}$.

One would notice that $\widehat{\mathbf{V}}$ is actually the Fisher information matrix for the MAP estimator, hence it can be used to bound the covariance matrix of the estimator according to the Cramér-Rao Inequality. From this point of view, we would expect that $\widehat{\mathbf{V}}$ is an underestimate of the posterior covariance matrix $\mathbf{V}$ in the sense that the difference $\mathbf{V}-\widehat{\mathbf{V}}$ is positive semidefinite.

4 MCMC sampling

4.1 Metropolis–Hastings (MH) sampling

The Metropolis–Hastings (MH) sampling strategy (Metropolis et al. 1953; Hastings 1970) is the most popular and the most investigated MCMC method. It has also been used and proved to be effective in several parameter estimation applications in hydrology (Kuczera and Parent 1998; Bates and Campbell 2001; Vrugt et al. 2003; Feyen et al. 2007; Blasone et al. 2008). The MH sampling strategy starts from an initial sample and then evolves according to the following steps:

(1) Generate a new sample $\boldsymbol{\theta}_{k+1} \in \boldsymbol{\Theta}$ from a candidate generating density function, $p(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1})$, where $\boldsymbol{\theta}_k$ is the current sample, and $p(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1})$ should be from a distribution that can be directly sampled.

(2) Calculate the ratio

$$r = \frac{\pi(\boldsymbol{\theta}_{k+1})p(\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_k)}{\pi(\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1})} \tag{9}$$

and accept the new sample with probability $\alpha = \min\{r, 1\}$.

If we choose a symmetric candidate generating density function $p(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1})$ such that $p(\boldsymbol{\theta}, \boldsymbol{\theta}_{k+1}) = p(\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_k)$, then the ratio in Eq. 9 can be simplified to $r = \frac{\pi(\boldsymbol{\theta}_{k+1})}{\pi(\boldsymbol{\theta}_k)}$ and $\alpha = \min\{\frac{\pi(\boldsymbol{\theta}_{k+1})}{\pi(\boldsymbol{\theta}_k)}, 1\}$.

The Metropolis–Hastings algorithm is quite straightforward and easy to program. One modification to this algorithm would be to run multiple parallel chains starting from different initial samples. This modification needs little change to the original algorithm and is especially useful when one can take advantage of a multi-processor computer.

4.2 Adaptive direction sampling (ADS)

For the multiple-chain Metropolis–Hastings algorithm, each chain starts from its own initial sample and goes through a different path of samples to converge to the same target distribution, $\pi(\boldsymbol{\theta})$. During this process, each chain does not gain any information from the other chains, hence it does not know where the other chains are and what paths the other chains have traveled through. Communication among chains could improve efficiency. For example, when a chain gets trapped at a minor local mode of the target distribution, information from other chains could allow it to break away. On the other hand, the communication among the chains has to be carefully designed to retain the desired properties of the Markov chain. Gilks et al. (1994) propose a method called adaptive direction sampling (ADS) for this purpose, and in this paper, we will avoid the theories but review the algorithm below.

Adaptive Direction Sampling (ADS) generates several chains of samples in parallel. Hence, instead of one initial sample, it starts from a population of m samples and all samples in the population evolve from one generation to the next. To generate the next generation, a current point $(\boldsymbol{\theta}_k^{(c)})$—the point to be moved—is chosen randomly from the current generation, and an anchor point $(\boldsymbol{\theta}_k^{(a)})$ is chosen independently and uniformly from the rest of the current generation k (Fig. 1). Then, the current point is updated by a new point sampled along the random direction determined by the two chosen points, i.e., $\mathbf{e}_k = (\boldsymbol{\theta}_k^{(c)} - \boldsymbol{\theta}_k^{(a)})/\left\|\boldsymbol{\theta}_k^{(c)} - \boldsymbol{\theta}_k^{(a)}\right\|$. The location of the new point is determined by a scalar l drawn from distribution

$$f(r) \propto |l|^{d-1}\pi(\boldsymbol{\theta}_k^{(a)} + l\mathbf{e}_k) \tag{10}$$

where d is the dimension of $\boldsymbol{\theta}$ and π is the target distribution. After that, the current population is updated by $\boldsymbol{\theta}_{k+1}^{(c)} = \boldsymbol{\theta}_k^{(a)} + l_k\mathbf{e}_k$, and $\boldsymbol{\theta}_{k+1}^{(j)} = \boldsymbol{\theta}_k^{(j)}$ for $j \neq c$. In the distribution in Eq. 10, the target distribution is adjusted in a way such that a penalty is enforced on the points around the anchor point. The adjustment is a must to ensure that the stationary distribution of the chain is such that each sample from the current population is independently drawn from distribution $\pi(\boldsymbol{\theta})$, or

$$P(\boldsymbol{\theta}_k^{(1)}, \boldsymbol{\theta}_k^{(2)}, \ldots, \boldsymbol{\theta}_k^{(m)}) = \prod_{i=1}^{m} \pi(\boldsymbol{\theta}_k^{(i)}) \tag{11}$$

The line sampling along the random direction can be conducted in various ways, for example, the Griddy-Gibbs (Ritter and Tanner 1992) sampling that approximates the 1-D distribution with numerical integration. Here in this paper, we use the multiple-try Metropolis (MTM) method developed by Liu et al. (2000) because it does not need the numerical integration that is usually computationally intensive. It proceeds as follows.

First define

$$w(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta})T(\boldsymbol{\theta}, \boldsymbol{\theta}')\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}') \tag{12}$$

where $\pi(\boldsymbol{\theta})$ is the target pdf specified up to a multiplicative constant, $T(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is the candidate generating density function (corresponding to the function $p(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in Eq. 9), and $\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is a nonnegative symmetric function in $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. In this equation, the first two parts are commonly seen as in the Metropolis–Hastings algorithm, and the third part, as we will see later, adds a lot of flexibility to this algorithm. The MTM algorithm proceeds as follows.

1. At current state $\boldsymbol{\theta}$, draw t ($t = 8$, for example) iid candidates $\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, \ldots \boldsymbol{\theta}'_t$, from $T(\boldsymbol{\theta}, \cdot)$, compute $w(\boldsymbol{\theta}, \boldsymbol{\theta}'_i)$ for $i = 1, 2, \ldots t$;
2. select $\boldsymbol{\theta}^*$ among the candidates $\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, \ldots \boldsymbol{\theta}'_t$ with probability proportional to $w(\boldsymbol{\theta}, \boldsymbol{\theta}'_i)$;
3. draw $t - 1$ iid candidates $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_{t-1}$ from $T(\boldsymbol{\theta}^*, \cdot)$, and let $\boldsymbol{\theta}_t = \boldsymbol{\theta}$;
4. accept $\boldsymbol{\theta}^*$ with probability

$$r_g = \min\left\{1, \frac{\sum_{i=1}^{t} w(\boldsymbol{\theta}'_i, \mathbf{x})}{\sum_{j=1}^{t} w(\boldsymbol{\theta}_j, \boldsymbol{\theta}^*)}\right\} \tag{13}$$

where r_g is called the generalized M–H ratio.

The choice of $\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is flexible. Liu et al. (2000) propose the so-called MTM(II) algorithm in which

$$\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}') = \left(\frac{T(\boldsymbol{\theta}, \boldsymbol{\theta}') + T(\boldsymbol{\theta}', \boldsymbol{\theta})}{2}\right)^{-1} \tag{14}$$

where $T(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is a symmetric candidate generating density function, such that

$$r_g = \min\left\{1, \frac{\sum_{i=1}^{t} \pi(\boldsymbol{\theta}'_i)}{\sum_{j=1}^{t} \pi(\boldsymbol{\theta}_j)}\right\}. \tag{15}$$

However, this formation is numerically problematic when $\pi(\boldsymbol{\theta})$ is calculated as its logarithm (or negative logarithm) to avoid computational overflow when $\pi(\boldsymbol{\theta})$ is too big. Also loss of accuracy may occur when $\pi(\boldsymbol{\theta})$ is too small. In this paper a new $\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is proposed as

$$\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}') = (\pi(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta}'))^{-1}T(\boldsymbol{\theta}, \boldsymbol{\theta}')^{-1}. \tag{16}$$

It satisfies the requirements for $\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}')$: (1) it is symmetric when we choose a symmetric $T(\boldsymbol{\theta}, \boldsymbol{\theta}')$; and (2) $\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}') > 0$ when $T(\boldsymbol{\theta}, \boldsymbol{\theta}') > 0$.

It leads to

$$\begin{aligned} w(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \pi(\boldsymbol{\theta})T(\boldsymbol{\theta}, \boldsymbol{\theta}')\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}') \\ &= \frac{\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta}')} \\ &= \left[1 + \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})}\right]^{-1} \\ &= [1 + \exp(\ln(\pi(\boldsymbol{\theta}')) - \ln(\pi(\boldsymbol{\theta})))]^{-1} \end{aligned} \tag{17}$$

and the generalized M–H ratio can be calculated using Eq. 13.

The choice for $\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is robust when the absolute value of $\pi(\boldsymbol{\theta})$ is small relative to the computational precision, while the difference between $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}')$ is relatively larger.

4.3 Starting samples

There is no doubt that the choice of the starting samples affects the performance of MCMC sampling. For example, having a starting sample at the tail of the density function means it will take a long time to reach the mode area of the density function and many of the samples at the beginning of the chain will have to be discarded (the burn-in period). The most intuitive way of choosing a starting sample is to use the prior mean values, or a local mode search can be conducted with the prior estimates to get a starting sample. However, this only gives one starting sample while multiple starting samples are needed in parallel MH sampling and ADS sampling. What is more, when multiple starting samples are used, we require them to be relatively scattered. An ideal choice would be that the starting samples are from the target distribution. Unfortunately, the later is unknown, and a straightforward choice would be to sample from a multivariate normal distribution with mean and covariance from the MAP estimates. (MCMC sampling are much more computationally intensive than the MAP method and it is supposed to provide more accurate estimation of the parameters, so it makes sense to use the results from MAP method as a starting point.) However, samples generated in such a way are often not adequately dispersed, especially when the target distribution has more than one modes and minor local modes are not insignificant compared with the major mode. As Cowles and Carlin (1996) have shown, serious convergence test errors can be made when the starting samples may not be adequately dispersed. Here we will review a process suggested by Gelman and Rubin (1992) as follows.

1. Conduct a set of local mode searches w.r.t. Eq. 7 from various initial points, and evaluate the Hessian matrices at the local modes.
2. Generate u samples from a mixture of normal approximations of the target distribution with the local modes as the mean values and the inverse Hessian matrices as covariance matrices. The weight of each normal distribution in the mixture is proportional to the probability density evaluations at the local modes, hence minor local modes with density evaluations significantly smaller than that of the major mode can be ignored. If there is only one mode (i.e., the MAP estimate), sample from the MAP normal approximation.
3. Divide the samples by a scalar random variable $\sqrt{\chi^2_\eta/\eta}$, where ${\chi^2}_\eta$ is a Chi-Square variate with degrees of freedom η (say $\eta = 4$, as suggested by Gelman and Rubin (1992)), hence resulting in n samples that are from a mixture of Student's t-distribution and probability density function $\tilde{\pi}_t(\boldsymbol{\theta})$.
4. Draw m starting samples from the u samples without replacement and with probability proportional to the importance ratio $\frac{\pi(\boldsymbol{\theta})}{\tilde{\pi}_t(\boldsymbol{\theta})}$.

4.4 Candidate generating density functions

The candidate generating density function affects the performance of MCMC sampling in several ways. First, as in importance sampling, the acceptance rate is affected by the candidate generating density function. Having a candidate generating density function that is significantly different from the target density function can significantly decrease the acceptance rate. Second, the candidate generating density function defines the relationship between two consecutive samples in the chain, hence the autocorrelation function (ACF) of the Markov chain. Third, the candidate generating density function defines the size of the region where the next sample will be generated. Consequently, it affects the rate at which the Markov chain traverses the support domain.

There are two major groups of candidate generating density functions. The first group uses random walks, meaning the candidate is a random increment added to the current sample. The other group avoids random walk, meaning that the candidates are either calculated from a deterministic function (Hybrid MC; Duane et al. 1987), or iid samples from a proxy distribution as in importance sampling.

We can use the prior distribution to get sample candidates (either random walk or iid); however, when the number of observations is large or the observation error is small, the likelihood function will have much larger weight in the posterior density function than the prior distribution does, thus the posterior distribution will be generally significantly different from the prior distribution. When that is the case, the acceptance rate will be very low. Another choice is to use the covariance matrix $\widehat{\mathbf{V}}$ from the MAP estimation (Eq. 8), i.e., we do a Cholesky factorization w.r.t. to $\widehat{\mathbf{V}}$ to get an upper triangular matrix $\mathbf{R}$ such that $\mathbf{R}^T\mathbf{R} = \widehat{\mathbf{V}}$. In case the $\widehat{\mathbf{V}}$ is not positive definite, a modified Cholesky factorization by adding positive values to the diagonal components of $\widehat{\mathbf{V}}$ (Gill and Murray 1981) can be used. Then a sample candidate $\boldsymbol{\theta}_{k+1}$ is generated through $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mathbf{R}^T\mathbf{e}$ (random walk) or $\boldsymbol{\theta}_{k+1} = \hat{\boldsymbol{\theta}} + \mathbf{R}^T\mathbf{e}$ (iid), where $\hat{\boldsymbol{\theta}}$ is the MAP best estimates of the parameters, and $\mathbf{e}$ is a vector of iid standard normal random numbers.

To generalize, we propose here an autoregressive equation to generate candidate samples as

$$\boldsymbol{\theta}_{k+1} = \rho\boldsymbol{\theta}_k + (1-\rho)\hat{\boldsymbol{\theta}} + \mu\boldsymbol{\zeta}_{k+1} + \lambda\boldsymbol{\delta}_{k+1} \tag{18}$$

where $\boldsymbol{\theta}_k$ is the current sample, $\boldsymbol{\zeta}_{k+1}$ is a multivariate normal vector with zero mean and covariance matrix $\widehat{\mathbf{V}}$, $\boldsymbol{\delta}_{k+1}$ is a vector of normal random numbers with zero mean and covariance from the prior covariance matrix $\mathbf{Q}$, and ρ, μ, and λ are tuning coefficients that will be discussed below.

Equation 18 is a combination of the two previously mentioned groups of candidate functions. We see that when $\rho = 1$, it is a random walk process; and when $\rho = 0$, it generates iid samples from a scaled multivariate normal distribution from the MAP estimation. ρ takes values between -1 to 1 and it weights the random walk part and the iid part in the candidates. When $\rho = -1$, it is the antithetic sampling, which is known as a variance reduction technique. μ is a scaling factor which defines the step size for candidate generation. The λ term is used to compensate the fact that the MAP covariance estimate $\widehat{\mathbf{V}}$ might not represent the posterior covariance matrix well. In case $\widehat{\mathbf{V}}$ is a good approximation of the posterior covariance, λ can be near 0.

Random walk and iid samples have their own advantages. When random walk is used, the candidate sample is always generated from the neighborhood of the current sample, which means that once the chain hits a feature (mode) of the distribution, that feature will be extensively explored before the chain leaves it. On the contrary, when iid samples are used, the surrogate distribution is independent of the target distribution and all the accepted samples are still mutually independent, hence the autocorrelation of the Markov chain is generally low, which is a desirable property for MCMC samples.

On the other hand, these methods have also disadvantages. For the random walk method, sometimes (when the step size is too small, for instance) it takes a long time for the chain to escape a minor feature, and for the iid sampling method, some areas of the target distribution will practically never be explored when the surrogate distribution is significantly different from the target distribution (for example, when the surrogate distribution has long tails at the areas where the target distribution has probability away from 0).

Equation 18 gives one the flexibility to lean to either one of the two methods depending on the values of ρ and μ. When ρ is large, one tends to use a smaller μ value to down-scale the search area when the Markov chain is away from the MAP best estimate, and vice versa.

Another criterion for selecting the proper ρ, μ and λ is to adjust the values in test runs to get the desired acceptance rate (for instance, 25% as suggested by Chib and Greenberg (1995)) .

The candidate generating density function for Eq. 18 as in Eq. 9 can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{\theta}') \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}' - \rho\boldsymbol{\theta} - (1-\rho)\hat{\boldsymbol{\theta}}\right)^T \left(\mu^2\widehat{\mathbf{V}} + \lambda^2\mathbf{Q}\right)^{-1} \times \left(\boldsymbol{\theta}' - \rho\boldsymbol{\theta} - (1-\rho)\hat{\boldsymbol{\theta}}\right)\right). \tag{19}$$

This equation is not symmetric unless $\rho = \pm 1$, hence it can not be dropped from Eq. 9.

4.5 Diagnosis of MCMC samples

There are some simple tests we can perform to diagnose the samples even when the statistical properties of the target distribution is completely unknown. In this paper, we promote two different posterior sample diagnosis methods that test the samples in two important aspects.

The autocorrelation of the samples as a function of lag distance is an important measure of the quality of the samples. Given a sequence of samples $\{x_i, i = 1,\ldots, n\}$, the autocorrelation coefficient of x as a function of lag distance d is defined as

$$\gamma(d) = \frac{1}{n-d}\sum_{i=1}^{n-d}\frac{(x_i - \mu_q)(x_{i+h} - \mu_s)}{\sigma_q\sigma_s}, \tag{20}$$

where n is the number of samples in $\{x_i\}$; μ_q and σ_q are respectively the mean and standard deviation of the first $n - d$ samples in the chain, and μ_s and σ_s are those of the last $n - d$ samples. Here the maximum value of d is $n - 1$. However, the larger h is, the less samples there are in the two sub-sequences, and hence the less accurate $\gamma(d)$ is. Therefore, we limit the lag distance to be less than $n/2$ so that the two sub-sequences together will always include all the samples in the original chain.

Generally, for a MCMC chain, the autocorrelation coefficient should approach and then stabilize around 0 as a function of the lag distance. In practice, one can define a threshold autocorrelation coefficient (0.2, for example) and record the lag distance where the autocorrelation coefficient hits the threshold value. That lag distance can be used as a measure of the quality of the samples in the sense of autocorrelation.

Furthermore, for any type of iterative modeling such as MCMC sampling, we need to know when to terminate the iteration, and usually we stop the iteration when convergence is reached. Generally there are various ways to define convergence. Cowles and Carlin (1996) reviewed up to 13 convergence diagnostic methods and sorted them according to properties such as single/multiple chains, theoretical basis, applicability, and ease of use. Among all these methods, the one proposed by Gelman and Rubin (1992) is widely used due to its wide applicability and ease of implementation. According to Cowles and Carlin (1996), Gelman and Rubin's (1992) method is based on large-sample normal theory; it is quantitative; and it uses multiple chains. Furthermore, Gelman and Rubin (1992) show that lack of convergence can not be generally examined from a single chain, hence they propose a convergence test for multiple sequences. In this method, a scale reduction factor (SRF) (it is actually $\sqrt{\widehat{R}}$ that is called the scale reduction factor, but we will work on $\widehat{R}$ instead.) is calculated as

$$\widehat{R} = \frac{\widehat{V}}{W}\frac{df+3}{df+1} = \left(\frac{n-1}{n} + \frac{m+1}{mn}\frac{B}{W}\right)\frac{df+3}{df+1} \tag{21}$$

Before explaining what the various terms mean in the equation above, we will give a brief introduction of the process to acquire the SRF.

1. Generate m starting samples as described previously in subsection “Candidate Generating Density Functions”.
2. Run m independent MCMC chains with $2n$ samples in each chain. We will work on the last n samples in each chain only.
3. Choose a scalar statistic x that is a function of one sample (for instance, the quantity we are trying to estimate using the Monte Carlo approach, or $\mathcal{L}(\boldsymbol{\theta})$ in equation Eq. 7 as suggested by Gelman and Rubin (1992)); calculate
   - $\frac{B}{n} = \frac{1}{m-1}\sum_{i=1}^{m}(x_{i.} - x_{..})^2$, the variance of the m chain means $\bar{x}_{i.}$, where $x_{..}$ is the mean of all samples in all chains.
   - $W = \frac{1}{m}\sum_{i=1}^{m} s_i^2$, the mean of the variances within each chain, where $s_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j} - x_{i.})^2$, and $x_{i,j}$ is the jth component in the ith chain.
4. Calculate $df = \frac{\widehat{V^2}}{\widehat{\mathrm{Var}}(\widehat{V})}$ where

$$\widehat{V} = \frac{B}{mn} = \frac{n-1}{n}W + \frac{(m+1)B}{mn}$$

and

$$\widehat{\mathrm{Var}\left(\widehat{V}\right)} = \left(\frac{n-1}{n}\right)^2\frac{1}{m}\widehat{\mathrm{Var}(s_i^2)} + \left(\frac{m+1}{mn}\right)^2\frac{2}{m-1}B^2 + 2\frac{(m+1)(n-1)}{mn^2}\frac{n}{m}\left[\widehat{\mathrm{Cov}(s_i^2,\bar{x}_{i.}^2)} - 2\bar{x}_{..}\widehat{\mathrm{Cov}(s_i^2,\bar{x}_{i.})}\right]$$

5. The $\widehat{R}$ statistic is calculated as in Eq. 21

In a deterministic iterative process, such as the Gauss-Newton iteration, convergence is claimed when the improvement in the objective function value, gradient, step size, etc. on the current iteration is small. Similarly, we claim the convergence of an MCMC sampling process when the improvement in some statistic(s) on the current iteration is small. The idea behind Gelman and Rubin’s (1992) method is that, starting from an overdispersed distribution, the multiple chains will start from different areas of the target distribution. In the beginning, these chains should have drastically different statistical properties, however, as the chains evolve and more samples are generated, after a point, all the chains will have approximately traveled through the whole support domain of the target distribution such that the statistical properties of each separate chain are about the same as those of all the chains together. At this point, convergence can be claimed.

In the process above, we calculate W, the mean of the within-chain variance; and $\widehat{V}$, the cross-chain variance of all chains. Without the df term, the SRF is simply a ratio between the cross-chain variance and the within-chain variance. At convergence we should see the SRF stabilize around 1. In fact, df represents the sampling variability (Fisher 1953), and it is usually a large number when the number of samples is larger, hence it can practically be dropped when the number of samples is large.

Brooks and Gelman (1998) corrected an error on the df term in Eq. 21 made by Gelman and Rubin (1992) and argued that in addition to $\widehat{R}$ approaching 1, $\widehat{V}$ and W should also stabilize at convergence. They proposed an iterated graphical approach to monitor the convergence by dividing the t chains into batches of length b; then calculate $\widehat{V}(k), W(k)$, and $\widehat{R}(k)$ using the latter half of sub-chains of length $2kb$, $k = 1, 2, \ldots l/b$; and plot $\widehat{V}(k)$, $W(k)$, and $\widehat{R}(k)$ as a function of k. Thus on the plot, we should expect that the line of $\widehat{V}(k)$ is always on the top of that of $W(k)$ but the two lines get sufficiently close and stabilize at convergence, and at the same time, $\widehat{R}(k)$ should approach 1. The failure of either one indicates lack of convergence.

Cowles and Carlin (1996) claim that any single existing convergence monitoring strategy could fail under certain complex circumstances. Specifically, Cowles and Carlin (1996) showed that for the bimodal mixture of trivariate normal distributions, Gelman and Rubin’s method fails when there is not enough dispersion in the initial samples. This lack of dispersion in the initial samples generally could happen when one or more of the modes of a multimodal distribution is not detected. In practice, we know some prior information of the parameters to be estimated, hence deterministic mode searches can be conducted starting from multiple initial values. Thus unawareness of a mode can be avoided in most cases. In this paper, we use a combination of two simple methods for convergence monitoring. The autocorrelation tells us the quality of the samples, while SRF tells us whether we have sampled the whole support domain or not.

5 An example of DNAPL dissolution and transport

To exemplify the methodologies we propose in this paper, a simple test case is used. We use a modification of the semi-analytic source dissolution and dissolved-phase transport solution developed by Parker et al. (2008). This model has two main sub-modules—a source dissolution sub-module that calculates the net mass flow out of a source zone, and an advection-dispersion-reaction (ADR) sub-module that calculates the transport of the dissolved contaminant plume within the aquifer. The model also simulates effects of remedial actions such as partial source-zone mass removal. In this paper, we will focus on pre-remediation stage parameter estimation.

5.1 Source dissolution sub-module

We utilize a modified version of the Parker and Park (2004) model of field-scale source-zone dissolution to simulate flux from a source zone that can include multiple dissolution architectures (e.g. residual DNAPL and pools). Parker and Park (2004) use an exponential formula to describe non-equilibrium DNAPL dissolution:

$$J_i(t) \approx J_{o,i}\left(\frac{M_i(t)}{M_{o,i}}\right)^{\beta_i} \tag{22}$$

where, for source architecture type i, J_i [M/T] is the mass dissolution rate from the non-aqueous phase to the dissolved phase within the source zone; $J_{o,i}$ [M/T] is the initial mass flow rate at the time of site contamination (t_o); $M_i(t)$ [M] is the current remaining mass of DNAPL; $M_{o,i}$ [M] is the initial mass of NAPL at the time of contamination; and β_i [−] is a mass depletion exponent. β_i measures the speed that DNAPL dissolves, which is greater than 1 for finger-dominated residual and less than 1 for DNAPL pools and lenses.

By solving a mass conservation equation

$$\frac{dM_i(t)}{dt} = -J_i(t) \tag{23}$$

subject to the initial condition $M_i(0) = M_{0,i}$, the analytic solution for source mass remaining in architecture i is:

$$M_i(t) = \begin{cases} \left[\left(M_{o,i}\right)^{1-\beta_i} - (1-\beta_i)\frac{J_{o,i}}{M_{o,i}^{\beta_i}}(t-t_o)\right]^{1/(1-\beta_i)} & \text{for } \beta_i \neq 1 \\ M_{0,i}\exp\left(-\frac{J_{0,i}}{M_{0,i}}(t-t_0)\right) & \text{for } \beta_i = 1 \end{cases} \tag{24}$$

Substituting Eq. 24 back to Eq. 22, yields an analytical solution for $J_i(t)$. Summing up all $J_i(t)$ leads to

$$J_{tot}(t) \approx \sum_i J_i(t). \tag{25}$$

5.2 Advection-dispersion sub-module

The source dissolution sub-module calculates the total mass flow rate of dissolved contaminant that enters the aquifer over time. We assume a contaminant source of width L_y perpendicular to the groundwater flow direction in an aquifer of thickness L_z. Applying the depth averaged solution of Domenico (1987) yields:

$$C(x,y,t) = \int_0^t \frac{J_{tot}(t-\tau)}{4L_z L_y \phi (\pi A_L v\tau)^{1/2}} \exp\left(-\frac{(x-v\tau)^2}{4A_L v\tau}\right) \times \left[\operatorname{erf}\left(-\frac{y-L_y/2}{2(A_T v\tau)^{1/2}}\right) - \operatorname{erf}\left(-\frac{y+L_y/2}{2(A_T v\tau)^{1/2}}\right)\right] d\tau \tag{26}$$

where J_{tot} is the source zone discharge rate from Eq. 25, ϕ is the porosity [−], A_L and A_T are the longitudinal and transverse dispersivities [L], respectively, and v is the aquifer pore velocity [L/T] (specific discharge q_w [L/T] divided by porosity ϕ), assumed to be in the x direction.

5.3 Parameter estimation problem setup

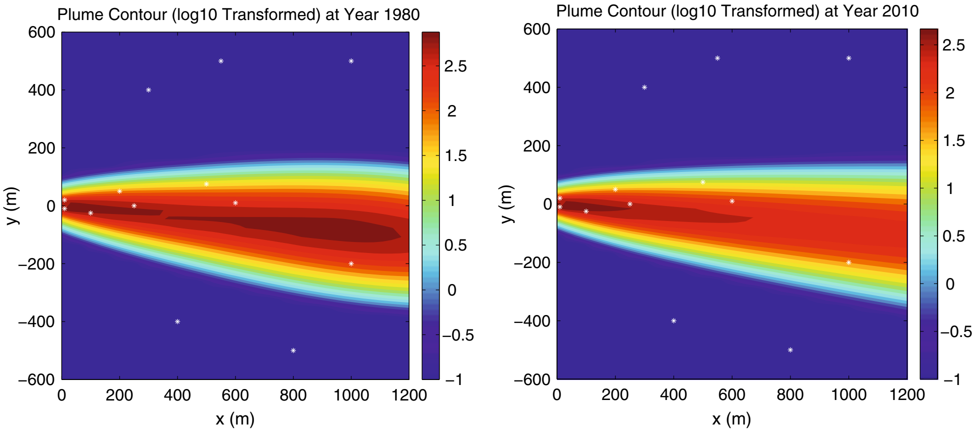
In this paper, we test the parameter estimation scheme on a synthetic problem. A DNAPL source is assumed to occur in 1965. The source zone is comprised of two architectures, one representing pools and the other one representing residual DNAPL. The true groundwater plume extents in 1980 and 2010, and the positions of monitoring wells are shown in Fig. 2. Monitoring wells are assumed to be sampled from 1980 to 2010, totalling 140 samples. DNAPL concentration in the samples is calculated through Eq. 26 and Gaussian noise of various levels (ε of 0.1, 0.01, and 0.001 for the log-concentration) is added to the calculated concentration to represent measurement error.

The true values of the parameters are listed in Table 1 along with pdf of the prior distributions. To test the effects of measurement error on parameter estimation, three different levels of noise (ε of 0.1, 0.01, and 0.001) were added to natural logarithms of the measurements to represent measurement error and conceptual model deviations.

We enforce a physical constraint on the porosity (ϕ) such that it is between 0 and 1. In addition, to avoid the un-identifiability issue between the two architectures, we enforce another constraint such that $M_{0,1} > M_{0,2}$. The preceding constraints fully define the support domain of the parameters ($\boldsymbol{\Theta}$) and it will be included in the prior distribution of the parameters.

In addition to the prior distributions, the likelihood function, or the conditional distribution of the measurements (log-transformed dissolved contaminant concentrations) given a set of parameter values follows a multi-variate normal distribution. Its expectations are calculated through Eq. 26 and its covariance matrix $\epsilon^2\mathbf{I}_{140\times 140}$ is diagonal. Approximately, $\varepsilon = 0.1$ represents a noise level of 10%, $\varepsilon = 0.01$ of 1%, and $\varepsilon = 0.001$ of 0.1%. The last case is rare in practice, however, it serves well as a numerical exercise because under this situation, the posterior distribution is almost the same as the likelihood function, which is highly nonlinear and generally difficult to analyze using classical methods.

Fig. 2 Plumes of DNAPL in the aquifer at years 1980 and 2010. The *white dots* represent the observation points



(a) Plume extent at 1980

(b) Plume extent at 2010

Table 1 The variables that will be estimated in the model and their true values and prior estimates

Variable	Unit	Description	Model	True	Prior	Prior σ
$M_{0,1}$	kg	Initial contaminant mass deposited in architecture 1	LN	8.41	7.82	0.60
$M_{0,2}$	kg	Initial contaminant mass deposited in architecture 2	LN	6.91	6.21	0.60
$J_{0,1}$	kg/d	Initial flux out of architecture 1	LN	−2.30	−2.53	0.20
$J_{0,2}$	kg/d	Initial flux out of architecture 2	LN	−0.69	−0.92	0.20
β_1	–	Mass depletion exponent for architecture 1	LN	−0.51	−0.80	0.35
β_2	—	Mass depletion exponent for architecture 2	LN	0.26	0.34	0.20
A_L	m	Longitudinal dispersivity	LN	3.00	2.71	0.20
A_T	m	Transverse dispersivity	LN	0.69	0.83	0.10
L_y	m	Width of source zone	LN	3.00	3.14	0.10
q_w	m/d	Ground water Darcy velocity	LN	−2.66	−2.53	0.35
ϕ	–	Aquifer porosity	N	0.30	0.32	0.01
Y_0	y	Time of initial contaminant mass deposition	N	1965.00	1963.00	1.00
α	°	Direction of region flow	N	5.00	0.00	4.00

Table 2 MAP estimates and the estimation uncertainty

Variable	True	MAP 1 ($\varepsilon = 0.1$)		MAP 2 ($\varepsilon = 0.01$)		MAP 3 ($\varepsilon = 0.001$)	
		Estimate	σ	Estimate	σ	Estimate	σ
$M_{0,1}$	8.41	8.27	3.63E-01	8.23	1.99E-01	8.30	2.29E-03
$M_{0,2}$	6.91	7.03	1.32E-01	6.99	7.40E-02	6.98	1.50E-03
$J_{0,1}$	−2.30	−2.37	1.17E-01	−2.28	6.48E-02	−2.27	1.63E-03
$J_{0,2}$	−0.69	−0.83	1.28E-01	−0.63	5.18E-02	−0.63	2.23E-03
β_1	−0.51	−0.91	3.30E-01	−0.88	2.86E-01	−0.77	8.30E-04
β_2	0.26	0.26	1.58E-01	0.30	9.21E-02	0.28	2.28E-03
A_L	3.00	2.75	9.10E-02	2.99	1.76E-02	3.00	1.17E-03
A_T	0.69	0.70	1.65E-02	0.69	2.26E-03	0.69	1.90E-04
L_y	3.00	3.16	6.69E-02	3.00	1.61E-02	3.00	1.35E-03
q_w	−2.66	−2.64	4.53E-02	−2.61	2.71E-02	−2.61	1.04E-03
ϕ	0.30	0.32	9.88E-03	0.32	8.65E-03	0.32	3.55E-04
Y_0	1965.00	1963.63	7.37E-01	1965.04	1.22E-01	1965.01	9.40E-03
α	5.00	5.03	5.03E-02	5.00	5.33E-03	5.00	4.84E-04

6 Results

We first apply the active-set quasi-Newton minimization method (as "fmincon" in MATLAB does) on Eq. 7 to find the MAP best estimate ($\hat{\mathbf{x}}$) and then compute the sensitivity matrix of Eq. 26 at $\hat{\mathbf{x}}$. Using the sensitivity matrix, we get the linearized covariance matrix $\hat{\mathbf{V}}$. We did several local minimum searches starting at a set of rather sparse points and found that for cases $\varepsilon = 0.1$ and $\varepsilon = 0.01$, all of them converged to the same point. For the case $\varepsilon = 0.001$, it was very difficult for the program to converge and the minimization procedures ended at slightly different points due to very small line search step sizes ($\sim 10^{-10}$). For this case, we used the point with the smallest objective function value as our MAP estimate. The MAP results are presented in Table 2.

Using the MAP results and following the procedure previously introduced, we generated 4 over-dispersed starting samples for each case ($\varepsilon = 0.1, 0.01, 0.001$) that would be used for both MH sampling and ADS sampling. In MH sampling, we simply generated one MCMC chain on each of four computing nodes. There is no communication between the nodes, thus the speed-up is linear. Each chain contains 420000 samples, hence 1680000 samples in total, and the computation time is about 150 min. For the ADS sampling, we insert one adaptive line sampling step between every 20 MH steps, and we use four master nodes that will do most of the calculation and eight slave nodes

Table 3 Major statistics of the samples from MH sampling ($\varepsilon = 0.1$)

Variable	True	MH sequence 1			MH sequence 2			MH sequence 3			MH sequence 4		
		Mean	Mode	σ	Mean	Mode	σ	Mean	Mode	σ	Mean	Mode	σ
$M_{0,1}$	8.41	8.39	8.23	3.67E-01	8.41	8.25	3.90E-01	8.39	8.24	3.68E-01	8.40	8.23	3.82E-01
$M_{0,2}$	6.91	7.06	7.02	1.37E-01	7.06	7.07	1.36E-01	7.06	7.04	1.37E-01	7.06	7.05	1.36E-01
$J_{0,1}$	−2.30	−2.38	−2.40	1.12E-01	−2.39	−2.39	1.11E-01	−2.39	−2.41	1.10E-01	−2.39	−2.41	1.11E-01
$J_{0,2}$	−0.69	−0.82	−0.83	1.29E-01	−0.82	−0.81	1.29E-01	−0.82	−0.85	1.29E-01	−0.82	−0.85	1.29E-01
β_1	−0.51	−0.91	−0.94	3.28E-01	−0.92	−0.89	3.30E-01	−0.91	−0.95	3.31E-01	−0.92	−0.90	3.28E-01
β_2	0.26	0.31	0.27	1.64E-01	0.31	0.30	1.63E-01	0.31	0.29	1.62E-01	0.31	0.28	1.63E-01
A_L	3.00	2.76	2.76	9.00E-02	2.76	2.76	9.08E-02	2.76	2.76	8.96E-02	2.75	2.75	9.13E-02
A_T	0.69	0.70	0.70	1.64E-02	0.70	0.70	1.65E-02	0.70	0.70	1.63E-02	0.70	0.71	1.66E-02
L_y	3.00	3.15	3.17	6.93E-02	3.15	3.16	6.91E-02	3.15	3.17	6.85E-02	3.15	3.16	6.89E-02
q_w	−2.66	−2.64	−2.63	4.44E-02	−2.64	−2.63	4.49E-02	−2.64	−2.63	4.47E-02	−2.64	−2.64	4.50E-02
ϕ	0.30	0.32	0.32	9.79E-03	0.32	0.32	9.76E-03	0.32	0.32	9.87E-03	0.32	0.32	9.88E-03
Y_0	1965.00	1963.68	1963.57	7.13E-01	1963.69	1963.68	7.16E-01	1963.68	1963.58	7.12E-01	1963.67	1963.57	7.19E-01
α	5.00	5.03	5.02	5.10E-02	5.03	5.02	5.11E-02	5.03	5.03	5.09E-02	5.03	5.03	5.08E-02

Table 4 Major statistics of the samples from ADS sampling ($\varepsilon = 0.1$)

Variable	True	ADS sequence 1			ADS sequence 2			ADS sequence 3			ADS sequence 4		
		Mean	Mode	σ	Mean	Mode	σ	Mean	Mode	σ	Mean	Mode	σ
$M_{0,1}$	8.41	8.39	8.29	3.71E-01	8.37	8.30	3.55E-01	8.39	8.29	3.62E-01	8.41	8.26	3.88E-01
$M_{0,2}$	6.91	7.06	7.07	1.34E-01	7.06	7.06	1.36E-01	7.06	7.07	1.39E-01	7.07	7.09	1.38E-01
$J_{0,1}$	−2.30	−2.39	−2.41	1.09E-01	−2.39	−2.39	1.12E-01	−2.39	−2.39	1.11E-01	−2.39	−2.39	1.09E-01
$J_{0,2}$	−0.69	−0.82	−0.85	1.29E-01	−0.82	−0.82	1.30E-01	−0.82	−0.84	1.30E-01	−0.82	−0.82	1.29E-01
β_1	−0.51	−0.91	−0.99	3.27E-01	−0.92	−0.91	3.30E-01	−0.92	−0.99	3.28E-01	−0.92	−0.95	3.24E-01
β_2	0.26	0.30	0.28	1.63E-01	0.31	0.31	1.61E-01	0.31	0.30	1.66E-01	0.31	0.33	1.68E-01
A_L	3.00	2.76	2.77	9.05E-02	2.75	2.77	9.25E-02	2.75	2.73	9.07E-02	2.75	2.75	9.07E-02
A_T	0.69	0.70	0.70	1.64E-02	0.70	0.70	1.64E-02	0.70	0.70	1.65E-02	0.70	0.70	1.64E-02
L_y	3.00	3.15	3.15	6.88E-02	3.15	3.16	6.93E-02	3.15	3.13	6.85E-02	3.15	3.16	6.84E-02
q_w	−2.66	−2.64	−2.64	4.45E-02	−2.64	−2.63	4.46E-02	−2.64	−2.64	4.50E-02	−2.64	−2.63	4.53E-02
ϕ	0.30	0.32	0.32	9.86E-03	0.32	0.32	9.89E-03	0.32	0.32	9.87E-03	0.32	0.32	9.75E-03
Y_0	1965.00	1963.67	1963.58	7.18E-01	1963.66	1963.79	7.28E-01	1963.68	1963.73	7.19E-01	1963.66	1963.70	7.23E-01
α	5.00	5.03	5.02	5.13E-02	5.03	5.03	5.07E-02	5.03	5.02	5.09E-02	5.03	5.02	5.13E-02

that exclusively evaluate density values in the multiple-try line sampling procedure. With this parallelization strategy, the computation time of ADS sampling to generate the same amount of samples is about 170 min.

For the case $\varepsilon = 0.1$, we show in Table 3 the major statistics for each chain of the samples from MH sampling, and in Table 4 we show those from ADS sampling. These tables, together with the data from the other cases that are not shown here, indicate that all chains give similar results and the difference between chains is subtle, however, it is hard to evaluate the quality of the chains simply based on these tables.

In Figs. 3 and 4, we show for case $\varepsilon = 0.1$ the autocorrelation plots from both sampling methods. We see that the samples are weakly correlated and after tens of steps, the autocorrelation drops to a rather low level. Furthermore, the figures show that samples from both methods have similar autocorrelation patterns. In fact, the samples from ADS sampling are slightly less correlated than those from MH sampling.

Figures 5 and 6 show for case $\varepsilon = 0.1$ a comparison of the SRF plots of the samples from both sampling methods. First of all, these plots display that both methods have converged with rather tight convergence criteria. Second, we can see from the SRF plots in Fig. 5 that with the ADS method, the SRF stabilizes earlier and gets close to 1 than with the MH method. A similar patter can be seen in Fig. 6 that the within-chain variance and cross-chain variance converge and stabilizes earlier with the ADS method than with the MH method.

In Fig. 7 we present for case $\varepsilon = 0.1$ the histograms of the samples from all four chains in ADS sampling. In

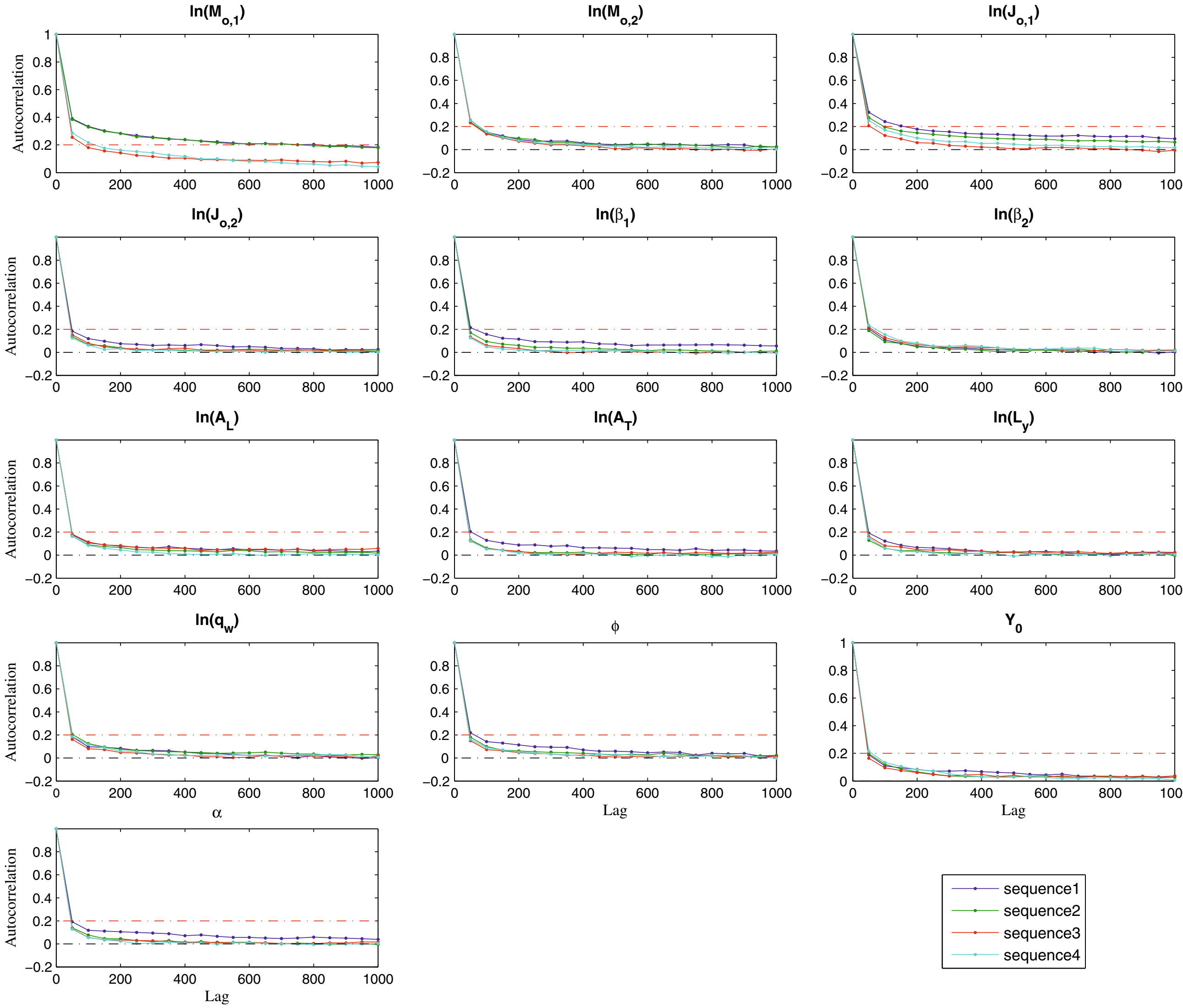


Fig. 3 ACF function plot of the samples from MH sampling ($\varepsilon = 0.1$)

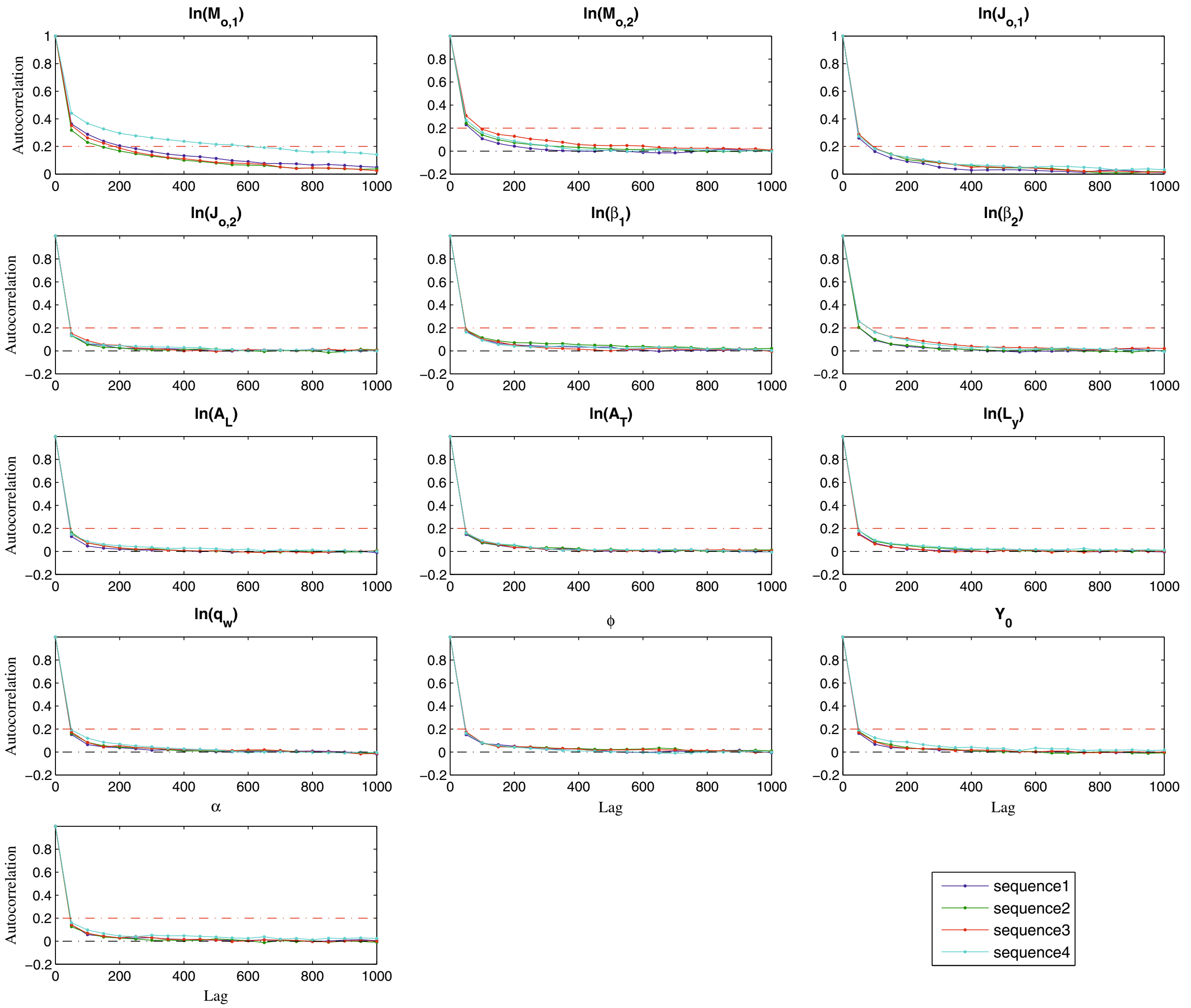
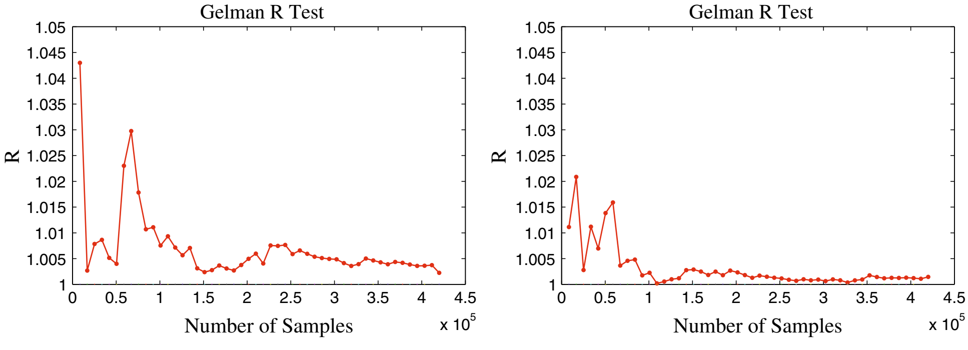


Fig. 4 ACF function plot of the samples from ADS sampling ($\varepsilon = 0.1$)

Fig. 5 SRF plot of the samples from MH and ADS sampling ($\varepsilon = 0.1$)

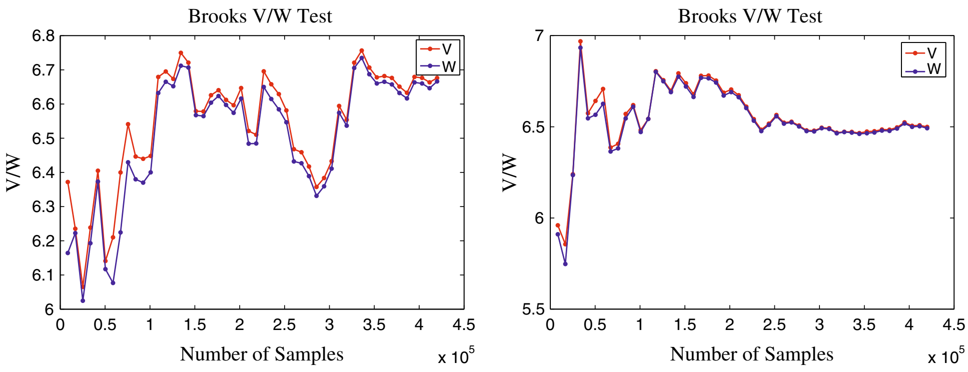


(a) MH sampling **(b)** ADS sampling

these figures, we also plot the prior marginal density functions and MAP approximated marginal density functions. We see from Fig. 7 that the MAP method satisfactorily approximates the posterior distribution. In fact, for most of the parameters, the approximation is nearly exact.

Figure 8 displays for case $\varepsilon = 0.01$ the histograms of the samples from all four chains in ADS sampling. In

Fig. 6 Cross-chain/within-chain variance plot of the samples from MH and ADS sampling ($\varepsilon = 0.1$)



(a) MH sampling **(b)** ADS sampling

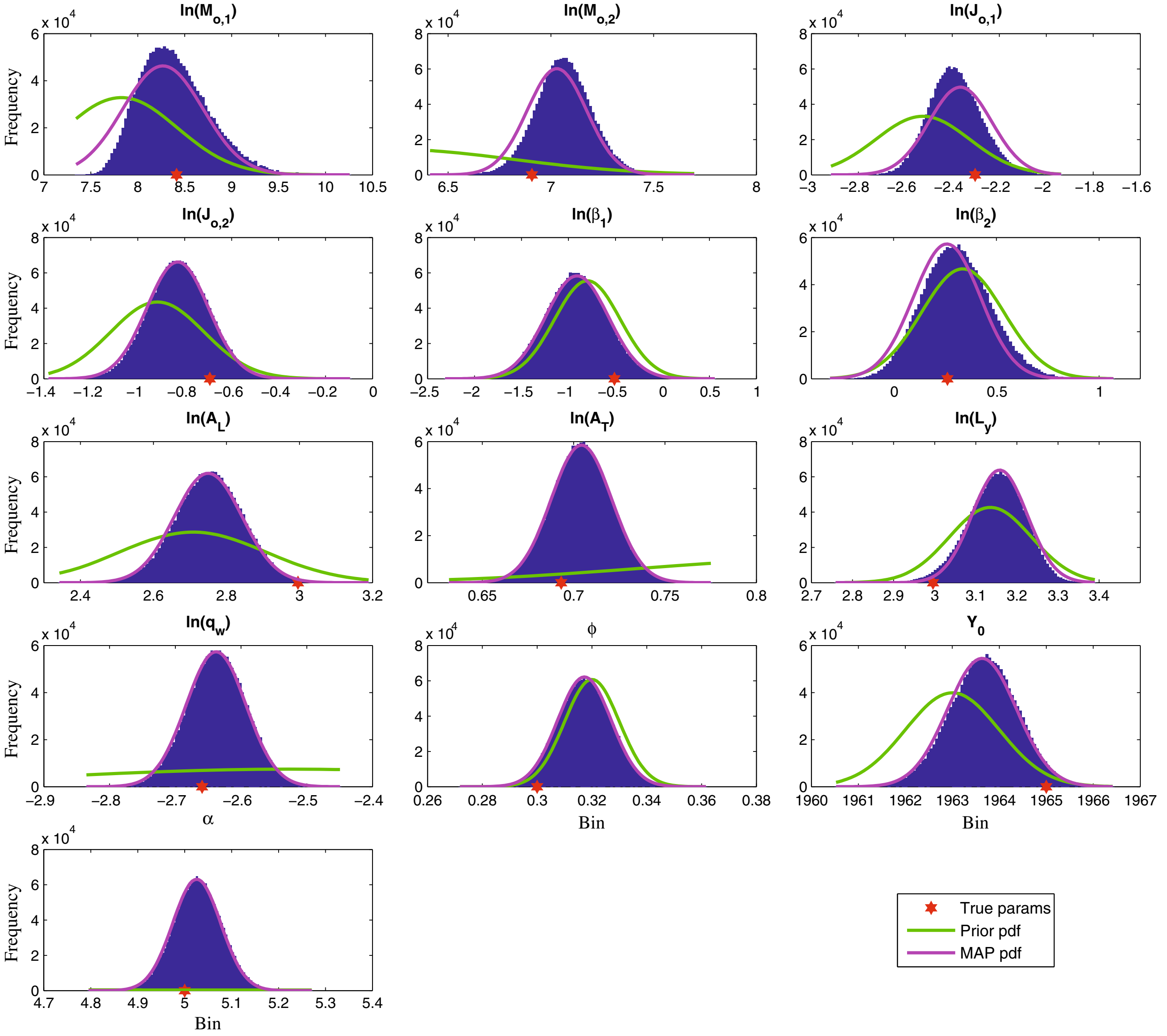


Fig. 7 Histograms of the samples from ADS sampling ($\varepsilon = 0.1$)

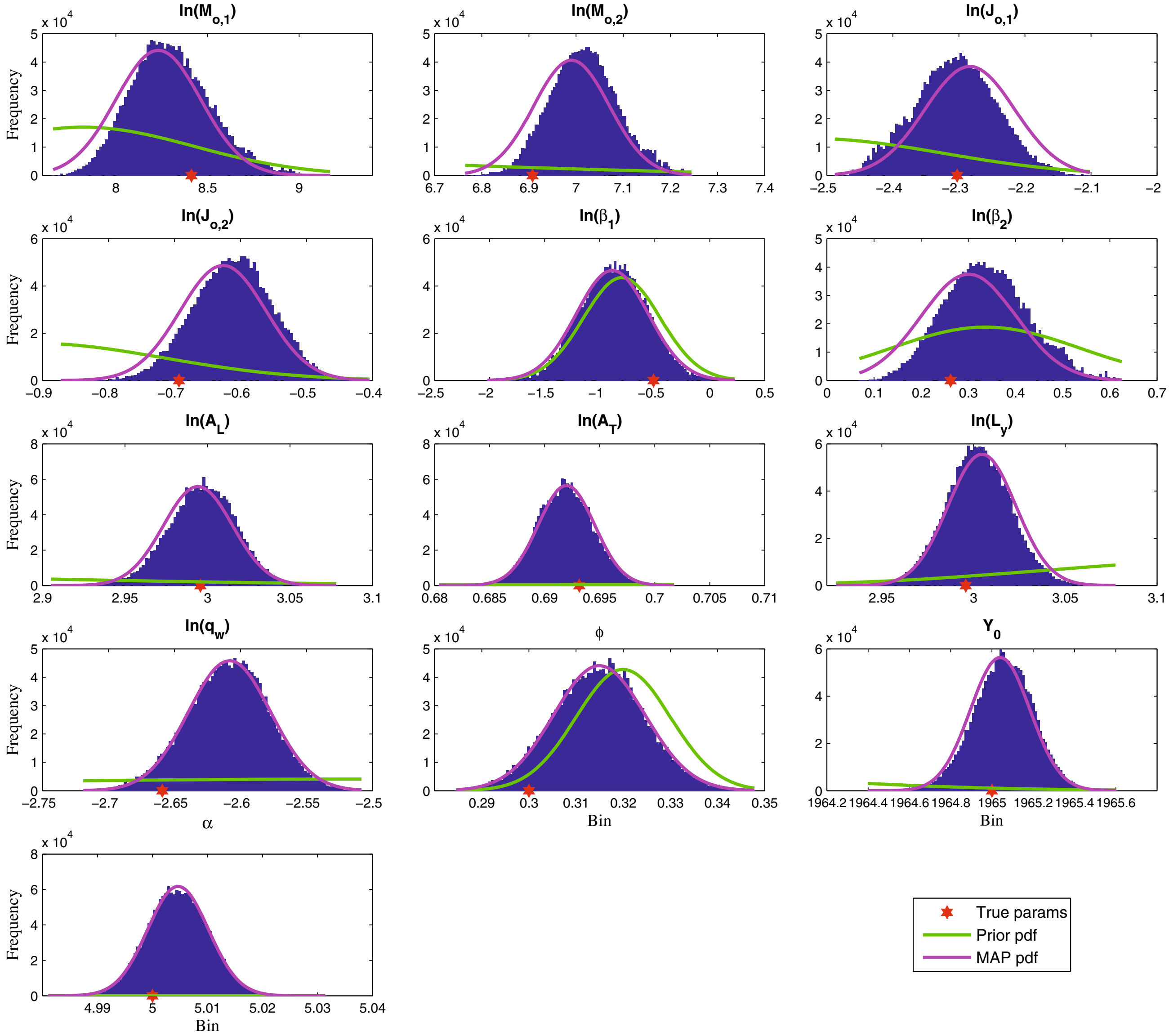


Fig. 8 Histograms of the samples from ADS sampling ($\varepsilon = 0.01$)

this figure, we clearly see that as the noise level in the data decreases, and the posterior distribution leans more to the likelihood function than in the previous case, and the MAP method provides acceptable but less accurate estimation than it does in the previous case. When the noise level in the data gets even smaller, as shown in Fig. 9 for case $\varepsilon = 0.001$, many of the MAP estimates are biased, however, the variance estimation is still acceptable. What is more, we generated 10 times more samples for this case to get the sampling process converge. For the two cases mentioned in this paragraph, the convergence tests are shown in Figs. 10, 11, 12, and 13, and they show similar phenomena as the other case aforementioned.

7 Conclusions and discussion

In this paper, we reviewed several parameter estimation methods and showed an application to a semi-analytical DNAPL dissolution/transport model. We also tested a relatively new MCMC sampler, the ADS sampler with the sample problem in this paper. Our results showed that generally, the MAP with a Gaussian model approximated the posterior distribution quite well. As the noise level got lower, the MAP approximation slightly deviated from the posterior distribution. In the extremely case with very small measurement noise, on one hand, it was difficult for the MAP method to converge to the right mode; on the other hand, it is difficult for

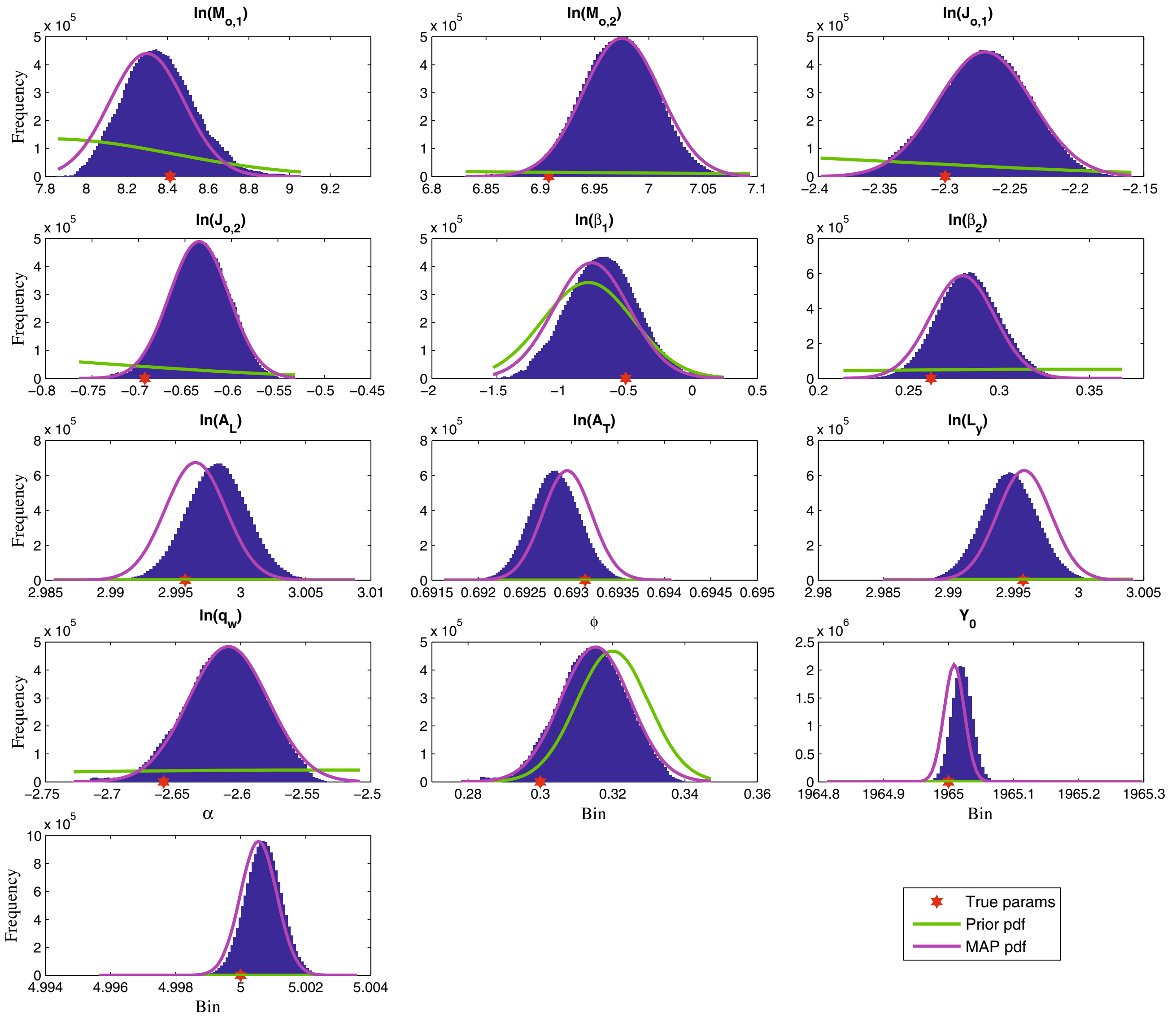
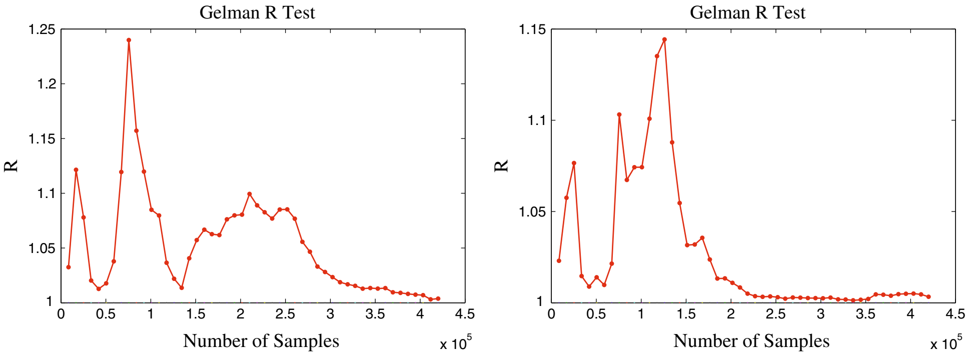


Fig. 9 Histograms of the samples from ADS sampling ($\varepsilon = 0.001$)

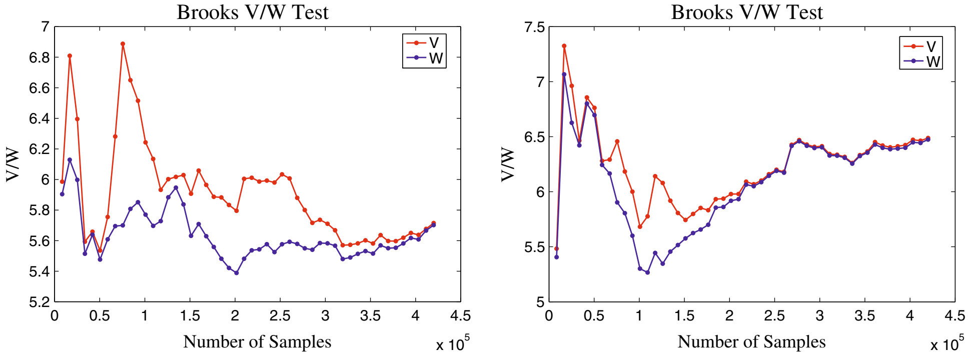
Fig. 10 SRF plot of the samples from MH and ADS sampling ($\varepsilon = 0.01$)



(a) MH sampling **(b)** ADS sampling

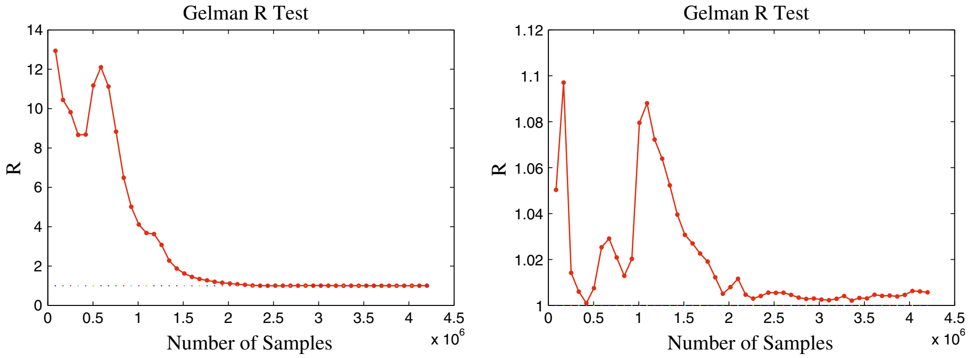
the MCMC method to converge too. Furthermore, the MAP method still provides acceptable variance estimation.

We introduced several methods to diagnose the MCMC samples and under that framework, we compared the performance of the MH sampler and the ADS sampler. We



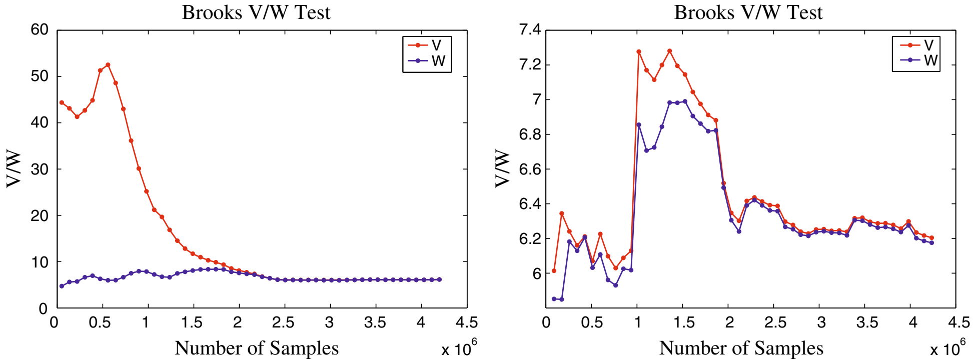
(a) MH sampling **(b)** ADS sampling

Fig. 11 Cross-chain/within-chain variance plot of the samples from MH and ADS sampling ($\varepsilon = 0.01$)



(a) MH sampling **(b)** ADS sampling

Fig. 12 SRF plot of the samples from MH and ADS sampling ($\varepsilon = 0.001$)



(a) MH sampling **(b)** ADS sampling

Fig. 13 Cross-chain/within-chain variance plot of the samples from MH and ADS sampling ($\varepsilon = 0.001$)

found that the ADS method was superior to the MH sampling method in both autocorrelation of the samples and convergence rate.

The benefit of using multiple chains in this paper is twofold. First, it fits into the convergence analysis frame proposed by Gelman and Rubin (1992); second, this strategy is easily parallelizable, hence the efficiency of sampling can be increased to a rather large extent. The parallelization of the MH sampling is rather easy because there is no communication between computing nodes. For the ADS sampling, communication between computing nodes is needed but it only requires minor modifications to the original single-node model.

Acknowledgements This research was funded by the U.S. Department of Defense Strategic Environmental Research and Development Program (SERDP) Environmental Restoration Focus Area managed by Andrea Leeson under project ER-1611 entitled “Practical Cost Optimization of Characterization and Remediation

Decisions at DNAPL sites with Consideration of Prediction Uncertainty.” Additional funding was provided by the Stanford Center for Computational Earth and Environmental Science.

References

Bard Y (June 1973) Nonlinear parameter estimation. Academic Press, London, ISBN 0120782502

Bates BC, Campbell EP (2001) A markov chain Monte carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. Water Resour Res 37(4):937–947

Blasone RS, Madsen H, Rosbjerg D (2008) Uncertainty assessment of integrated distributed hydrological models using glue with markov chain Monte carlo sampling. J Hydrol 353(1–2):18–32

Brooks SP, Gelman A (1998) General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 7(4):434–455

Butler JJ, McElwee CD, Bohling GC (1999) Pumping tests in networks of multilevel sampling wells: motivation and methodology. Water Resour Res 35(11):3553–3560

Carrera J, Neuman SP (1986) Estimation of aquifer parameters under transient and steady-state conditions. 1. maximum-likelihood method incorporating prior information. Water Resour Res 22(2):199–210

Chib S, Greenberg E (1995) Understanding the Metropolis–Hastings algorithm. Am Stat 49(4):327–335

Cowles MK, Carlin BP (1996) Markov chain Monte carlo convergence diagnostics: a comparative review. J Am Stat Assoc 91(434):883–904

Domenico PA (1987) An analytical model for multidimensional transport of a decaying contaminant species. J Hydrol 91(1–2):49–58

Dorrie H (1965) 100 Great problems of elementary mathematics, Dover Publications, NY, pp 73–77

Duane S, Kennedy AD, Pendleton BJ, Roweth D (1987) Hybrid Monte carlo. Phys Lett B 195(2):216–222. doi:10.1016/0370-2693(87)91197-X

Feyen L, Vrugt JA, Nuallain BO, van der Knijff J, De Roo A (2007) Parameter optimisation and uncertainty assessment for large-scale streamflow simulation with the lisflood model. J Hydrol 332:276–289

Fisher AR (1953) The design of experiments, 6th edn. Hafner Pub. Co, NY

Franssen HJH, Gomez-Hernandez J, Sahuquillo A (2003) Coupled inverse modelling of groundwater flow and mass transport and the worth of concentration data. J Hydrol 281(4):281–295

Franssen HJH, Alcolea A, Riva M, Bakr M, van der Wiel N, Stauffer F, Guadagnini A (2009) A comparison of seven methods for the inverse modelling of groundwater flow. application to the characterisation of well catchments. Adv Water Resour 32(6):851–872

Frind EO, Pinder GF (1973) Galerkin solution of inverse problem for aquifer transmissivity. Water Resour Res 9(5):1397–1410

Fu JL, Gomez-Hernandez JJ (2009) Uncertainty assessment and data worth in groundwater flow and mass transport modeling using a blocking markov chain Monte carlo method. J Hydrol 364(3–4):328–341

Gelman A, Rubin DB (1992) A single series from the gibbs sampler provides a false sense of security. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian statistics, 4th edn. Oxford University Press, New York, pp 625–631

Geman S, Geman D (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6(6):721–741.

Gilks WR, Roberts GO, George EI (1994) Adaptive direction sampling. Stat 43(1):179–189

Gill PE, Murray W (1981) Practical optimization. Academic Press, London, pp 113–113

Gomez-Hernandez JJ, Sahuquillo A, Capilla JE (1997) Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data—i. theory. J Hydrol 203(1–4):162–174

Gomez-Hernandez JJ, Franssen HJH, Cassiraga EF (2001) Stochastic analysis of flow response in a three-dimensional fractured rock mass block. International J Rock Mech Mini Sci 38(1):31–44

Hastings WK (1970) Monte carlo sampling methods using markov chains and their applications. Biometrika 57(1):97–109. ISSN 00063444

Kentel E, Aral MM (2005) 2D Monte carlo versus 2D fuzzy Monte carlo health risk assessment. Stoch Environ Res Risk Assess 19(1):86–96. doi:10.1007/s00477-004-0209-1

Kitanidis PK, Lane RW (1985) Maximum-likelihood parameter-estimation of hydrologic spatial processes by the gauss-newton method. J Hydrol 79(1–2):53–71

Kitanidis PK, Vomvoris EG (1983) A geostatistical approach to the inverse problem in groundwater modeling (steady-state) and one-dimensional simulations. Water Resour Res 19(3):677–690

Kuczera G, Parent E (1998) Monte carlo assessment of parameter uncertainty in conceptual catchment models: the metropolis algorithm. J Hydrol 211(1–4):69–85

Lavenue M, de Marsily G (2001) Three-dimensional interference test interpretation in a fractured aquifer using the pilot point inverse method. Water Resour Res 37(11):2659–2675

Liu JS, Liang FM, Wong WH (2000) The multiple-try method and local optimization in metropolis sampling. J Am Stat Assoc 95(449):121–134

Liu X, Illman WA, Craig AJ, Zhu J, Yeh TCJ (2007) Laboratory sandbox validation of transient hydraulic tomography. Water Resour Res 43(5):13. doi:10.1029/2006WR005144

Marshall L, Nott D, Sharma A (2004) A comparative study of markov chain Monte carlo methods for conceptual rainfall-runoff modeling. Water Resour Res 40(2). doi:10.1029/2003WR002378

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092. doi:10.1063/1.1699114

Michalak AM (2008) A gibbs sampler for inequality-constrained geostatistical interpolation and inverse modeling. Water Resour Res 44(9)W09437

Michalak AM, Kitanidis PK (2003) A method for enforcing parameter nonnegativity in bayesian inverse problems with an application to contaminant source identification. Water Resour Res 39(2):SBH 7/1–SBH 7/14

Oliver DS, Cunha LB, Reynolds AC (1997) Markov chain Monte carlo methods for conditioning a permeability field to pressure data. Math Geol 29(1):61–91

Parker JC, Park E (2004) Modeling field-scale dense nonaqueous phase liquid dissolution kinetics in heterogeneous aquifers. Water Resour Res 40(5). doi:10.1029/2003WR002807

Parker JC, Park E, Tang G (2008) Dissolved plume attenuation with dnapl source remediation, aqueous decay and volatilization—analytical solution, model calibration and prediction uncertainty. J Contam Hydrol 102(1–2):61–71. doi:10.1016/j.jconhyd.2008.03.009

Ramarao BS, Lavenue AM, Demarsily G, Marietta MG (1995) Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields. 1. theory and computational experiments. Water Resour Res 31(3):475–493

Ritter C, Tanner MA (1992) Facilitating the gibbs sampler - the gibbs stopper and the griddy-gibbs sampler. J Am Stat Assoc 87(419):861–868

Sahuquillo A, Capilla JE, Gómez-Hernández JJ, Andreu J (1992) Conditional simulation of transmissivity fields honoring piezometric data. In: Blain WR, Cabrera E (eds) Fluid flow modeling, hydraulic engineering software, vol 2 of 4. Elsevier, Oxford, pp 201–214

Smith TJ, Marshall LA (2008) Bayesian methods in hydrologic modeling: a study of recent advancements in markov chain Monte Carlo techniques. Water Resour Res 44. doi:10.1029/2007WR006705

van den Bos A (2007) Parameter estimation for scientists and engineers. Wiley-Interscience, Hoboken, NJ, 273 pp. ISBN 0470147814

von Neumann J (1951) Various techniques used in connection with random digits. In: Householder AS, Forsythe GE, Germond HH (eds) Monte Carlo method. National Bureau of Standards Applied Mathematics Series, vol 12. U.S. Government Printing Office, Washington, DC, pp 36–38

Vrugt JA, Gupta HV, Bouten W, Sorooshian S (2003) A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour Res 39(8) 1.1–1.16

Vrugt JA, ter Braak CJF, Clark MP, Hyman JM, Robinson BA (2008) Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with markov chain Monte carlo simulation. Water Resour Res 44. doi:10.1029/2007WR006720

Vrugt JA, ter Braak CJF, Gupta HV, Robinson BA (2009) Equifinality of formal (DREAM) and informal (GLUE) bayesian approaches in hydrologic modeling? Stoch Environ Res Risk Assess 23(7):1011–1026. doi:10.1007/s00477-008-0274-y

Yeh TCJ, Liu SY (2000) Hydraulic tomography: development of a new aquifer test method. Water Resour Res 36(8)2095–2105