

Dynamic Data Driven Simulations in Stochastic Environments

C. Douglas, Lexington, Y. Efendiev, R. Ewing, College Station,
V. Ginting, Fort Collins, and R. Lazarov, College Station

Received September 13, 2005; revised March 20, 2006

Published online: June 6, 2006

© Springer-Verlag 2006

Abstract

To improve the predictions in dynamic data driven simulations (DDDAS) for subsurface problems, we propose the permeability update based on observed measurements. Based on measurement errors and *a priori* information about the permeability field, such as covariance of permeability field and its values at the measurement locations, the permeability field is sampled. This sampling problem is highly nonlinear and Markov chain Monte Carlo (MCMC) method is used. We show that using the sampled realizations of the permeability field, the predictions can be significantly improved and the uncertainties can be assessed for this highly nonlinear problem.

AMS Subject Classifications: 65N99.

Keywords: MCMC, porous media flow, uncertainty, permeability, DDDAS.

1. Introduction

Dynamic data driven simulations (DDDAS) are important for many practical applications. Consider an extreme example of a disaster scenario in which a major waste spill occurs in a subsurface near a clean water aquifer. Sensors can now be used to measure where the contamination is, where the contaminant is going to go, and to monitor the environmental impact of the spill.

One of the objectives of dynamic data driven simulations is to incorporate the sensor data into real-time simulations that run continuously. Unlike traditional approaches, in which a static input data set is used as initial conditions only, our approach assimilates many sets of data and corrects computed errors above a given level (which can change during the course of the simulation) as part of the computational process. Many important issues are involved in DDDAS for this application and some of them are described in [3].

Subsurface formations typically exhibit heterogeneities over a wide range of length scales whereas the sensors are usually located at sparse locations and sparse data from these discrete points in a domain is broad-casted. Because the sensor data usually contains noise, it can be imposed either as a *hard* or a *soft constraint*. Previously, to incorporate the sensor data into the simulations, we introduced a multiscale interpolation operator. This is done in the context of general nonlinear

parabolic operators that include many subsurface processes. The main idea of this interpolation is that we do not alter the heterogeneities of the random field that drives the contaminant. Instead, based on the sensor data, we rescale the solution in a manner that it preserves the heterogeneities. This interpolation technique fits nicely with a new multiscale framework for solving nonlinear partial differential equations.

The interpolation technique is only a temporary remedy because it does not correct the error sources that occur in the initial data, as well as in the permeability field. Previously, we addressed the initial data correction. However, one of the main sources of the errors is the permeability field. Indeed, permeability fields are typically given by their covariance matrix and some local measurements. In this setting, one can have large uncertainties, and, consequently, it is important to reduce the uncertainties by incorporating the additional data, such as the available information about the contaminant.

In this paper, our goal is to study the permeability correction. The permeability represents the properties of the porous media. Its correction is substantially different from the correction of the initial data. In particular, the problem of permeability correction is strongly nonlinear and stochastic.

In this paper, we also assume that the covariance of the permeability field and some of its values at the measurement points are known. Using Karhunen-Loève expansion, the permeability is expanded based on the covariance matrix. This allows some dimension reduction, which can be further reduced by incorporating the permeability values at sensor locations. Furthermore, based on measurement errors and *a priori* information about the permeability field, we consider the sampling of the permeability field. This sampling problem is highly nonlinear and the posterior distribution is multimodal. We use the Markov Chain Monte Carlo (MCMC) method to sample from this multimodal distribution. We show that using the sampled realizations of the permeability field, the predictions can be significantly improved. Moreover, the proposed technique allows assessment of the uncertainties for this highly nonlinear problem. The proposed approach is general and can be applied to more complicated porous media problems.

The paper is organized as follows. In the next section, we describe the methodology. Section 3 is devoted to numerical results. Conclusions are presented in Sect. 4.

2. Methodology

We study the problem of contaminant transport in heterogeneous porous media. The model equations are

$$\frac{\partial C}{\partial t} + v \cdot \nabla C - \nabla \cdot (D \nabla C) = 0 \quad \text{in } \Omega, \quad (1)$$

where by Darcy's Law, $v = -k \nabla p$, in which the pressure p satisfies

$$-\nabla \cdot (k \nabla p) = 0. \quad (2)$$

Here, k is a generated permeability with a certain statistical variogram and correlation structure, and D is the diffusion coefficient. One of the problems in dynamic data driven simulation is the estimation of the permeability field $k(x)$ given a set of spatially sparse concentration measurements at certain times.

Before presenting the procedure, we shall introduce several relevant notations. Let N_s be the number of sensors installed in various points in the porous medium and $\{x_j\}_{j=1}^{N_s}$ denote such points. Let N_t be the number of times the concentration is measured in time and $\{t_k\}_{k=1}^{N_t}$ denote such times. Furthermore, let $\gamma_j(t_k)$ denote the measured concentration at the sensor located in x_j and at time t_k . We set

$$M(\gamma) = \{\gamma_j(t_k), j = 1, \dots, N_s, k = 1, \dots, N_t\}. \quad (3)$$

Due to measurement errors, the data obtained from the sensors will not necessarily be imposed exactly. Hence, the general idea is to draw a sample of the initial condition from its posterior distribution, which we denote by $P(k(x)|M(\gamma))$. From Bayes' theorem we have

$$P(k(x) | M(\gamma)) \propto P(M(\gamma) | k(x))P(k(x)), \quad (4)$$

where $P(M(\gamma) | k(x))$ is the likelihood probability distribution, and $P(k(x))$ is the prior probability distribution.

Our main goal is to sample the posterior distribution correctly. The posterior distribution, which is a conditional probability distribution, maps the observed data into the permeability fields. This mapping is multi-valued because many permeability fields can give the same observed data. Thus, the posterior distribution is often multimodal. To reduce the effect of multimodality, one can use the prior information. In particular, if the prior has a large weight, then the posterior distribution may have one minimum. However, for the examples that are considered in the paper the posterior distribution is multi-modal, and its efficient sampling is a difficult and computationally expensive task.

To draw a sample from the posterior distribution, we will be using the Markov Chain Monte Carlo (MCMC) approach with the Metropolis-Hasting rule. This way, the minimization procedure is carried out in a Bayesian framework. The MCMC scheme can be carried out by updating the permeability field using the Metropolis-Hasting algorithm. To describe the algorithm, we denote $\pi(k) = P(k(x)|M(\gamma))$.

Algorithm (Metropolis-Hasting MCMC [7])

- Step 1: At k_n generate k from $q(k|k_n)$ (instrumental proposal distribution).
- Step 2: Accept k as a sample with probability

$$P(k_n, k) = \min \left(1, \frac{q(k_n|k)\pi(k)}{q(k|k_n)\pi(k_n)} \right),$$

i.e., $k_{n+1} = k$ with probability $p(k_n, k)$ and $k_{n+1} = k_n$ with probability $1 - p(k_n, k)$.

The main idea of Metropolis-Hasting MCMC is to create a Markov chain which converges to the steady state distribution $P(k(x)|M(\gamma))$. It was shown (e.g., [7]) that the realizations accepted in Metropolis-Hasting MCMC will constitute the Markov chain which converges to the desired probability distribution. One can view Metropolis-Hasting rule as a selection criteria which allows us to sample from the desired probability distribution. For example, if $q(k_n|k) = q(k|k_n)$, then Metropolis-Hasting MCMC will accept realizations that reduce the value of the objective function, while it will also accept the realizations that increase the value of the objective function. The acceptance probability of the latter will be equal to the rate of the increase of the objective function. For general instrumental distributions, one also needs to take into account the ratio $q(k_n|k)/q(k|k_n)$.

To perform the sampling, we use the Karhunen-Loève expansion to parameterize the permeability field. We assume the permeability field is given on a $N \times N$ grid with a prescribed covariance matrix. Consequently, the permeability field is the vector of size N^2 . Because the covariance structure of the permeability is known, we can reduce the dimensionality of permeability space by considering the eigenvectors of the covariance matrix.

Using Karhunen-Loève expansion [5], the permeability field will be expanded in terms of an optimal L^2 basis. We will truncate the expansion using the eigenvalue structure of the covariance matrix. To impose the hard constraint (the values of the permeability at the prescribed locations), we will find a linear subspace of our parameter space (a hyperplane) that yields the corresponding values of the permeability field. First, we will briefly recall the essentials of the Karhunen-Loève expansion. Denote $Y(x, \omega) = \log[k(x, \omega)]$, where $x \in \Omega = [0, 1]^2$, and ω is a random element in a probability space. Suppose $Y(x, \omega)$ is a second-order stochastic process, that is, $Y(x, \omega) \in L^2(\Omega)$ with a probability of one. We will assume that $E[Y(x, \omega)] = 0$. Given an arbitrary orthonormal basis $\{\phi_k\}$ in $L^2(\Omega)$, we can expand $Y(x, \omega)$ as $Y(x, \omega) = \sum_{k=1}^{\infty} Y_k(\omega)\phi_k(x)$, where

$$Y_k(\omega) = \int_{\Omega} Y(x)\phi_k(x)dx$$

are random variables. We are interested in the special L^2 basis $\{\phi_k\}$, which makes Y_k uncorrelated, $E[Y_i Y_j] = 0$ for all $i \neq j$. Denote the covariance function of Y as $R(x, y) = E[Y(x)Y(y)]$. Then such basis functions $\{\phi_k\}$ satisfy

$$E[Y_i Y_j] = \int_{\Omega} \phi_i(x)dx \int_{\Omega} R(x, y)\phi_j(y)dy = 0, \quad i \neq j.$$

Because $\{\phi_k\}$ is a complete basis in $L^2(\Omega)$, it follows that $\phi_k(x)$ are eigenfunctions of $R(x, y)$:

$$\int_{\Omega} R(x, y)\phi_k(y)dy = \lambda_k\phi_k(x), \quad k = 1, 2, \dots, \quad (5)$$

where $\lambda_k = E[Y_k^2] > 0$. Furthermore, we have

$$R(x, y) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(y). \tag{6}$$

Denote $\theta_k = Y_k / \sqrt{\lambda_k}$, then θ_k satisfies $E(\theta_k) = 0$ and $E(\theta_i \theta_j) = \delta_{ij}$. Then

$$Y(x, \omega) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} \theta_k(\omega) \phi_k(x), \tag{7}$$

where ϕ_k and λ_k satisfy (5). We assume that eigenvalues λ_k are ordered so that $\lambda_1 \geq \lambda_2 \geq \dots$. The expansion (7) is called the Karhunen-Loève expansion (KLE). In (7), the L^2 basis functions $\phi_k(x)$ are deterministic and resolve the spatial dependence of the permeability field. The randomness is represented by the scalar random variables θ_k . Generally, we only need to keep the leading order terms (quantified by the magnitude of λ_k) and still capture most of the energy of the stochastic process $Y(x, \omega)$. For a N -term KLE approximation $Y_N = \sum_{k=1}^N \sqrt{\lambda_k} \theta_k \phi_k$, we define the energy ratio of the approximation as

$$e(N) := \frac{E \|Y_N\|^2}{E \|Y\|^2} = \frac{\sum_{k=1}^N \lambda_k}{\sum_{k=1}^{\infty} \lambda_k}.$$

If λ_k decays very fast, then the truncated KLE would be a good approximation of the stochastic process in the L^2 sense.

Suppose the permeability field $k(x, \omega)$ is a log normal homogeneous stochastic process, then $Y(x, \omega)$ is a Gaussian process and θ_k are independent standard Gaussian random variables. We assume that the covariance function of $Y(x, \omega)$ bears the form

$$R(x, y) = \sigma_Y^2 \exp\left(-\frac{|x_1 - y_1|^2}{2L_1^2} - \frac{|x_2 - y_2|^2}{2L_2^2}\right). \tag{8}$$

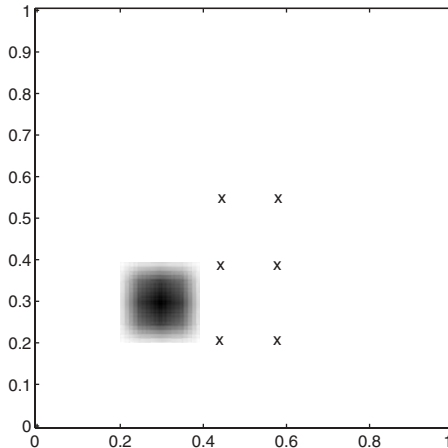


Fig. 1. The initial condition profile and sensors

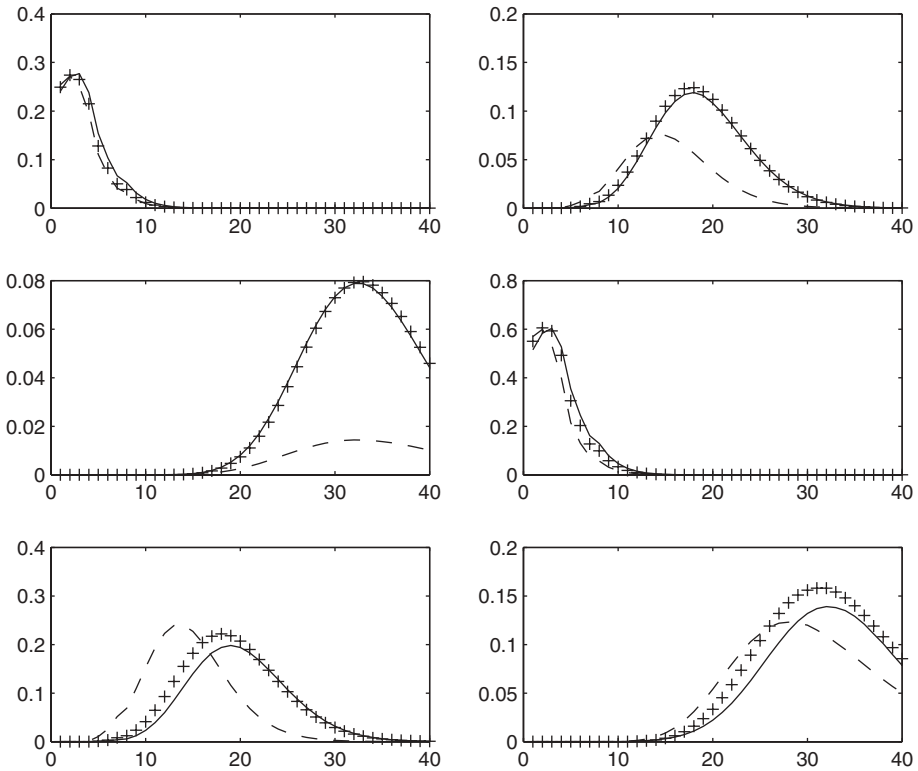


Fig. 2. Concentration at different time instances for random walk sampler with $\sigma = 0.01$ – the solid line designates the observed concentration, the dashed line designates the first match, and the data marked with “+” designates the concentration after the measurement information is incorporated into the simulations

In the above formula, L_1 and L_2 are the correlation length in each dimension and $\sigma_Y^2 = E(Y^2)$ is a constant. We first solve the eigenvalue problem (5) numerically and obtain the eigenpairs $\{\lambda_k, \phi_k\}$. Hence, the truncated KLE should approximate the stochastic process $Y(x, \omega)$ fairly well. Therefore, we can sample $Y(x, \omega)$ from the truncated KLE (7) by generating Gaussian random variables θ_k .

The Karhunen-Loève expansion is used only for the purpose of efficient parameterization of permeability field. Typically, the permeability fields are defined over the underlying grid where the number of grid blocks are large. Direct parameterization of the permeability field over large number of grid blocks results to the very large dimensional parameter spaces that can not be handled in rigorous sampling procedures. Karhunen-Loève expansion allows us to reduce the parameter space dimension by using only dominant eigenvectors of the covariance matrix.

3. Numerical Results

In the simulation, we first generate a true (reference) permeability field using all eigenvectors and compute corresponding observed data at sensor locations. To

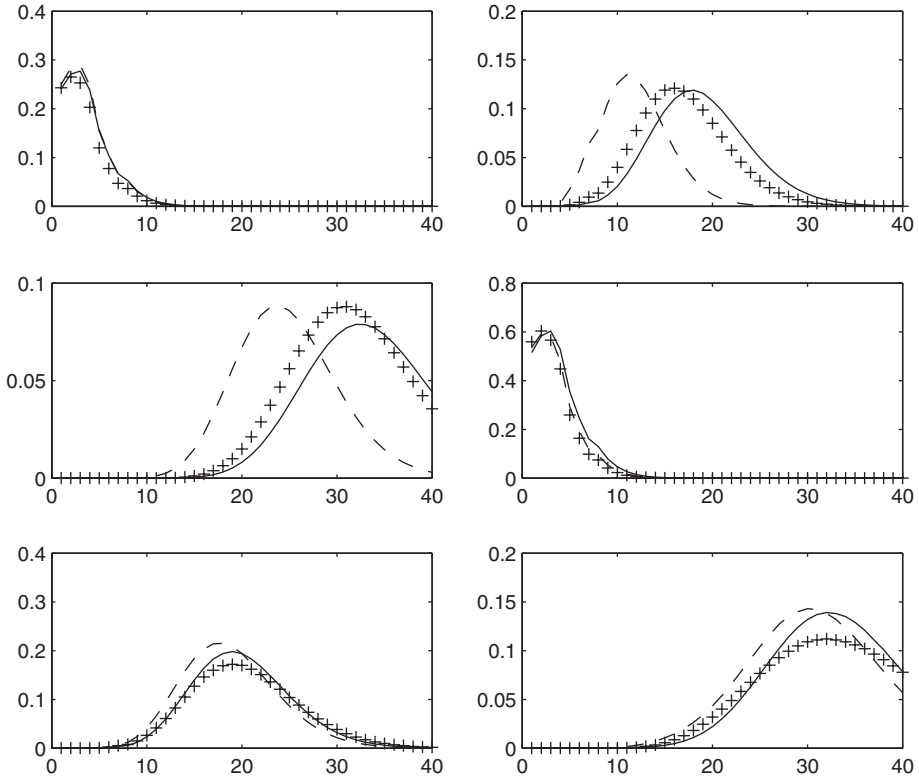


Fig. 3. Concentration at different time instances for independent sampler with $\sigma = 0.01$ – the solid line designates the observed concentration, the dashed line designates the first match, and the data marked with “+” designates the concentration after the measurement information is incorporated into the simulations

propose permeability fields from the prior distribution, we assume that at sensor locations, the permeability field is known. This condition is imposed by setting

$$\sum_{k=1}^{20} \sqrt{\lambda_k} \theta_k \phi_k(x_j) = \alpha_j, \tag{9}$$

where α_j ($j = 1, \dots, 6$) are prescribed constants. In our simulations, we propose 14 θ_i and calculate the rest of θ_i from (9). The permeability values at points x_j are assumed to be 1. The prior distribution is taken to be Gaussian with correlation lengths $L_1 = 0.4$, $L_2 = 0.1$, and $\sigma^2 = 2$.

Using (7) expansion of the permeability field, the problem reduced the sampling from

$$\pi(\theta) \propto P(M(\gamma)|k(x))P(k(x)),$$

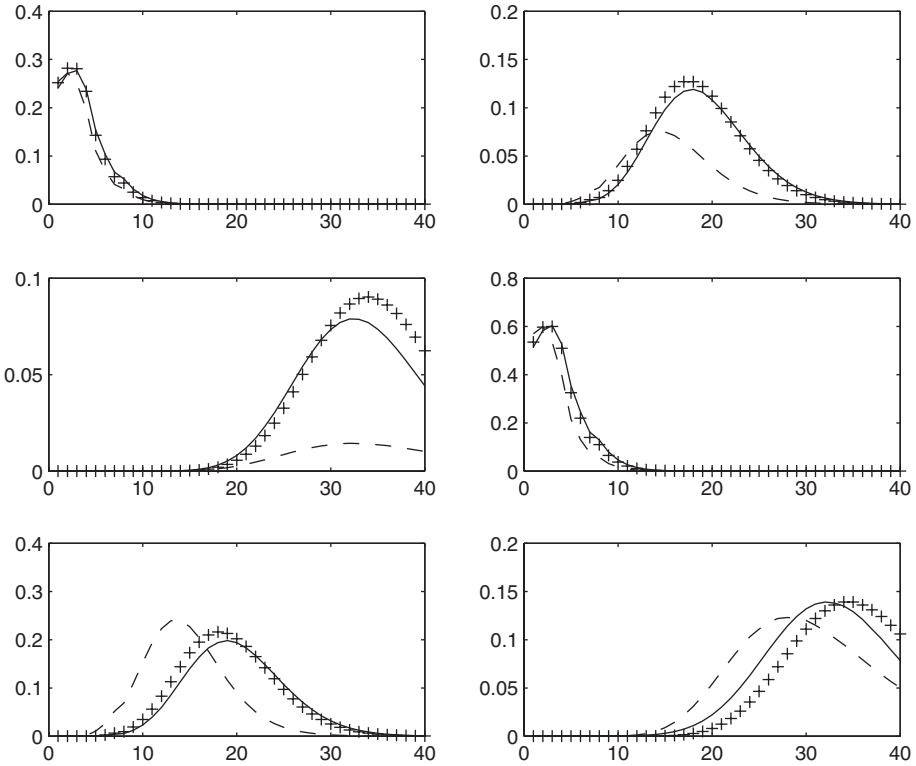


Fig. 4. Concentration at different time instances for random walk sampler with $\sigma = 0.005$ – the solid line designates the observed concentration, the dashed line designates the first match, and the data marked with “+” designates the concentration after the measurement information is incorporated into the simulations

where $M(\gamma)$ is computed by (3) that involves the solution of (1). As for the likelihood, we take

$$P(M(\gamma)|k(x)) = \exp\left(-\frac{\sum_{k=1}^{N_t} \sum_{j=1}^{N_s} (C(x_j, t_k) - \gamma_j(t_k))^2}{\sigma}\right).$$

Note that $C(x_j, t_k)$ depends on θ in a strongly nonlinear manner.

Next, we briefly describe the sampling procedure. At each iteration, new permeability field is proposed and the response (the concentration at sensor location as a function of time) is computed. Further, the acceptance probability is computed based on the values of objective function. Thus, each iteration requires a forward solution of (2.1)–(2.2). In our numerical examples, we use independent sampler and random walk sampler as a proposal. Random walk sampler has high acceptance rate, however, it can get stuck in a local extrema. One can use gradient based proposals, such as Langevin proposals, in sampling the posterior distribution. Langevin proposals require the computation of the gradient of objective functions and the proposal

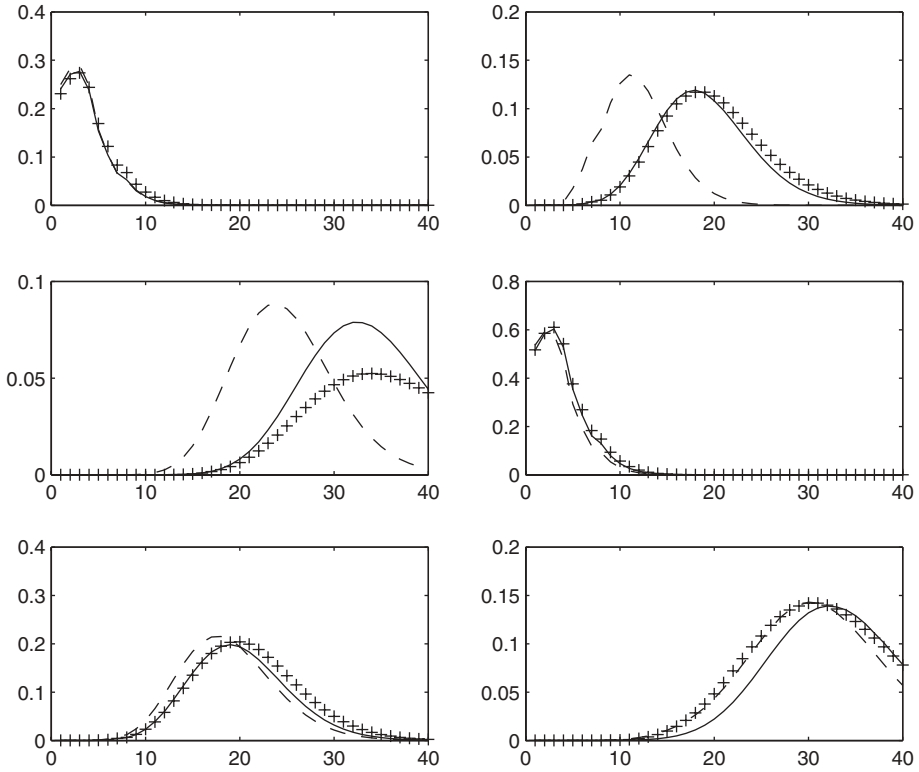


Fig. 5. Concentration at different time instances for independent sampler with $\sigma = 0.005$ – the solid line designates the observed concentration, the dashed line designates the first match, and the data marked with “+” designates the concentration after the measurement information is incorporated into the simulations

is computed using this gradient information. This is in a sense similar to gradient based minimization procedures with the goal to sample the posterior distribution.

In the case of independent samplers, the proposal distribution $q(k|k_n)$ is chosen to be independent of k_n and equal to the prior (unconditioned) distribution. In the random walk sampler, the proposal distribution depends on the previous value of the permeability field and is given by

$$q(k|k_n) = k_n + \epsilon_n, \tag{10}$$

where ϵ_n is a random perturbation with a prescribed variance. If the variance is chosen to be very large, then the random walk sampler becomes similar to the independent sampler. Although the random walk sampler allows us to accept more realizations, it often gets stuck in the neighborhood of a local maximum of the target distribution. In all the simulations, we use 10000 proposals.

As for boundary conditions for (1), we assume that the equations are solved in a square domain $[0, 1]^2$ with the following boundary conditions: $p = 1$ at $x = 0$; $p = 1$ at $x = 1$; no flow boundary conditions on the lateral boundaries ($y = 0$ and



Fig. 6. Permeability realizations after each update. Random walk sampler with $\sigma = 0.01$ is used

$y = 1$) for the pressure equation. As for the concentration equation, we assume no flow boundary conditions. The initial distribution of the contaminant is taken to be $C = 1$ in a subregion, as it is shown in Fig. 1. In this figure, we also illustrate the sensor locations.

To solve the contaminant transport described by (2.1)–(2.2), we use finite volume methods. First, the pressure equation (2.1) is solved and the edge velocity field is computed on the grid. This velocity field is used to solve the transport equation. The convection term is treated explicitly using upwind methods with higher-order limiter, while the diffusion term is treated implicitly in order to avoid diffusive time step restriction.

For our first set of numerical results, we plot the contaminant concentration at sensor locations in Figs. 2–5. In these figures, the solid line designates the observed measurement, the dashed line designates the measurement results obtained using our initial permeability field sampled only from the prior distribution, and the line marked with “+” designates the contaminant concentration for a sampled realization by using MCMC. Each figure represents measurement data (contaminant concentration) as a function of time for each sensor. In these figures, we use the

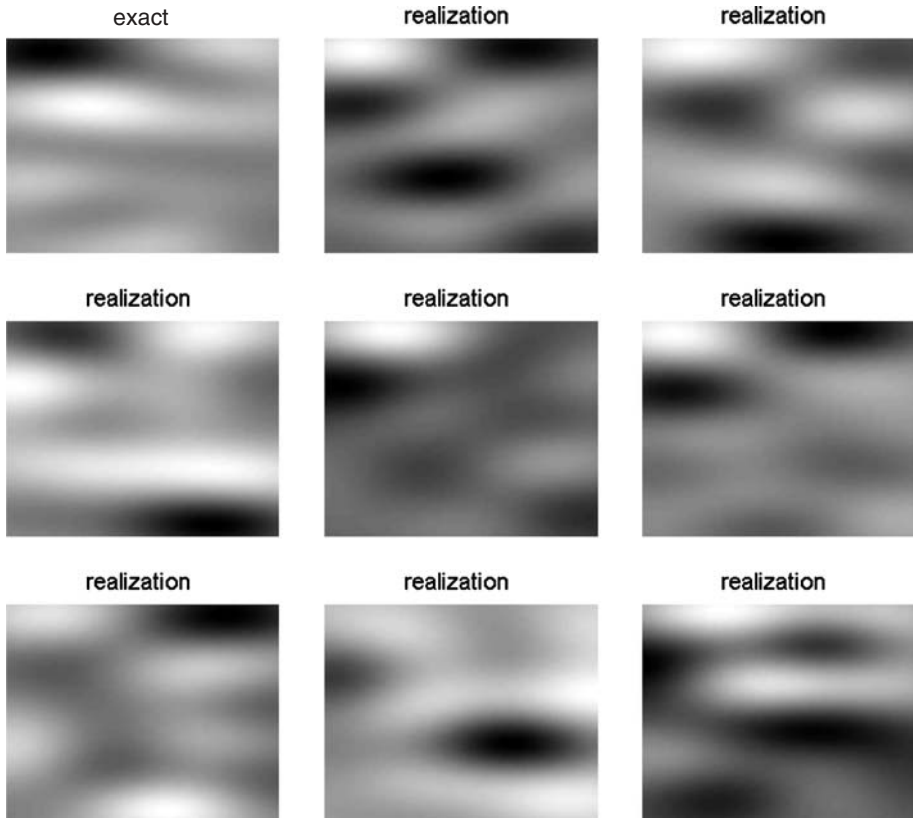


Fig. 7. Permeability realizations after each update. Independent sampler with $\sigma = 0.01$ is used

random walk sampler and independent sampler for measurement errors $\sigma = 0.01$ and $\sigma = 0.005$. It is clear from these figures that the sampled realizations provide good agreement with the observed measurement. Contrary, realizations sampled from prior distribution that are not sampled from $\pi(\theta)$ are not accurate. Moreover, comparing the results for different measurement errors, we observe that smaller measurement errors provide closer agreement with observed data. However, as we mentioned before, the measurement errors are associated with sensor errors, which are dictated by experiments.

In Figs. 6 and 7, we plot the permeability samples obtained using MCMC. We have used $\sigma = 0.01$ for both the independent sampler and random walk sampler. We can observe from this figure that some of the permeability samples are very close to the true permeability field. Typically, using the independent sampler, the acceptance ratio (total number of accepted realizations divided by the total number of proposals) is smaller compared to the case of the random walk sampler. However, the samples obtained using the independent sampler differ from each other significantly.

Finally, we present prediction results in Fig. 8. The left two columns are the results obtained from the MCMC samples, and the right column presents the results

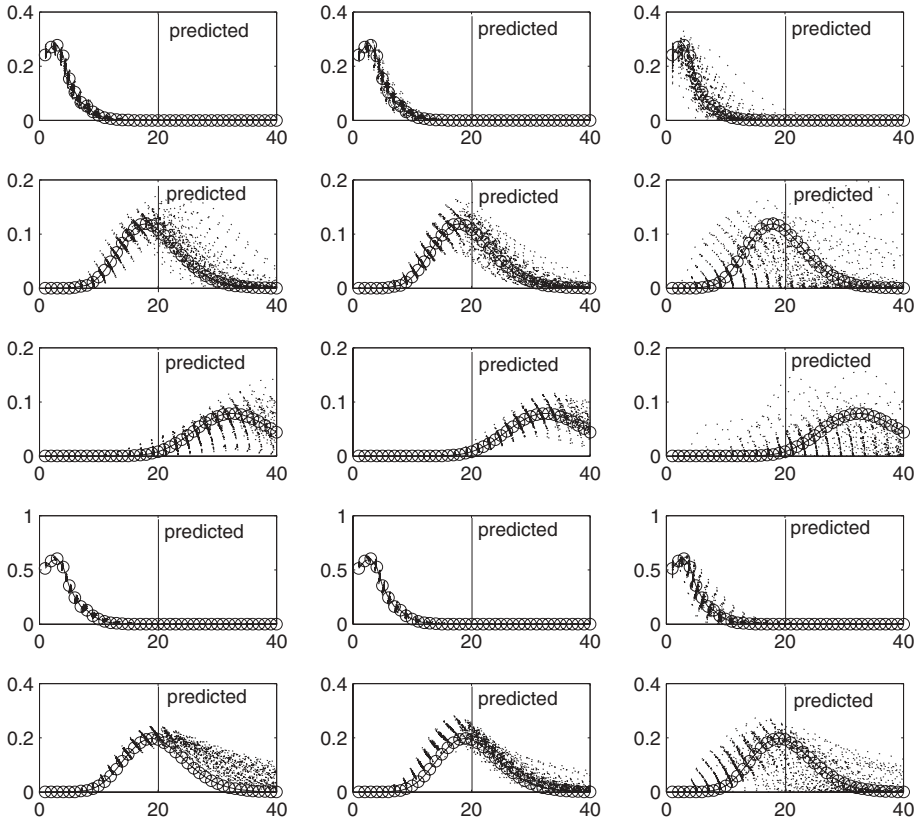


Fig. 8. Each row corresponds to the time history $[0, 20]$ of the concentration at the measurement point and forward prediction. The (o) denotes the measured concentration, whereas the dotted line corresponds to the predicted concentration using the sampled permeability. The first column uses MCMC with $\sigma = 10^{-2}$, the second column uses MCMC with $\sigma = 0.005$, and the third column uses realizations from the prior distribution

obtained without the use of the MCMC samples. In particular, for the results on the left two columns, forward simulations are run using properly sampled permeability fields based on measurements on $[0, 20]$ with two different measurement errors, $\sigma = 0.005$ and $\sigma = 0.01$. As we see from this figure, our method allows us to make accurate predictions for both contaminant concentration as well as possible spread. The latter will allow us to assess the uncertainties in our predictions. Because the sampling problem is highly nonlinear, there is no unique permeability field which gives the measured data. Consequently, it is important to properly sample the posterior distribution to make an accurate assessment of the uncertainties.

4. Conclusions

In this paper, we proposed the procedure for the permeability update based on observed measurements, such as contaminant concentration. Based on measurement errors and *a priori* information about the permeability field, such as covariance of

permeability field and its values at the measurement locations, the permeability field is sampled from complicated, multimodal, posterior distribution. This sampling problem is highly nonlinear and the Markov Chain Monte Carlo (MCMC) method is used. We show that using the sampled realizations of the permeability field, the predictions can be significantly improved and the uncertainties can be assessed for this highly nonlinear problem. In the future, we plan to use preconditioners for the MCMC sampling to reduce the cost. In particular, the solutions of deterministic inverse problems can be used for this purpose.

Acknowledgements

This work is supported in part by NSF grants EIA-0219627, EIA-0218721, EIA-0218229, ACI-0305466, ACI-0324876, CNS-0540178, CNS-0540136 and OISE-0405349.

References

- [1] Deutsch, C. V., Journel, A. G.: *GSLIB: Geostatistical software library and user's guide*, 2nd edition. New York: Oxford University Press 1998.
- [2] Douglas, C. C., Shannon, C., Efendiev, Y., Ewing, R., Ginting, V., Lazarov, R., Cole, M., Jones, G., Johnson, C., Simpson, J.: A note on data-driven contaminant simulation. *Lecture Notes in Computer Science*, vol. 3038. Springer 2004, pp. 701–708.
- [3] Douglas, C. C., Efendiev, Y., Ewing, R., Lazarov, R., Cole, M. R., Johnson, C. R., Jones, G.: Virtual telemetry middleware for DDDAS. *Computational Sciences – ICCS 2003* (Silot, P. M. A., Abramson, D., Dongarra, J. J., Zomaya, A. Y., and Gorbachev, Yu. E., eds.), vol. 4, pp. 279–288.
- [4] Douglas, C. C., Shannon, C., Efendiev, Y., Ewing, R., Ginting, V., Lazarov, R., Cole, M. R., Jones, G., Johnson, C. R., Simpson, J.: Using a virtual telemetry methodology for dynamic data driven application simulations. *Dynamic data driven applications systems* (Darema, F., ed.). Amsterdam: Kluwer 2004.
- [5] Loève, M.: *Probability theory*, 4th edition: Berlin: Springer 1977.
- [6] Oliver, D., Cunha, L., Reynolds, A.: Markov Chain Monte Carlo methods for conditioning a permeability field to pressure data. *Math. Geology* 29, 61–91 (1997).
- [7] Robert, C., Casella, G.: *Monte Carlo statistical methods*. New York: Springer 1999.

C. Douglas
Department of Computer Science
University of Kentucky
773 Anderson Hall
Lexington, KY 40506-0046
USA

and

Department of Computer Science
Yale University
P.O. Box 208285
New Haven, CT 06520-8285
USA
e-mail: craig.douglas@yale.edu

R. Ewing
Institute for Scientific Computation and
Department of Mathematics
Texas A&M University
612 Blocker Hall
College Station, TX 77843-3404
USA
e-mail: richard-ewing@tamu.edu

Y. Efendiev and R. Lazarov
Institute for Scientific Computation and
Department of Mathematics
Texas A&M University
612 Blocker Hall
College Station, TX 77843-3404
USA
e-mails: {efendiev; lazarov}@math.tamu.edu

V. Ginting
Department of Mathematics
Colorado State University
101 Weber Building
Fort Collins
CO 80523-1874, USA
e-mail: ginting@math.colostate.edu