

The use of sampling techniques in dynamic data driven simulations*

C. 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*

Deng Li

*Japan Research Institute, Limited, Engineering Department
Kudan Bldg. 1-5-3 Kudan-Minami, Chiyoda-ku, Tokyo, 102-0074, Japan and
Tokyo Institute of Technology, Graduate School of Science and Engineering
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan and
Department of Computer Science, University of Kentucky,
773 Anderson Hall, Lexington, KY 40506-0046 USA
E-mail: li.deng@uky.edu*

Y. Efendiev, R. Ewing, V. Ginting, R. Lazarov

*Institute for Scientific Computation & Department of Mathematics
Texas A&M University
College Station, TX 77843-3404
E-mail: efendiev, ewing, ginting, lazarov @math.tamu.edu*

In this paper we discuss some numerical procedures involved in dynamic data driven simulations (DDDAS). The main objective of this paper is the recovery of the initial data as well as the media properties based on dynamic data. We consider the contaminant transport in porous media with a number of sensors placed at some locations. Based on measured data we propose to recover the permeability field that characterizes the porous media as well as initial data. Our approaches use Markov chain Monte Carlo methods. Numerical examples are presented.

Keywords: MCMC, DDDAS

AMS Subject classification: Primary 65N99

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

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 while the sensors are usually located at sparse locations and sparse data from these discrete points in a domain is broadcasted. Since 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 have 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. The main idea of this rescaling is the use of local problems. This interpolation technique fits nicely with a new multiscale framework for solving nonlinear partial differential equations. The combination of the interpolation and multiscale frameworks provides a robust and fast simulation technique.

The interpolation technique is only a temporary remedy because it does not fix the error sources which occur in the initial data as well as in the permeability field. The main objective of this paper is to propose a method for adjusting initial data as well as permeability field. In our previous work, we have presented dynamic least-squares technique for initial data recovery. In this paper, we will mostly concentrate on the use of sampling techniques for initial data recovery as well as permeability recovery. Because of the noise in the measurements, the initial data as well as permeability recovery can be treated as a statistical sampling problem. The goal is to sample the parameters given the measurements. We can use Bayes' theorem to formulate the corresponding posterior distribution. However, this distribution is complicated and the rigorous sampling from

it can be achieved using Markov chain Monte Carlo (MCMC) methods. In particular, we employ Metropolis-Hasting rule.

In the initial recovery problem, the initial data is considered as a linear combination of compactly supported functions. This reduces the problem into the finite dimensional problem. In the case of the permeability recovery, we use Karhunen-Loeve expansion to represent the permeability fields given by their covariance structure. In particular, the permeability field is assumed to have a two-point geostatistics with prescribed correlation lengths and variance. Furthermore, at the sensor locations, the permeability field is known. We incorporate this information into our prior distribution. The estimation of permeability field is carried out using Markov chain Monte Carlo approach. As we mentioned earlier, MCMC allows the sampling from a complicated distribution, such as the one obtained in our application. In particular, the estimation of the permeability field based on concentration measurements is strongly nonlinear problem. MCMC allows us to sample rigorously from the posterior distribution.

Some numerical examples are presented. In the problem of initial data recovery, we show that our knowledge about the initial data can be significantly improved by incorporating the measurement data. For the permeability estimation problem, we use a set of measurements to improve our predictions on the permeability distributions. In particular, these approaches will be used in future to develop the techniques that employ Kalman filters.

2. Initial data recovery

The model that we consider is a convection-diffusion problem:

$$\frac{\partial C}{\partial t} + v \cdot \nabla C - \nabla \cdot (D \nabla C) = 0 \text{ in } \Omega \quad (2.1)$$

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

$$-\nabla \cdot (k \nabla p) = 0. \quad (2.2)$$

Here k is a generated permeability with 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 initial condition $C^0(x)$ given a set of spatially sparse concentration measurements at certain times.

Before presenting the procedure, we shall introduce several pertaining 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 be such points. Let N_t be the number of how many times the concentration is measured in time and

$\{t_k\}_{k=1}^{N_t}$ denote such times. Furthermore let $\gamma_j(t_k)$ denotes the measured concentration at sensor located in x_j and at time t_k . We set

$$M(\gamma) = \{\gamma_j(t_k), j = 1, \dots, N_s, k = 1, \dots, N_t\}. \quad (2.3)$$

Suppose we have a set of N_c possible initial conditions $\{\tilde{C}_i^0(x)\}_{i=1}^{N_c}$, where $\tilde{C}_i^0(x)$ are basis functions with support determined a priori, or $\tilde{C}_i^0(x)$ can be functions of certain form determined a priori. Furthermore, we designate $\tilde{C}_i(x, t)$ the solution of (2.1) using an initial condition $\tilde{C}_i^0(x)$. Let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_{N_c})$ be a vector of real numbers, and write

$$\tilde{C}^0(x, t) = \sum_{i=1}^{N_c} \alpha_i \tilde{C}_i^0(x). \quad (2.4)$$

Then obviously the solution of (2.1) with initial condition (2.4) has the following form:

$$\tilde{C}(x, t) = \sum_{i=1}^{N_c} \alpha_i \tilde{C}_i(x, t). \quad (2.5)$$

Due to measurement errors, the data obtained from the sensors will not be necessarily imposed exactly. Hence, the general idea is to draw sample of initial condition from its posterior distribution, which we denote by $P(\tilde{C}^0(x) | M(\gamma))$. From Bayes' theorem we have

$$P(\tilde{C}^0(x) | M(\gamma)) \propto P(M(\gamma) | \tilde{C}^0(x))P(\tilde{C}^0(x)), \quad (2.6)$$

where $P(M(\gamma) | \tilde{C}^0(x))$ is the likelihood probability distribution, and $P(\tilde{C}^0(x))$ is the prior probability distribution.

Using the formulation described above, (2.6) may be expressed as

$$P(\alpha | M(\gamma)) \propto P(M(\gamma) | \alpha)P(\alpha). \quad (2.7)$$

In other words, we may transform the task of estimating the initial condition of (2.1) into a problem of finding the "best" α such that $\tilde{C}^0(x) \approx C^0(x)$. First we introduce the following target function:

$$F(\alpha) = \sum_{k=1}^{N_t} \sum_{j=1}^{N_s} \left(\sum_{i=1}^{N_c} \alpha_i \tilde{C}_i(x_j, t_k) - \gamma_j(t_k) \right)^2 + \kappa \sum_{i=1}^{N_c} (\alpha_i - \beta_i)^2, \quad (2.8)$$

where κ is the penalty coefficient for an a priori vector $\beta = (\beta_1, \beta_2, \dots, \beta_{N_c})$. Using this target function, the posterior probability can be written as a zero-mean Gaussian distribution:

$$P(\alpha | M(\gamma)) = \exp\left(-\frac{F(\alpha)}{2\sigma^2}\right), \quad (2.9)$$

where σ^2 is the variance of the distribution. We note that the first and second term in (2.8) correspond to the likelihood probability and prior probability, respectively.

To draw a sample from the posterior distribution, we will be using Markov Chain Monte Carlo (MCMC) approach with Metropolis-Hasting rule. This way the minimization procedure is carried out in Bayesian framework. MCMC scheme can be carried out by updating α using the Metropolis-Hasting algorithm. In the single step of this algorithm, α^* is generated from a pre-specified proposal distribution $q(\alpha^* | \alpha)$ for a given α . Then the proposed α^* is accepted with probability of acceptance

$$P_r = \min \left\{ 1, \frac{P(\alpha^* | M(\gamma)) q(\alpha | \alpha^*)}{P(\alpha | M(\gamma)) q(\alpha^* | \alpha)} \right\}. \quad (2.10)$$

We note that for linear problems the likelihood probability can be determined using the pre-computed $\tilde{C}(x, t)$ prior to MCMC simulation. MCMC approaches can be applied to nonlinear problems (such as NAPL infiltration) and this approach can give some advantage since it avoids solving linear system for the minimization.

One of the drawbacks of MCMC approaches in subsurface applications is that the acceptance rate can be small. For the linear problems, we follow the idea presented in [6] to increase the acceptance rate. In particular, using the minimization problem one achieve the acceptance probability to be one. We would like to note that this is only true for linear problems, and does not work for nonlinear problems. Next, we will discuss the algorithm where the acceptance probability is one for linear problems. The main idea of this algorithm is to generate samples that have high acceptance probability. We will sample γ and β from some distribution and use them to explore the space of uncertainties. The algorithm is as follows:

1. Propose $\tilde{\gamma}$ and $\tilde{\beta}$ from the following distribution:

$$f(\tilde{\gamma}, \tilde{\beta}) \propto \exp \left(-\frac{1}{2\sigma^2} \left(\sum_{k=1}^{N_t} \sum_{j=1}^{N_s} (\tilde{\gamma}_j(t_k) - \gamma_j(t_k))^2 + \kappa \sum_{i=1}^{N_c} (\tilde{\beta}_i - \beta_i)^2 \right) \right) \quad (2.11)$$

2. Minimize $F(\alpha)$ in (2.8) where $\tilde{\gamma}$ and $\tilde{\beta}$ are used replacing γ and β , respectively.

In this algorithm, we do not need to implement the acceptance step, since the acceptance probability is 1, [6]. We have tested the algorithm and observed the significant improvement of the initial data, reduction of the uncertainty after each update, and better predictions. The numerical

results along with the rigorous analysis of the approach will be presented elsewhere.

3. Permeability recovery

In this section, we consider the sampling permeability field based on measurements. We assume the permeability field is given on a $N \times N$ grid with prescribed covariance matrix. Consequently, the permeability field is the vector of size N^2 . Since the covariance structure of the permeability is known, we can reduce the dimensionality of permeability space by considering the eigenvectors of covariance matrix. Next we describe this procedure.

Using Karhunen-Loeve expansion [5] the permeability field will be expanded in terms of an optimal L^2 basis. We will truncate the expansion using eigenvalues that are less than five percent of the largest eigenvalue. To impose the hard constraint (the values of the permeability at prescribed locations) we will find a linear subspace of our parameter space (a hyperplane) which yields the corresponding values of the permeability field. First we will briefly recall the essentials of Karhunen-Loeve 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 probability 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

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

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

$$\int_{\Omega} R(x, y)\phi_k(y)dy = \lambda_k\phi_k(x), \quad k = 1, 2, \dots, \quad (3.1)$$

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

$$R(x, y) = \sum_{k=1}^{\infty} \lambda_k\phi_k(x)\phi_k(y). \quad (3.2)$$

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

$$Y(x, \omega) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} \theta_k(\omega) \phi_k(x), \quad (3.3)$$

where ϕ_k and λ_k satisfy (3.1). We assume that eigenvalues λ_k are ordered so that $\lambda_1 \geq \lambda_2 \geq \dots$. The expansion (3.3) are called Karhunen-Loeve expansion (KLE). In (3.3), 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 decay very fast, then the truncated KLE would be good approximations of the stochastic process in 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

$$R(x, y) = \sigma^2 \exp\left(-\frac{|x_1 - y_1|^2}{2L_1^2} - \frac{|x_2 - y_2|^2}{2L_2^2}\right). \quad (3.4)$$

In the above formula, L_1 and L_2 are the correlation length in each dimension, and $\sigma^2 = E(Y^2)$ is a constant. We first solve the eigenvalue problem (3.1) 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 (3.3) by generating Gaussian random variables θ_k .

In the simulation, we first generate true (reference) permeability field using all eigenvectors and compute corresponding measurement data. To propose permeability fields from prior distribution, we assume that at 7 distinct points, the permeability field is known. This condition is imposed by setting

$$\sum_{k=1}^{20} \sqrt{\lambda_k} \theta_k \phi_k(x_j) = \alpha_j, \quad (3.5)$$

where α_j ($j = 1, \dots, 8$) are prescribed constants. In our simulations we propose 13 θ_i and calculate the rest of θ_i from (3.5). The permeability

values at points x_j are assumed to be 1. In all the simulations, we propose 20000 realizations.

Next we formulate Metropolis-Hasting MCMC for permeability sampling. From Bayes' theorem we have

$$P(k(x)|M(\gamma)) \propto P(M(\gamma)|k(x))P(k(x)).$$

This relation can be written in terms of θ_i , because the permeability field can be represented in terms of θ_i . The prior distribution is taken to be Gaussian with correlation length $L_1 = 0.4$, $L_2 = 0.1$ and $\sigma^2 = 2$ and with permeability values fixed at sensor locations. In particular, we assume the permeability is one at sensor locations. As for the likelihood, we take the L_2 norm of the difference between the sensor information obtained by solving the equations (2.1) and the observed sensor information. This relationship is strongly nonlinear. We take the measurement precision to be $7e - 4$ and use the random walk sampler as a proposal. In Figure 1, we plot the measurements at each sensor location as a function of time instants. Solid line designates the observed values of the concentration at the sensor locations, dashed line designates the initial predictions of the concentration at the sensor locations, and the lines marked with + designates the concentration after 10000 MCMC iterations at these locations. As we see from this Figure, using MCMC approach, we can obtain adequate permeability samples which provide us with accurate predictions. In Figure 2 we plot the permeability samples obtained using MCMC. We can observe from this figure that some of permeability samples are very close to the true permeability field.

4. Conclusions

In this paper, we used sampling techniques, such as MCMC, in dynamic data driven applications. Our results show that MCMC approaches can reduce uncertainty and allow us to achieve better predictions. More extensive study of these methods in dynamic data driven applications will be reported elsewhere.

References

- [1] C. V. DEUTSCH AND A. G. JOURNAL, *GSLIB: Geostatistical software library and user's guide, 2nd edition*, Oxford University Press, New York, 1998.
- [2] C. C. DOUGLAS, C. SHANNON, Y. EFENDIEV, R. EWING, V. GINTING, R. LAZAROV, M. COLE, G. JONES, C. JOHNSON, AND J. SIMPSON, *A note on data-driven contaminant simulation*, Lecture Notes in Computer Science, Springer-Verlag, 3038 (2004), pp. 701–708.

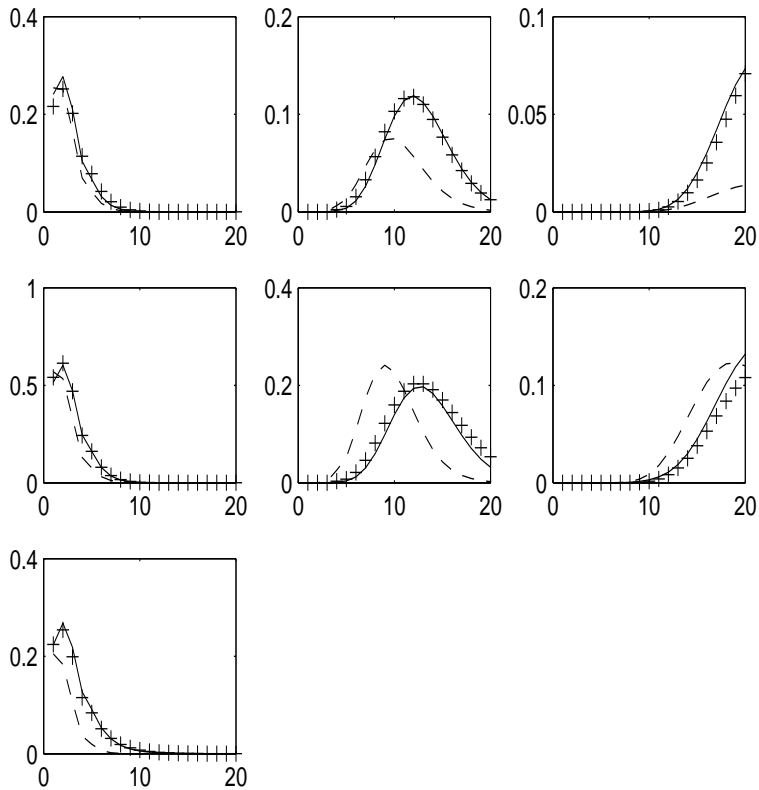


Figure 1. Concentration at different time instants - the solid line designates the observed concentration, the dashed line designates the first match, and the data marked with '+' designates the concentration after the measurement information is incorporated into the simulations

- [3] C. C. DOUGLAS, Y. EFENDIEV, R. EWING, R. LAZAROV, M. R. COLE, C. R. JOHNSON, AND G. JONES, *Virtual telemetry middleware for DDDAS*, Computational Sciences - ICCS 2003, P. M. A. Sloot, D. Abramson, J. J. Dongarra, A. Y. Zomaya, and Yu. E. Gorbachev (eds.), 4 (2003), pp. 279–288.
- [4] C. C. DOUGLAS, C. SHANNON, Y. EFENDIEV, R. EWING, V. GINTING, R. LAZAROV, M. R. COLE, G. JONES, C. R. JOHNSON, AND J. SIMPSON, *Using a virtual telemetry methodology for dynamic data driven application simulations*, in Dynamic Data Driven Applications Systems, F. Darema (ed.), Kluwer, Amsterdam, 2004.
- [5] M. LOEVE, *Probability Theory*, 4th ed., Springer, Berlin, 1977.
- [6] D. OLIVER, L. CUNHA, AND A. REYNOLDS, *Markov chain Monte Carlo methods for conditioning a permeability field to pressure data*, *Mathematical Geology*, 29 (1997).

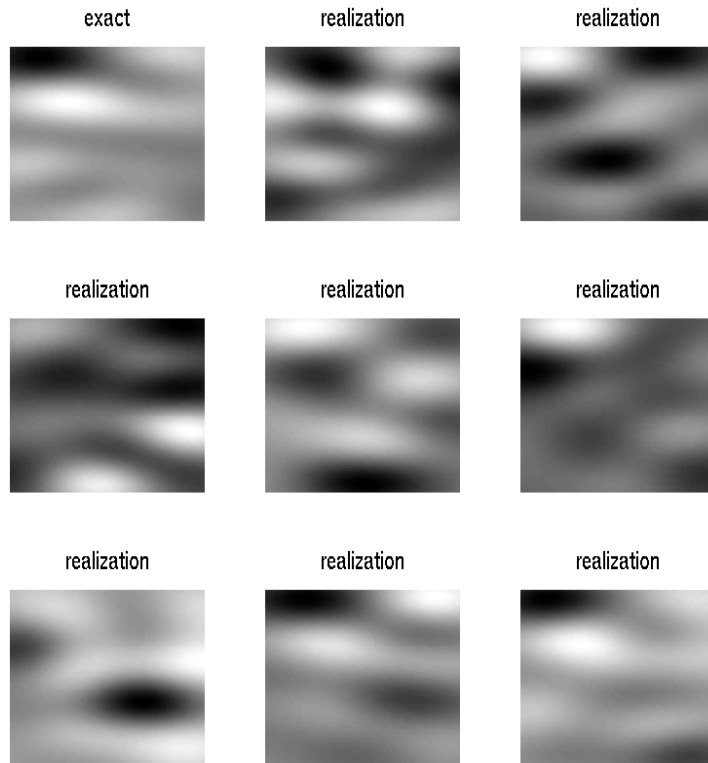


Figure 2. Permeability realizations after each update

- [7] C. ROBERT AND G. CASELLA, *Monte Carlo Statistical Methods*, Springer-Verlag, New-York, 1999.