Dynamic Reservoir Data Assimilation With an Efficient, Dimension-Reduced Kalman Filter

Dongxiao Zhang, SPE, U. of Oklahoma; **Zhiming Lu,** Los Alamos Natl. Laboratory; **Yan Chen,** SPE, U. of Oklahoma

Summary

Kalman filter-based methods have been widely applied for assimilating new measurements to continuously update the estimate of state variables, such as reservoir properties and responses. The standard Kalman filtering scheme requires computing and storing the covariance matrix of state variables, which is computationally expensive for large-scale problems with millions of gridblocks. In the ensemble Kalman filter (EnKF), this problem is alleviated with sampling from a limited number of realizations and computing the required subset of the covariance matrix at each update. However, the goodness of the (ensemble) covariance approximated from the limited ensemble depends on the number of realizations used and the representativity of a given ensemble. In this study, we propose an efficient, dimension-reduced Kalman filtering scheme based on Karhunen-Loeve (KL) and other orthogonal polynomial decompositions of the state variables. We consider flow in heterogeneous reservoirs with spatially variable permeability. The reservoir responses such as pressure are measured at some locations at various time intervals. The aim is to dynamically characterize the reservoir properties and to predict the reservoir performance and its uncertainty at future times. In our scheme, the covariance of the reservoir properties is approximated by a small set of eigenvalues and eigenfunctions using the KL decomposition and the reconstruction of the covariance from the KL decomposition can be done whenever needed. In each update, the forecast step is solved using the KL-based moment method, giving a set of functions from which the mean and covariance of the state variables can be constructed, when needed. The statistics of both the reservoir properties and the reservoir responses are then updated with the available measurements at this time using the auto- and cross-covariances obtained from the forecast step. The new approach is illustrated on a heterogeneous reservoir with dynamic measurements and the results are compared with those from the EnKF method, in terms of accuracy and efficiency.

Introduction

Owing to the high cost associated with direct measurements of reservoir properties, for instance permeability and porosity, the number of direct observations is always limited. However, the reservoir exhibits a high degree of spatial variability at all length scales resulted from its intrinsically complicated nature. This combination of significant spatial heterogeneity with a relatively small number of direct observations leads to uncertainty in characterizing reservoir properties, which in turn results in uncertainty in estimating or predicting the corresponding reservoir responses.

Large efforts have been made to take advantage of all the available observations, both the limited number of direct measurements of reservoir properties and a larger amount of observations of reservoir responses, to obtain better estimates of the primary parameters of the reservoir, thereby reducing the uncertainty associated with model predictions. These methods are usually named as history matching, in which the best estimate of reservoir properties is obtained by minimizing the mismatch between the estimation and the observation. The gradient-based approach, a commonly used method in the traditional history matching, needs substantial computational effort to calculate the gradient of the objective function, either with the adjoint method or other techniques in order to search for the minimum.

Recently, a number of sequential history matching methods have been proposed, which is basically extended from the optimal estimation theory of the Kalman filter. These methods include the Extended Kalman filter (EKF), the Ensemble Kalman filter (EnKF), and their variants. The common ground of these methods is their capability of incorporating observations sequentially, which results in a significant reduction of the number of data dealt at a certain time. An up-to-date system, consisting of an estimate of the unknown properties and the corresponding uncertainty is provided by these sequential methods.

The fundamental difference among these methods is how the statistics of the system state is propagated in time. The EKF uses the first-order linearization and needs to keep track of the whole covariance matrix, which is computationally expensive for large-scale problems. The EnKF employs a Monte Carlo method in which the covariance matrix is updated from a small-sized ensemble (a small number of realizations). The EnKF can alleviate the closure problem imbedded in the EKF and can better handle large-scale, highly nonlinear problems (Evensen, 1994, 2003).

Kalman filter method is only optimal when the dynamic system is linear and the state variables have Gaussian statistics. By using the Monte Carlo method as the forecast step and obtaining the statistics from the ensemble, the EnKF relaxes the linearity assumption to a certain extent. However, the update step of the EnKF still depends on the first two statistical moments and thus the EnKF is only suboptimal in the presence of strong nonlinearity and non-Gaussianity. There are other types of filter schemes analyzing the full probability distribution to obtain the maximum likelihood estimate. But owing to its straightforward conceptualization, relatively low computational cost and satisfactory performance, the EnKF has been used in a large number of applications in various fields such as meteorology, oceanography, hydrology, and reservoir engineering (Evensen, 1994, 2003; Mclaughlin, 2002; Nævdal et al., 2005; Wen and Chen, 2005; Liu and Oliver, 2005; Gu and Oliver, 2005; Chen and Zhang, 2006).

The size of the ensemble is crucial for the performance of the EnKF since the standard deviation of the sampling errors of the Monte Carlo method converges slowly with the sample size N (approximately equal to $1/\sqrt{N}$). In general, a small ensemble induces a large sampling error; a large ensemble leads to computational inefficiency. Various methods have been proposed to reduce the sampling error associated with the small-sized ensemble in the EnKF, such as double ensemble Kalman filter (Houtekamer and Mitchell, 1998), ensemble square root filter (Whitaker and Hamill, 2002), ensemble Kalman filter with forgetting factor (Anderson and Anderson, 1999). These methods showed promising results even with relatively small number of ensemble members. On the other hand, these methods are application dependent and some parameters are difficult to quantify. Extra effort may need to tune these parameters for obtaining reasonable results. The representativity of a given sized ensemble is another factor that controls the performance of the EnKF. If the ensemble members ideally sample the main directions of the probability space, the ensemble can better represent the underlying probability distribution, thus leading to better EnKF estimates. Although some initial ensemble selection and ensem-

Copyright © 2007 Society of Petroleum Engineers

This paper (SPE 95277) was revised for publication from paper SPE 95277, presented at the 2005 SPE Annual Technical Conference and Exhibition, Dallas, Texas, 9–12 October. Original manuscript received for review 14 July 2006. Revised manuscript received 14 October 2006. Manuscript peer approved 25 October 2006.

ble resampling schemes have been proposed (Xiu and Karniadakis, 2003; Evensen, 2004; Wen and Chen, 2005), the EnKF is usually carried out with randomly generated ensemble and the performance of the EnKF with a small-sized ensemble may vary with the particular realizations.

For high-resolution, large-scale problems, the dimension of the state is fairly large. One of the major problems of various Kalman filter-based methods is how to efficiently approximate the state error covariance matrix with a dimension-reduced approach in each update. The EnKF is one type of dimension-reduced approaches since the number of ensemble members is in general smaller than the dimension of the state. There are other attempts that estimate the state error covariance matrix by selecting its principal modes, and several of such methods are given next. The reduced rank square root filter (RRSQRT) (Verlaan and Heemink, 1997) is based on the factorization of the error covariance matrix where only the leading eigenvectors are selected in order to reduce the dimension of interest. The singular evolutive Kalman filter (SEEK) (Pham et al., 1998) approximates the state error covariance matrix by a singular lower-rank matrix. In this approach, the correction to the state is only made to the directions of the growing error and the directions of the correction change with time according to the model evolution. Partially orthogonal ensemble Kalman filter (POEnKF) and complementary orthogonal subspace filter for efficient ensemble (COFFEE) (Heemink et al., 2001) combine the RRSQRT and the EnKF together to obtain variance reduced estimation. RRSQRT is used as a variance reducer containing the information of the leading eigenvalues while the remaining state is represented by the EnKF.

In this work, a new type of dimension-reduced data assimilation method is proposed based on Karhunen-Loeve (KL) expansion and orthogonal polynomial decompositions, which will be referred to as Karhunen-Loeve-based Kalman filter (KLKF) in this paper. The major difference between KLKF and other dimension-reduced methods mentioned previously is that the approximation to the high-dimension covariance matrix is not performed directly as a matrix manipulation. The KL expansion is performed on the reservoir properties, to obtain the principle modes of the major source of the uncertainty, based on which an efficient KL-based moment-equation (KLME) approach is used to solve the stochastic flow equations (Zhang and Lu, 2003; Lu and Zhang, 2004, 2006). In order to construct the state error covariance, the covariance of reservoir properties is efficiently approximated by a small set of eigenvalues and eigenfunctions attributed to the mean square convergence of the KL decomposition; the covariance of reservoir responses is approximated by the coefficients of the polynomial expansions, the cross-covariance of the reservoir properties and responses is approximated by the eigenvalues, eigenfunctions and the coefficients of the polynomial expansions. In the later sections detailed formulations will be given on the construction of the covariance matrices. Compared to the full covariance matrix, approximating the covariance based on a finite number of modes represents a significant reduction in the random dimensions.

In KLKF, the forecast step is completed with the KLME method, from which the mean and covariance of the state variables can be constructed, when needed. The statistics of both the reservoir properties and reservoir responses are then updated with the observations available at the measurement time using the auto- and cross-covariance obtained from the forecast step. A synthetic 2D example is used to demonstrate the capability of this new method and to compare the results with those from the traditional EnKF method. Our results show that the Karhunen-Loeve-based Kalman filter (KLKF) is capable of significantly reducing the required computational resources with satisfactory accuracy, which indicates the potential applicability of this approach to high-resolution, large-scale predictive models.

Statement of Problem

We consider transient fluid flow in heterogeneous reservoirs satisfying the following governing equation:

$$\nabla \cdot [K_s(\mathbf{x})\nabla h(\mathbf{x},t)] + g(\mathbf{x},t) = S_s \frac{\partial h(\mathbf{x},t)}{\partial t}, \quad (1)$$

subject to the initial and boundary conditions:

$$h(\mathbf{x},0) = H_0(\mathbf{x}),\ \mathbf{x} \in D, \quad (2)$$

$$h(\mathbf{x},t) = H(\mathbf{x},t),\ \mathbf{x} \in \Gamma_D, \quad (3)$$

$$K_s(\mathbf{x})\nabla h(\mathbf{x},t) \cdot \mathbf{n}(\mathbf{x}) = -Q(\mathbf{x},t),\ \mathbf{x} \in \Gamma_N, \quad (4)$$

where $g(\mathbf{x},t)$ is the source/sink term, $h(\mathbf{x},t) = P(\mathbf{x})/(\rho g) + x_3$ is the pressure head (P being pressure, ρ density, g gravitational acceleration factor and x_3 elevation), $H_0(\mathbf{x})$ is the initial head in the domain D, $H(\mathbf{x},t)$ is the prescribed head on Dirichlet boundary segments Γ_D, $K_s(\mathbf{x}) = k(\mathbf{x})\rho g/\mu$ is the hydraulic conductivity (k is the intrinsic or absolute permeability and μ is the fluid viscosity), $Q(\mathbf{x},t)$ is the prescribed flux across Neumann boundary segments Γ_N, $\mathbf{n}(\mathbf{x})$ is an outward vector normal to the boundary Γ_N, and S_s is the specific storage. In this study, the hydraulic conductivity $K_s(\mathbf{x})$ is considered as a random space function, while specific storage S_s is treated as a deterministic constant. We usually work with log transformed hydraulic conductivity $Y = \ln K_s$. Note that the (co)variance of the log hydraulic conductivity is the same as that of the log absolute permeability $\ln k$ because $Y(\mathbf{x}) = \ln[\rho g/\mu] + \ln[k(\mathbf{x})]$ with ρ, g and μ being known constants.

Since $K_s(\mathbf{x})$ is a random function, the flow equations become stochastic partial differential equations, which can be solved with various approaches, such as the moment-equation method (Zhang, 2002) and Monte Carlo simulation. In this study the KL-based moment-equation (KLME) method is used in the KLKF. A brief description of the KLME will be given later. As a comparison, the Monte Carlo method is used in the EnKF.

The hydraulic conductivity $K_s(\mathbf{x})$ is assumed to follow a log normal distribution. We work with the log transformed variable $Y(\mathbf{x})$, given as:

$$Y(\mathbf{x}) = \ln[K_s(\mathbf{x})] = Y^{(0)}(\mathbf{x}) + Y'(\mathbf{x}), \quad (5)$$

where $Y^{(0)}(\mathbf{x})$ is the mean, representing a relatively smooth unbiased estimate of the unknown random function Y, and $Y'(\mathbf{x})$ is the zero-mean fluctuation.

Suppose we have N_Y direct measurements of the log hydraulic conductivity $Y_1, Y_2, \cdots, Y_{N_Y}$ taken at locations $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_Y}$ and N_h pressure head measurements $h_1, h_2, \cdots, h_{N_h}$ located at $\chi_1, \chi_2, \cdots, \chi_{N_h}$, both measured at some time intervals. The aim of this study is to provide an efficient algorithm to characterize reservoir properties by dynamically incorporating these measurements when they become available. During the data assimilation process the stochastic differential flow equation is solved forward with time; the mean and the fluctuation of hydraulic conductivity are modified together with the pressure head to honor the observations at various times sequentially, for obtaining up-to-date estimates of hydraulic conductivity field and the corresponding uncertainty.

Karhunen-Loeve Decomposition

For the Gaussian stochastic process $Y(\mathbf{x},\omega) = \ln[K_s(\mathbf{x},\omega)]$, where $\mathbf{x} \in D$ and $\omega \in \Omega$ (a probability space), the covariance function $C_Y(\mathbf{x},\mathbf{y}) = < Y'(\mathbf{x},\omega)Y'(\mathbf{y},\omega) >$ is bounded, symmetric, and positive definite, thus it can be decomposed as:

$$C_Y(\mathbf{x},\mathbf{y}) = \sum_{m=1}^{\infty} \lambda_m f_m(\mathbf{x}) f_m(\mathbf{y}), \quad (6)$$

where m is the index of modes, λ_m and f_m are eigenvalues and orthogonal deterministic eigenfunctions, respectively. λ_m are sorted in the descending order and the corresponding eigenfunctions f_m

exhibit a decreasing characteristic scale as the index m increases. The mean removed stochastic process $Y'(\mathbf{x},\omega)$ can be expanded with the KL expansion as:

$$Y'(\mathbf{x},\omega) = \sum_{m=1}^{\infty} \xi_m(\omega)\sqrt{\lambda_m} f_m(\mathbf{x}), \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$$

where ξ_m are orthogonal standard Gaussian random variables, i.e., $<\xi_m>=0$ and $<\xi_i\xi_j>=\delta_{ij}$, where δ_{ij} is the Kronecker delta function. It has been shown that the KL expansion is of mean square convergence and may be well approximated with a finite summation. In our study, the KL expansion is truncated up to the first M terms (M modes). The number of modes required for accurately approximating $Y'(\mathbf{x},\omega)$ depends on the ratio of the correlation length to the dimension of the domain (Zhang and Lu, 2003).

KL-based Moment Equations

We assume that the pressure head can be formally expressed as an infinite series: $h(\mathbf{x},t) = h^{(0)} + h^{(1)} + \cdots$, where the order of each term is with respect to σ_Y, the standard deviation of Y. Only the first two terms $h^{(0)}$ and $h^{(1)}$ are used in this work, thus the KLKF method proposed here is based on the first-order approximation of the pressure head. By substituting the expansions of $h(\mathbf{x},t)$ and Y into Eq. 1, we obtain the governing equations for zeroth- and first-order pressure head terms $h^{(0)}$ and $h^{(1)}$ as well as their corresponding boundary and initial conditions (Zhang, 2002; Zhang and Lu, 2003). We further assume that the first-order head term can be expressed as a polynomial expansion in terms of the orthogonal Gaussian random variable $\xi_m, m = 1,2,\cdots,M$:

$$h^{(1)}(\mathbf{x},t) = \sum_{m=1}^{M} \xi_m h_m^{(1)}(\mathbf{x},t). \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (8)$$

After substituting this expansion and the KL decomposition of the log hydraulic conductivity, i.e., Eq. 7, into equations for the first-order term $h^{(1)}$ and recalling the orthogonality of the Gaussian random variables $\xi_m, m = 1,2,\cdots,M$, we obtain M sets of deterministic governing equations for the coefficients $h_m^{(1)}(\mathbf{x},t)$. It is worthwhile to point out that the zeroth-order head equation and the M sets of equation for the coefficients $h_m^{(1)}(\mathbf{x},t)$ (in total, $M+1$ sets of equation) have the exactly same structure as the original flow equation (Lu and Zhang, 2006). By changing the input parameters, the KL-based moment equations can be solved easily with existing codes.

Once the coefficients $h_m^{(1)}(\mathbf{x},t)$ are solved, the head covariance as well as the cross-covariance between the head and the log hydraulic conductivity, required in the data assimilation process, can be approximated up to first order in terms of σ_Y^2 as:

$$C_{Yh}(\mathbf{x},\mathbf{y},t) = \sum_{m=1}^{M} \sqrt{\lambda_m} f_m(\mathbf{x}) h_m^{(1)}(\mathbf{y},t), \quad \ldots\ldots\ldots\ldots \quad (9)$$

$$C_h(\mathbf{x},t;\mathbf{y},\tau) = \sum_{m=1}^{M} h_m^{(1)}(\mathbf{x},t) h_m^{(1)}(\mathbf{y},\tau). \quad \ldots\ldots\ldots\ldots \quad (10)$$

Data Assimilation Methodology

Data assimilation is a process consisting of a series of update steps, each of which represents the time when observations become available and/or the updating process is operated. At each update step the current system state is combined with the new observations for the sake of obtaining a minimum variance estimation. In this section, the time symbol t is suppressed, since the discussion is based on any fixed update step. Similar to the EnKF, the KLKF requires to compute (some subsets of) the covariance functions C_Y, C_{Yh}, and C_h, as presented in Eqs. 6, 9, and 10. This implies that the eigenvalues λ_m, the eigenfunctions f_m and the first-order head coefficient $h_m^{(1)}$ have to be updated at each update step together with the mean hydraulic conductivity field $Y^{(0)}$ and the zeroth-order head $h^{(0)}$. Certainly, at each update step, the covariance function of the log hydraulic conductivity is nonstationary and its eigenvalues and eigenfunctions have to be solved numerically. It is well-known that solving the eigenvalue problem is computationally demanding for large-scale problems. For this reason, we developed an algorithm that requires solving the eigenvalue problem at the first time step and then efficiently updating the eigenvalues and eigenfunctions at the subsequent update steps.

Updating Mean Hydraulic Conductivity Field. At the update step, the updated mean and covariance of Y upon incorporating new observations can be derived from the cokriging technique:

$$\begin{aligned} {}^{(c)}Y^{(0)}(\mathbf{x}) &= Y^{(0)}(\mathbf{x}) + \sum_{i=1}^{N_Y} \alpha_i(\mathbf{x})[Y_{\text{Obs}}(\mathbf{x}_i) - Y^{(0)}(\mathbf{x}_i)] \\ &\quad + \sum_{i=1}^{N_h} \beta_i(\mathbf{x})[h_{\text{Obs}}(\chi_i) - h^{(0)}(\chi_i)], \quad \ldots\ldots\ldots \quad (11) \end{aligned}$$

$$\begin{aligned} {}^{(c)}C_Y(\mathbf{x}_i,\mathbf{x}_j) &= C_Y(\mathbf{x}_i,\mathbf{x}_j) - \sum_{n=1}^{N_Y} \alpha_n(\mathbf{x}_i) C_Y(\mathbf{x}_n,\mathbf{x}_j) \\ &\quad - \sum_{n=1}^{N_h} \beta_n(\mathbf{x}_i) C_{Yh}(\mathbf{x}_j,\chi_n) \\ &\quad i = 1,2,\ldots N_Y,\ j = 1,2,\ldots N_Y, \quad \ldots\ldots \quad (12) \end{aligned}$$

where the quantities with (without) superscript (c) stands for the values after (before) incorporating observations at this time step, and α_i and β_i are weighting functions, representing the relative importance of each measurement $Y(\mathbf{x}_i)$ and $h(\chi_i)$ in predicting the value of ${}^{(c)}Y^{(0)}(\mathbf{x})$ at location $\mathbf{x}$. The weighting functions are solutions of the following cokriging equations:

$$\sum_{i=1}^{N_Y} \alpha_i(\mathbf{x}) C_Y(\mathbf{x}_i,\mathbf{x}_j) + \sum_{i=1}^{N_h} \beta_i(\mathbf{x}) C_{Yh}(\mathbf{x}_j,\chi_i) = C_Y(\mathbf{x},\mathbf{x}_j) \quad j = 1,2,\ldots N_Y, \quad \ldots\ldots \quad (13)$$

$$\sum_{i=1}^{N_Y} \alpha_i(\mathbf{x}) C_{Yh}(\mathbf{x}_i,\chi_j) + \sum_{i=1}^{N_h} \beta_i(\mathbf{x}) C_h(\chi_i,\chi_j) = C_{Yh}(\mathbf{x},\chi_j) \quad j = 1,2,\ldots N_h. \quad \ldots\ldots \quad (14)$$

The observations $Y_{\text{Obs}}(\mathbf{x}_i)$ and $h_{\text{Obs}}(\chi_i)$ may include noises:

$$Y_{\text{Obs}}(\mathbf{x}_i) = Y^t(\mathbf{x}_i) + \zeta_i \sigma_{YN},\ i = 1,2,\ldots N_Y, \quad \ldots\ldots\ldots \quad (15)$$

$$h_{\text{Obs}}(\chi_i) = h^t(\chi_i) + \zeta_i \sigma_{hN},\ i = 1,2,\ldots N_h, \quad \ldots\ldots\ldots\ldots \quad (16)$$

where $Y^t(\mathbf{x}_i)$ and $h^t(\chi_i)$ are (unknown) true values, ζ_i are Gaussian random variables with zero mean and unit variance, and σ_{YN} and σ_{hN} are the standard deviations of measurements errors of the log hydraulic conductivity $Y(\mathbf{x})$ and pressure head $h(\chi)$, which are assumed to be known.

Since the set of eigenfunctions is complete, α_i and β_i can be expanded on the basis of these eigenfunctions:

$$\alpha_i(\mathbf{x}) = \sum_{m=1}^{\infty} \alpha_{im} f_m(\mathbf{x}),\ i = 1,2,\ldots N_Y, \quad \ldots\ldots\ldots\ldots \quad (17)$$

$$\beta_i(\mathbf{x}) = \sum_{m=1}^{\infty} \beta_{im} f_m(\mathbf{x}),\ i = 1,2,\ldots N_h. \quad \ldots\ldots\ldots\ldots\ldots \quad (18)$$

The equations for α_{im} and β_{im} can be derived from Eqs. 13 and 14 with Eqs. 17 and 18 by multiplying $f_m(\mathbf{x})$ on the both sides, and

integrating the equations with respect to $\mathbf{x}$ over the domain D:

$$\sum_{i=1}^{N_Y} \alpha_{im} C_Y(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{N_h} \beta_{im} C_{Yh}(\mathbf{x}_j, \chi_i) = \lambda_m f_m(\mathbf{x}_j) \ , \ . \ . \quad (19)$$

$$\sum_{i=1}^{N_Y} \alpha_{im} C_{Yh}(\mathbf{x}_i, \chi_j) + \sum_{i=1}^{N_h} \beta_{im} C_h(\chi_i, \chi_j) = \sqrt{\lambda_m} h_m^{(1)}(\chi_j) \ , \quad (20)$$

where the cross-covariance $C_{Yh}(\mathbf{x}_i, \chi_j)$, auto-covariance of pressure head $C_h(\chi_i, \chi_j)$ and auto-covariance of log hydraulic conductivity $C_Y(\mathbf{x}_i, \mathbf{x}_j)$ are given by Eqs. 9, 10 and 12, respectively.

Updating Pressure Head Fields. The zeroth-order pressure head can be updated with both types of observations in the same manner:

$$\begin{aligned} {}^{(c)}h^{(0)}(\mathbf{x}) &= h^{(0)}(\mathbf{x}) + \sum_{i=1}^{N_Y} \mu_i(\mathbf{x}) [Y_{\text{Obs}}(\mathbf{x}_i) - Y^{(0)}(\mathbf{x}_i)] \\ &+ \sum_{i=1}^{N_h} \eta_i(\mathbf{x}) [h_{\text{Obs}}(\chi_i) - h^{(0)}(\chi_i)] \ , \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad (21) \end{aligned}$$

where μ_i and η_i are subject to:

$$\sum_{i=1}^{N_Y} \mu_i(\mathbf{x}) C_Y(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{N_h} \eta_i(\mathbf{x}) C_{Yh}(\mathbf{x}_j, \chi_i) = C_{Yh}(\mathbf{x}_j, \mathbf{x}) \ , \quad (22)$$

$$\sum_{i=1}^{N_Y} \mu_i(\mathbf{x}) C_{Yh}(\mathbf{x}_i, \chi_j) + \sum_{i=1}^{N_h} \eta_i(\mathbf{x}) C_h(\chi_i, \chi_j) = C_h(\mathbf{x}, \chi_j) \ . \quad (23)$$

α_{im}, β_{im}, $\mu_i(\mathbf{x})$ and $\eta_i(\mathbf{x})$ in Eqs. 19, 20, 22 and 23 can be computed through solving a set of linear algebraic equations with the same coefficient matrix. In this work, LU decomposition is used to solve these equations, the decomposition needs to be done only once, and all the coefficients can be obtained efficiently by changing the right-hand side vector. However, the condition number of this coefficient matrix can be extremely large, especially after several update steps, since assimilating observations results in a great reduction of (co)variance. In the presence of the ill-conditioned matrix the truncation error may be easily amplified and result in instability. Dietrich and Newsam (1989) analyzed the cause of the ill-conditioning and proposed that adding a relaxation term ε (or an explicit error matrix) to the matrix can resolve this problem. In general, a relatively large relaxation term improves the stability with the price of losing information and slowing down the rate of convergence, while a small value may lead to numerical instability. The error matrix can be obtained through a maximum likelihood approach for the purpose of improving the conditioning and minimizing the loss of information. Yeh et al. (1996) added a relaxation term to the diagonal components of the matrix to reduce the condition number of the matrix and assigned the relaxation term as a fraction of the maximum value of the matrix. In this paper, a constant relaxation term is added to the diagonal components of $C_h(\chi_i, \chi_j)$, which will be discussed further with illustrative examples.

Similar to the zeroth-order head term, the first-order pressure head is updated by:

$$\begin{aligned} {}^{(c)}h^{(1)}(\mathbf{x}) &= h^{(1)}(\mathbf{x}) - \sum_{i=1}^{N_Y} \mu_i(\mathbf{x}) Y'(\mathbf{x}_i) \\ &- \sum_{i=1}^{N_h} \eta_i(\mathbf{x}) h^{(1)}(\chi_i) \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad (24) \end{aligned}$$

Because both $Y'(\mathbf{x})$ and $h^{(1)}(\mathbf{x})$ can be expanded based on ξ_m, by substituting Eqs. 7 and 8 into Eq. 24, instead of updating $h^{(1)}(\mathbf{x})$ directly, the coefficients $h_m^{(1)}(\mathbf{x})$ can be updated as:

$$\begin{aligned} {}^{(c)}h_m^{(1)}(\mathbf{x}) &= h_m^{(1)}(\mathbf{x}) - \sum_{i=1}^{N_Y} \mu_i(\mathbf{x}) \sqrt{\lambda_m} f_m(\mathbf{x}_i) \\ &- \sum_{i=1}^{N_h} \eta_i(\mathbf{x}) h_m^{(1)}(\chi_i) \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad (25) \end{aligned}$$

It can be shown that to the first order, Eq. 25 recovers the usual cokriging equation for the head convariance similar to Eq. 12.

Updating Eigenvalues and Eigenfunctions. Because of the nonstationarity of the covariance matrix given in Eq. 12, the conditional eigenvalues and eigenfunctions have to be solved numerically. Here we follow the method in Lu and Zhang (2004) with some modification to incorporate the influence of the pressure head measurements.

By definition, the eigenvalues ${}^{(c)}\lambda_m$ and their corresponding eigenfunctions ${}^{(c)}f_m(\mathbf{x})$ can be solved from the following Fredholm equation of the second kind:

$$\int_{\Omega} {}^{(c)}C_Y(\mathbf{x}, \mathbf{y}) {}^{(c)}f(\mathbf{x}) d\mathbf{x} = {}^{(c)}\lambda {}^{(c)}f(\mathbf{y}) \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad (26)$$

Since the set of eigenfunctions $f_m(\mathbf{x})$ computed at the previous update step is complete, the eigenfunctions ${}^{(c)}f_m(\mathbf{x})$ at the current step can be expanded in terms of $f_m(\mathbf{x})$ as:

$${}^{(c)}f_m(\mathbf{x}) = \sum_{p=1}^{M} d_{mp} f_p(\mathbf{x}), \ m = 1, 2, \cdots, M \ , \ . \ . \ . \ . \ . \ . \ . \ . \quad (27)$$

where the coefficient matrix $\mathbf{D} = (d_{mp})_{M \times M}$ is to be determined. Substituting this expression and Eq. 12 into Eq. 26, multiplying $f_m(\mathbf{y})$ on the both sides of the derived equation, and integrating it with respect to $\mathbf{y}$ over domain D, yields:

$$\begin{aligned} \lambda_m d_m &- \sum_{k=1}^{M} \left(\sum_{i=1}^{N_Y} \alpha_{ik} \lambda_m f_m(\mathbf{x}_i) + \sum_{i=1}^{N_h} \beta_{ik} \sqrt{\lambda_m} h_m^{(1)}(\chi_i) \right) d_k \\ &= {}^{(c)}\lambda_m d_m, \quad m = 1, 2, \cdots, M \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad (28) \end{aligned}$$

It can also be expressed in a succinct matrix form as:

$$(\mathbf{A} - {}^{(c)}\lambda \mathbf{I}) = 0 \ , \ . \quad (29)$$

where

$$a_{km} = \lambda_m \delta_{km} - \left(\sum_{i=1}^{N_Y} \alpha_{ik} \lambda_m f_m(\mathbf{x}_i) + \sum_{i=1}^{N_h} \beta_{ik} \sqrt{\lambda_m} h_m^{(1)}(\chi_i) \right)$$

is the components of $\mathbf{A} = (a_{km})_{M \times M}$, ${}^{(c)}\lambda = \text{diag}\left[{}^{(c)}\lambda_1, \cdots, {}^{(c)}\lambda_M\right]$, and $\mathbf{I}$ is an $M \times M$ identity matrix. Therefore, the problem of finding the eigenvalues and eigenfunctions of a nonstationary covariance matrix ${}^{(c)}C_Y(\mathbf{x}, \mathbf{y})$ of size $N \times N$ reduces to the problem of finding the eigenvalues ${}^{(c)}\lambda$ and eigenvectors $\mathbf{d}$ of an $M \times M$ matrix for M times, where N is the number of gridblocks and M is the number of modes. Note that the number of gridblocks N is usually much larger than the number of modes M. Conditional eigenvector ${}^{(c)}f_m(\mathbf{x})$ corresponding to each conditional eigenvalue ${}^{(c)}\lambda_m$ can be constructed through d_{mp} in Eq. 27. The updated eigenvalues and eigenvectors as well as the updated pressure head terms in Eqs. 21 and 25 are fed into the KLME forward model for the next forecast step.

Since directly solving the Fredholm equation is computationally expensive, the proposed algorithm for updating the conditional eigenvalues and eigenfunctions at each update step has a significant advantage, because we only need to solve the Fredholm equation for a stationary covariance function at the first timestep rather than at each subsequent assimilation step for a nonstationary covariance function.

Karhunen-Loeve-Based Kalman Filter. The methodology described in the previous section has a different format from the traditional Kalman filter, where the Kalman gain and the update step are formulated as:

$$\mathbf{K} = \mathbf{P}^f \mathbf{H}^T [\mathbf{H}\mathbf{P}^f \mathbf{H}^T + \mathbf{R}]^{-1}, \quad (30)$$

$$\mathbf{S}^u = \mathbf{S}^f + \mathbf{K}[\mathbf{d}_{\text{Obs}} - \mathbf{H}\mathbf{S}^f], \quad (31)$$

$$\mathbf{P}^a = [\mathbf{I} - \mathbf{K}]\mathbf{P}^f \quad (32)$$

where $\mathbf{K}$ is the Kalman gain, $\mathbf{H}$ is the observation operator, $\mathbf{R}$ is the observation error covariance matrix, $\mathbf{d}_{\text{Obs}}$ is the observation vector, $\mathbf{S}^f$ is the forecast state vector, $\mathbf{S}^u$ is the updated state vector, and $\mathbf{P}^f$ is the covariance matrix of the forecast state vector. In fact, the cokriging based updating scheme is exactly the same as the traditional Kalman filter at each step. If we define the state vector as $\mathbf{S} = [Y(\mathbf{x}_1), \cdots, Y(\mathbf{x}_N), h(\mathbf{x}_1), \cdots, h(\mathbf{x}_N)]$, the corresponding Kalman gain can be constructed with the coefficients $\alpha_i(\mathbf{x})$, $\beta_i(\mathbf{x})$, $\mu_i(\mathbf{x})$ and $\eta_i(\mathbf{x})$ described before:

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_\alpha & \mathbf{K}_\beta \\ \mathbf{K}_\mu & \mathbf{K}_\eta \end{bmatrix}_{2\times N, N_Y+N_h},$$

$$\begin{aligned} \mathbf{K}_\alpha &= [\alpha_1(\mathbf{x}),\ \alpha_2(\mathbf{x}),\ \cdots\ \alpha_{N_Y}(\mathbf{x})] \\ \mathbf{K}_\beta &= [\beta_1(\mathbf{x}),\ \beta_2(\mathbf{x}),\ \cdots\ \beta_{N_h}(\mathbf{x})] \\ \mathbf{K}_\mu &= [\mu_1(\mathbf{x}),\ \mu_2(\mathbf{x}),\ \cdots\ \mu_{N_Y}(\mathbf{x})] \\ \mathbf{K}_\eta &= [\eta_1(\mathbf{x}),\ \eta_2(\mathbf{x}),\ \cdots\ \eta_{N_h}(\mathbf{x})], \end{aligned} \quad (33)$$

where N is total number of gridblocks and N_Y and N_h are respectively the number of log hydraulic conductivity and pressure head measurements.

The major difference between the KLKF approach and the cokriging method is that the KLKF is a sequential or on-line method, while cokriging is a statistical interpolation method, which only operates at a fixed time. Sometimes cokriging is performed iteratively to account for the possible nonlinear effects in order to obtain a reasonable estimation, but the iteration is still based on a certain time (Yeh et al., 1996). The KLKF uses the KLME method for advancing the system with time and incorporates the observations at the time when they become available to adjust the model state. After the current update step, the updated reservoir responses are taken as the initial conditions for the next forecast step. The updated model is then run until the next set of observations becomes available, at which the update step will be performed again.

Illustrative Examples

In this section, we use a synthetic 2D example to demonstrate the applicability of the KLKF method for estimating the hydraulic conductivity field by assimilating pressure head and hydraulic conductivity measurements. The results are compared with those of the EnKF method in terms of both computational cost and accuracy.

The flow domain is a square of size $L_x = L_y = 800$ m, uniformly discretized into 40×40 square elements, as shown in **Fig. 1**. Starting at $t = 0$ day, a pumping well with a volumetric flow rate of 150 m^3/day is placed at (260 m, 240 m) and an injection well with the same flow rate is placed at (560 m, 540 m). These two wells are active throughout the whole study period with constant flow rates. The two lateral boundaries are no-flow boundaries, while the left and the right are Dirichlet boundaries with prescribed pressure head of 202 m and 198 m, respectively. Storage coefficient is assumed to be a constant taken as 0.0001.

The log hydraulic conductivity field is treated as a Gaussian random function with zero mean and unit variance. Thus, the geometric mean of the saturated hydraulic conductivity is $K_G = 1.0$ m/day. The unconditional hydraulic conductivity field is also assumed to be second-order stationary characterized by the separable exponential covariance function, which is defined as

$$C_Y(\mathbf{h}) = \sigma_Y^2 \exp\left[-\frac{|h_x|}{\lambda_x} - \frac{|h_y|}{\lambda_y}\right], \quad (34)$$

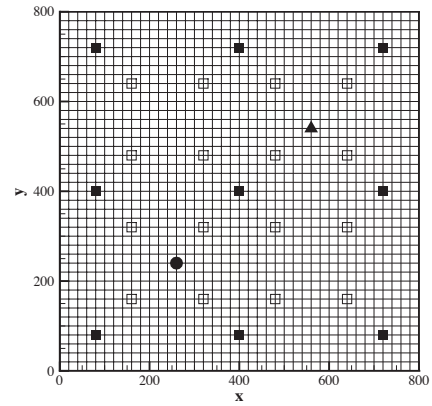

Fig. 1—The flow domain and the observation locations of $\ln K_s$ (the nine filled squares) and the pressure head (all the 25 squares). The circle and triangle are the pumping and injection well, respectively.

where correlation lengths $\lambda_x = 200$ m, $\lambda_y = 100$ m, variance $\sigma_Y^2 = 1.0$, and $\mathbf{h} = (h_x, h_y)^T$ is the separation vector.

In this synthetic example, we generate an unconditional log hydraulic conductivity field using the aforementioned statistics and take this field as the reference field (see **Fig. 2**). We then take nine samples from this reference field at selected locations as shown in **Fig. 1** (filled squares) and consider these samples as direct measurements of the log hydraulic conductivity field. We run a forward simulation using the reference field and take 25 pressure head measurements at selected locations (all the well locations in **Fig. 1**) with specified time intervals. The log hydraulic conductivity measurements are assumed to be perfect. While the pressure head observations are noisy and the error follows a normal distribution with mean equal to zero and standard deviation equal to 0.05 m.

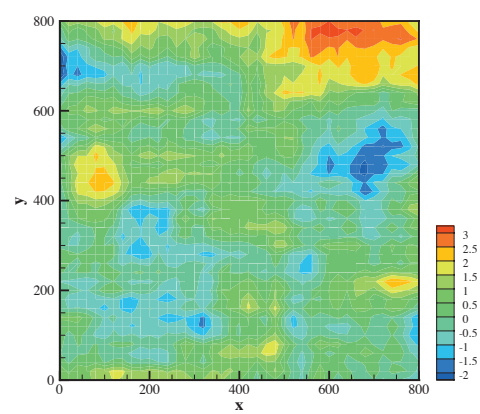

Fig. 2—The reference $\ln K_s$ field.

For this model setup, the fluid flow reaches steady state at about $t = 10$ day. This period is chosen as the simulation duration, which is subdivided into 50 equally sized timesteps of 0.2 day interval. The pressure head measurements are available at an equally spanned period with $\Delta t = 0.6$ day. The hydraulic conductivity measurements are assimilated with the first set of pressure head measurements at $t = 0.2$ day, and after that only pressure head measurements are assimilated every 0.6 day up to 10 day. For both the EnKF and the KLKF methods, the initial pressure head is assumed to be known without uncertainty.

Karhunen-Loeve-Based Kalman Filter. The KLKF is initialized with statistics slightly different from the reference. The mean of the log hydraulic conductivity is still zero while the variance is set to be 1.2. Separable exponential covariance function was used with correlation lengths $\lambda_x = 220$ m, $\lambda_y = 120$m.

50, 100 and 200 modes has been used for testing the KLKF. Before showing results obtained from KLKF with different number of modes involved, it is of interest to examine how much energy is carried by these leading modes. **Fig. 3** shows the normalized eigenvalues (λ_m/D, where D is the size of the domain) with respect to the indices of the modes and the fraction of energy contained by the cumulative eigenvalues ($\sum_m \lambda_m/(D\sigma_Y^2)$) as a function of the modes included. The magnitude of eigenvalues decreases with the increase of the index of the modes, and by using the first 100 modes, we can capture about 85% energy of the system of correlation length to domain size ratios $\lambda_x/L_x = 220/800$ and $\lambda_y/L_y = 120/800$. As a general rule, the number of modes needed decreases as such ratios increase.

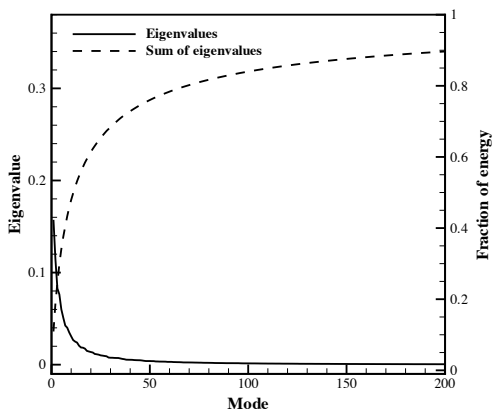

Fig. 3—The normalized eigenvalues and their normalized accumulative values.

Two commonly used metrics are used to quantify the goodness of the estimates. The root mean square error (RMSE) shows the deviation of the estimated mean Y field from the reference field,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[Y^*(\mathbf{x}_i) - Y^t(\mathbf{x}_i)]^2} \quad \ldots\ldots\ldots\ldots\ldots \quad (35)$$

where $Y^*(\mathbf{x}_i)$ stands for the estimated mean values and $Y^t(\mathbf{x}_i)$ stands for the reference values.

The SPREAD is the estimated uncertainty represented by the ensemble, which can be defined as:

$$\mathrm{SPREAD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathrm{VAR}_{\mathrm{en}}(\mathbf{x}_i)} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (36)$$

where $\mathrm{VAR}_{\mathrm{en}}(\mathbf{x}_i)$ is the ensemble variance for point $\mathbf{x}_i$. The estimated uncertainty of the KLKF can be computed similarly with the calculated variance instead of the ensemble variance.

Fig. 4 and **Fig. 5** show the RMSE and SPREAD of the KLKF with 50, 100, and 200 modes involved, where a relaxation term $\varepsilon = 0.3$ is used for all these three cases. It is seen that overall the RMSE decreases as more observations are incorporated in time while the RMSE for KLKF with 50 modes exhibits some oscillations. The KLKF with 200 modes gives the lowest RMSE since with more modes included the hydraulic conductivity field can be better characterized. The RMSE quantifies the deviation of the estimation from the reference (Eq. 35) and the SPREAD gives the estimated variation of the estimation (Eq. 36). For a good estimate, the SPREAD should be similar to the RMSE. From **Fig. 4** and **Fig. 5**, it can be seen that the SPRAD systematically underestimates the RMSE. It is worthwhile to note that although the KLKF with 100 and 200 modes gives similar RMSE, the KLKF with 200 modes shows a better estimation of the uncertainty.

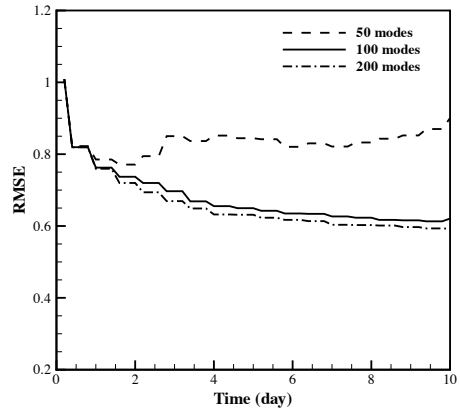

Fig. 4—The RMSE for the KLKF with different number of modes used.

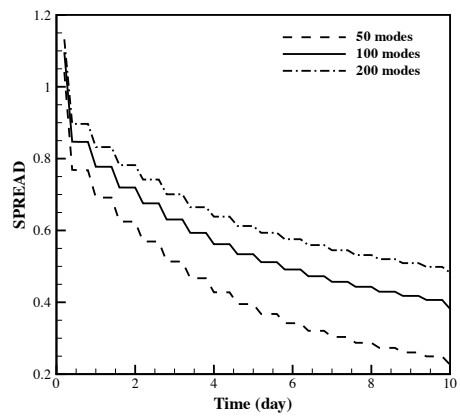

Fig. 5—The SPREAD for the KLKF with different number of modes used.

Fig. 6 compares the reference field and the estimated mean log hydraulic conductivity field from the KLKF with 100 modes at several different times. It shows that with time (and available observations), the estimated field becomes closer to the reference field.

The relaxation term ε used in the KLKF reduces the condition number of the coefficient matrix in Eqs. 19, 20, 22 and 23, and therefore improves stability of the KLKF solution. Our numerical experiments show that the performance of the KLKF is sensitive to the choice of relaxation term when the number of modes is small, for instance, being 50. However, reasonable results can be achieved by cautiously choosing the relaxation term for the KLKF with 50 modes. With larger number of modes, the KLKF has satisfactory performance as long as the relaxation term is changing within a reasonable range. Results of sensitivity runs on the basis of 100 modes are shown in **Table 1**. In our approach, the log hydraulic conductivity measurements are used only at the first assimilation step, and the relaxation term is only added to the pressure head covariance part. If the direct measurements are assimilated with the pressure

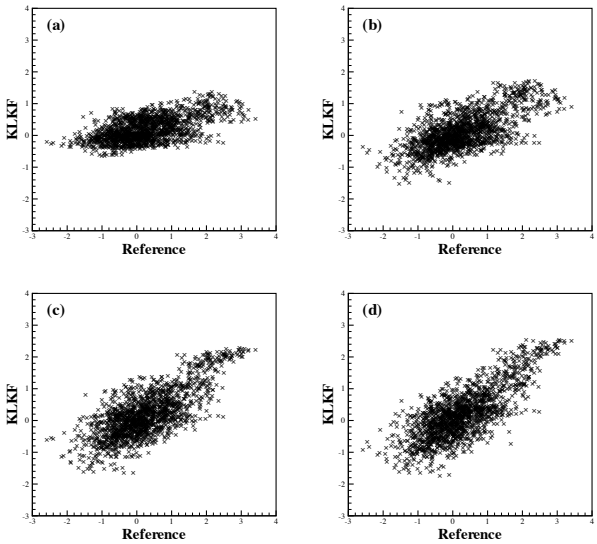

Fig. 6—Comparison between the estimated mean $\ln K_s$ from the KLKF with 100 modes and the reference at different times: (a) 0.2 day, (b) 2.0 day, (c) 5.0 day and (d) 10.0 day.

TABLE 1—RMSE AT DAY 10 FOR KLKF WITH 100 MODES							
ε	0.01	0.05	0.2	0.3	0.4	0.5	1.0
RMSE	0.632	0.621	0.608	0.621	0.653	0.633	0.676

head observations at every assimilation step, it is expected that relaxation terms need to be added to the diagonal components of both log hydraulic conductivity and pressure head covariance matrices.

Two sensitivity runs have also been done for testing the performance of the KLKF in the presence of the incorrect initial statistics. As mentioned in the beginning of this section, all the results shown so far are obtained with initial statistics slightly different from the reference which showed a minor impact on the results. In the following two cases, we further weaken the accuracy of the initial statistics. In the first case the initial mean is set to be -1.5 instead of zero; in the second case the spatial correlation structure is reversed with $\lambda_x = 120$ m and $\lambda_y = 220$ m. All the other statistics for these two cases are the same as the previous setting and the KLKF with 100 modes is used for these sensitivity studies. The RMSE evolution of these two cases is shown in **Fig. 7**. With the incorrect prior mean the RMSE is large initially, but it drops about 50% with only the first set of observation assimilated. For both of these two cases KLKF is able to reduce the RMSE to a certain extent, while the SPREAD still underestimates the RMSE.

Ensemble Kalman Filter. One thousand unconditional realizations of the log hydraulic conductivity field are generated by Karhunen-Loeve expansion (Zhang and Lu, 2003) for performing the EnKF. The statistics for generating the initial ensemble is chosen to be the same as the initial statistics for the KLKF, which is given in the beginning of the last section.

Since the performance of the EnKF is realization dependent, on the basis of the generated 1000 realizations we run 10 sets of the EnKF with 100 ensemble members, 5 sets of the EnKF with 200 ensemble members and one run with all the 1000 realizations. For the sake of comparison, the same relaxation term 0.3 is added to the corresponding part of the matrix when performing the EnKF. **Fig. 8** and **Fig. 9** show the RMSE and the SPREAD for the EnKF runs. The RMSE shows a large variation among different runs with the same number of realizations used while the SPREAD has similar behavior for different runs. As in the KLKF, the SPREAD underestimates the RMSE systematically. For instance, some runs of the EnKF with 100 realizations exhibit oscillations after the first several assimilation steps, while the ensemble spread keeps decreasing, indicating that the ensemble is converging to a wrong solution. The EnKF with 1000 realizations (see **Fig. 10**) shows a satisfactory match between the RMSE and the SPREAD. This implies that the mean field can be approximated with a small set of realizations or modes while for precisely estimating the associated uncertainty more realizations or modes need to be considered.

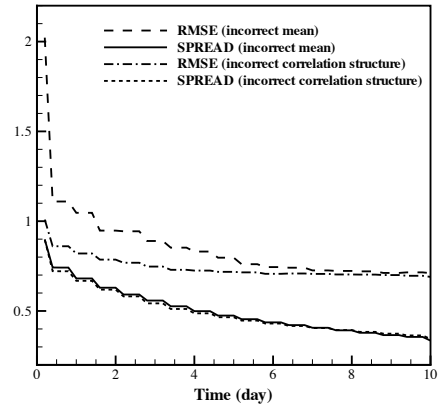

Fig. 7—RMSE and SPREAD of KLKF with 100 modes in the presence of incorrect initial statistics.

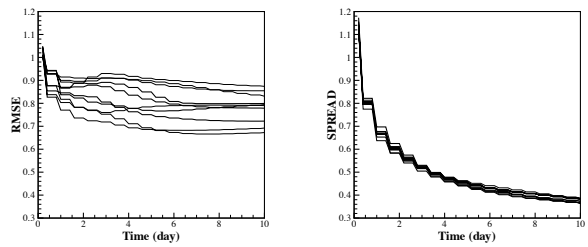

Fig. 8—RMSE and SPREAD for the EnKF with 100 realizations.

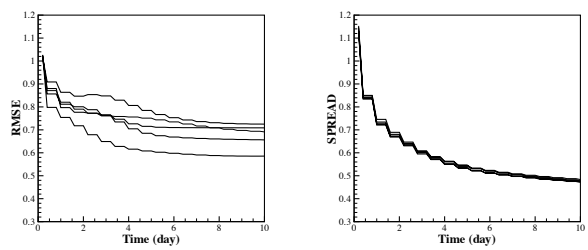

Fig. 9—RMSE and SPREAD for the EnKF with 200 realizations.

Comparison of Accuracy. The ensembles with the best results from the EnKF in the last section are selected to compare with those of the KLKF. **Fig. 11** compares the RMSE and the SPREAD of the KLKF and the best EnKF results from the previously mentioned sets of runs. The KLKF with 100 modes gives better results in terms of both the RMSE and the SPREAD than does the (best) EnKF with 100 realizations. The (best) EnKF shows the lowest RMSE with 200 realizations, which is different from what we would expect, since in general the more realizations used the better estimation we would obtain. On the other hand, there exists a large variation among the five ensembles of size 200 shown in **Fig. 9**.

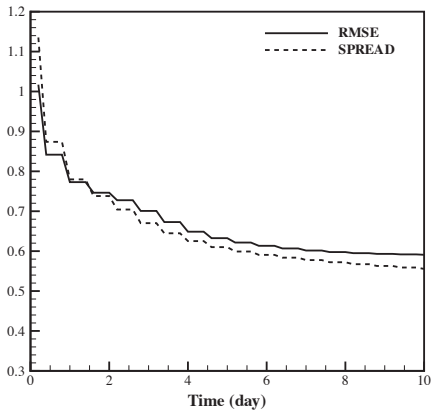

Fig. 10—RMSE and SPREAD for the EnKF with 1000 **realizations.**

This indicates that the performance of the EnKF not only depends on the number of the realizations used but also strongly depends on the representativeness of those realizations. The EnKF may be made more efficient if the initial realizations can be somehow selected more intelligently. In terms of the uncertainty estimation, however, the EnKF with 1000 realizations and the KLKF with 200 modes outperform the EnKF with 200 realizations.

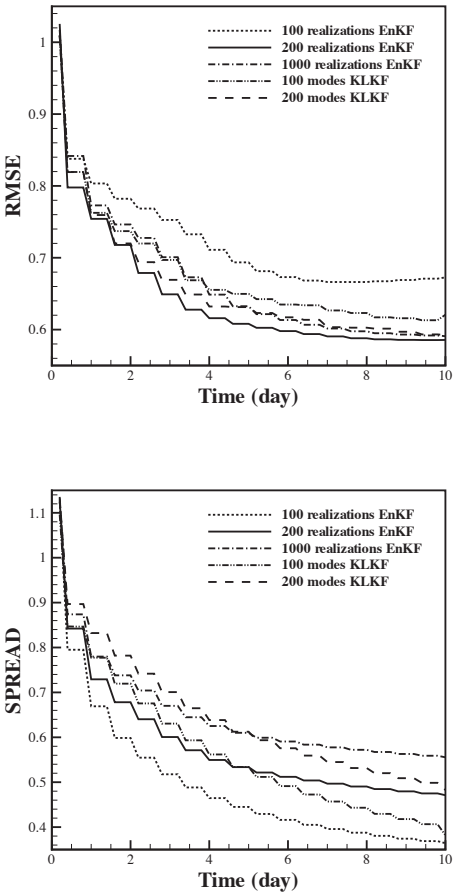

Fig. 11—Comparison of the RMSE and the SPREAD between the EnKF and the KLKF.

The contours of the mean log hydraulic conductivity fields estimated from the KLKF with 100 and 200 modes and from the EnKF with 1000 realizations are shown in the left panel of **Fig. 12**. Compared to the reference field (shown in **Fig. 2**), all these three cases are able to identify the major patterns of the reference field. The estimated variance fields are shown in the right panel of **Fig. 12**. Comparing the KLKF results with 100 modes and 200 modes reveals that the estimated variance is higher if more modes are included in data assimilation. The EnKF with 1000 realizations shows the highest variances which is more consistent with the actual deviations between the estimated mean and the reference, as illustrated in **Fig. 10**. The major patterns of the estimated variance fields from the KLKF and the EnKF are similar, where the variances are lowest at the nine $\ln K_s$ observation locations. Also the variances are lower along the major flow pathes since more heterogeneities have been explored in these regions.

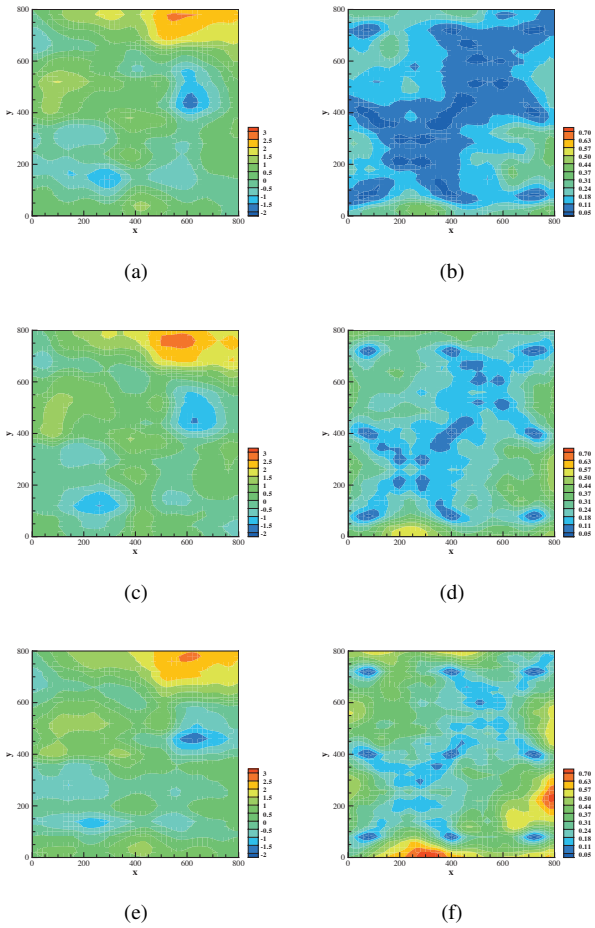

Fig. 12—The estimated mean $\ln K_s$ **fields and the corresponding variance fields: (a) and (b) KLKF with** 100 **modes, (c) and (d) KLKF with** 200 **modes and (e) and (f) EnKF with** 1000 **realizations.**

Predictability. In order to test the predictability of the model after incorporating all the observations, we run a deterministic flow simulation from $t = 0$ day to $t = 20$ day using the final estimated mean log hydraulic conductivity field (at day 10) from the KLKF with 100 modes as the initial input. The corresponding reference pressure head fields are obtained by running the flow simulation using the reference hydraulic conductivity field. Again the initial and boundary conditions are assumed to be known without uncertainty. The calibrated pressure head at day 5 and the predicted pressure head at day 20 are depicted in **Fig. 13** by the dashed lines to com-

pare with the reference pressure head fields shown by the solid lines. Apparently, the estimated head fields match the reference very well.

Computational Efficiency. The computational time for both the EnKF and the KLKF is the time required to run the forward models (flow equations) plus some overhead of matrices manipulations for the update step. For large-scale problems, the forward model run (flow equations) is the major computational burden. If M modes are used in the KLKF, the flow equation needs to be solved for $M+1$ times (once for the zeroth-order head term $h^{(0)}$ and M times for the first-order head terms $h_m^{(1)}$), while for the EnKF with the ensemble size of K, the same equation needs to be solved for K times. Since the CPU time for solving one pressure head term $h_m^{(1)}$ in the KLKF is about the same as that for solving one realization in the EnKF method, the computational efficiency for the two methods simply depends on how many modes (or realizations) are needed to approximate the statistics of the state. The illustrative example showed that the KLKF can achieve satisfactory estimation with a smaller computational cost.

Conclusions

The Kalman filter-based sequential data assimilation methods have been widely used in various fields. These methods are capable of updating the system parameters continuously and sequentially with the availability of the measurements of the system responses. Both the estimate and the corresponding uncertainty are advanced in time.

In this study, a Karhunen-Loeve-based Kalman filter (KLKF) is proposed to reduce the dimensionality of the uncertainty in implementing the Kalman filter scheme to large-scale single-phase flow problems. The permeability field is treated as a random spatial function and is decomposed using the KL expansion. The pressure head is expanded using the perturbative polynomial expansion. On the basis of these expansions, the higher-order terms are truncated and the KLKF is based on the first-order approximation of the pressure head. The KLKF utilizes only a small number of principal modes to propagate the statistics of the state variables, which greatly reduces the computational cost. Since the principal modes are chosen based on the significance in terms of the uncertainty representation (the magnitude of the eigenvalue of the major source of the uncertainty), they can approximate the underlying probability distribution more efficiently than would the same number of randomly generated realizations. The forecast step of the KLKF can be solved accurately and efficiently using the Karhunen-Loeve-based moment-equation (KLME) method, which is suitable for parallel computing using existing flow models (Lu and Zhang, 2006). The update step is operated based on the state statistics given by the forecast step and the observations.

A synthetic 2D example is investigated using the KLKF. The results were compared with those of the ensemble Kalman filter (EnKF), showing that with the same computational efforts the KLKF is able to achieve better estimate than the EnKF. The example indicates that the estimated hydraulic conductivity field using the KLKF method with 100 modes is reasonably close to the reference hydraulic conductivity field and the pressure head predicted using the estimated hydraulic conductivity field agrees well with the reference pressure head field. Although this dimension-reduced Kalman filter is developed and demonstrated for single-phase flow, the general idea of using Karhunen-Loeve and polynomial expansions to represent random fields for reducing the random dimensionality is applicable to multiphase flow.

Nomenclature

C_h = covariance of pressure head h
C_Y = covariance of $Y = \ln K_s$
C_{Yh} = cross-covariance between Y and h
$\mathbf{d}$ = observation vector
D = size of the simulation domain
f_m = eigenfuction
$h^{(0)}$ = zeroth-order pressure head
$h^{(1)}$ = first-order pressure head
$\mathbf{H}$ = observation operator
$\mathbf{K}$ = Kalman gain
K_G = geometric mean of hydraulic conductivity
K_s = hydraulic conductivity
M = number of modes used
$\mathbf{n}$ = an outward unit vector normal to boundaries
N = number of gridblocks
N_h = number of pressure head observations
N_Y = number of log hydraulic conductivity observations
$\mathbf{P}$ = covariance matrix
P = pressure
$\mathbf{S}$ = state vector
S_s = specific storage
t = time
$\mathbf{x}$ = Cartesian coordinate vector
Y = log hydraulic conductivity
Y' = zero-mean fluctuation of Y
ζ = standard Gaussian random variable
λ = eigenvalue
μ = fluid viscosity
ξ_m = orthogonal Gaussian random variables
ρ = fluid density
χ = Cartesian coordinate vector of head observation points
σ_{hN} = standard deviation of the h observation errors
σ_{YN} = standard deviation of the Y observation errors
σ_Y^2 = variance of the log hydraulic conductivity
Γ_D = Dirichlet boundary segment
Γ_N = Neumann boundary segment

Subscript

en = ensemble
Obs = observation

Superscript

(c) = conditioned
f = forecast
t = true
T = transpose
u = updated

Acknowledgments

This work is partially funded by Los Alamos National Lab through LDRD/DR 20040042, by American Chemical Society through its Petroleum Research Fund grant #42109-AC9, and by National Science Foundation through grant OISE-0511496. The authors would like to thank the anonymous reviewers for their constructive comments, which led to an improved paper.

References

Anderson, J. L. and S. L. Anderson, A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, *Monthly Weather Review*, **127**(12), 2741–2758, 1999.

Chen, Y. and D. Zhang, Data assimilation for transient flow in geologic formations via ensemble Kalman filter, *Advances in Water Resources*, **29**, 1107–1122, 2006.

Dietrich, C. and G. Newsam, A stability analysis of the geoststistical approach to aquifer transmissivity identification, *Stochastic Hydrol. Hydraul.*, **3**, 293–316, 1989.

Evensen, G., Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical Research*, **99**(C5), 10,143–10,162, 1994.

Evensen, G., The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dynamics*, **53**, 343–367, 2003.

Evensen, G., Sampling strategies and square root analysis schemes for the EnKF, *Ocean Dynamics*, **54**(6), 539–560, 2004.

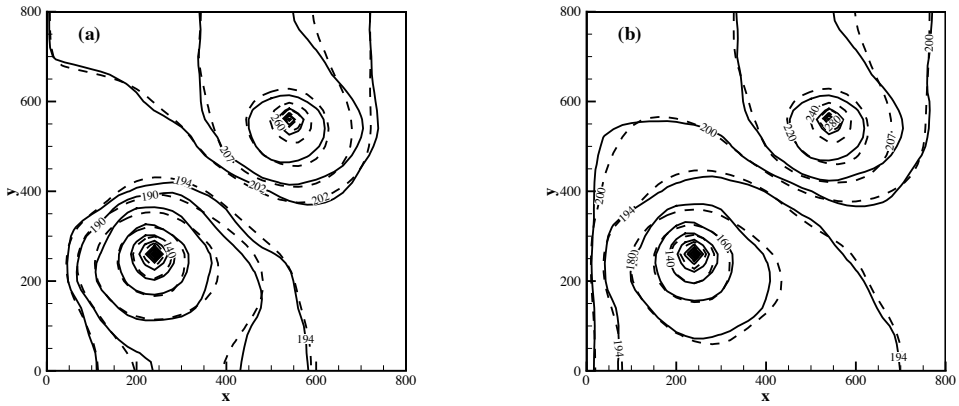

Fig. 13—Comparison of the pressure head fields computed using the reference $\ln K_s$ field (solid lines) and those computed using the estimated mean $\ln K_s$ field from KLKF with 100 modes (dashed lines): (a) pressure head at day 5 and (b) pressure head at day 20.

Gu, Y. and D. S. Oliver, History matching of the PUNQ-S3 reservoir model using the ensemble Kalman filter, *SPE Journal*, **10**(2), 51–65, 2005.

Heemink, A. W., M. Verlaan, and A. J. Segers, Variance reduced ensemble kalman filtering, *Monthly Weather Review*, **129**, 1718–1728, 2001.

Houtekamer, P. L. and H. L. Mitchell, Data assimilation using an ensemble kalman filter technique, *Monthly Weather Review*, **126**, 796–811, 1998.

Liu, N. and D. S. Oliver, Ensemble Kalman filter for automatic history matching of geologic facies, *Journal of Petroleum Science and Engineering*, **47**(3–4), 147–161, 2005.

Lu, Z. and D. Zhang, Conditional simulations of flow in randoml heterogeneous porous media using a kl-based moment-equation approach, *Advances in Water Resources*, **27**(9), 859–874, 2004.

Lu, Z. and D. Zhang, Accurate, efficient quantification of uncertainty for flow in heterogeneous reservoirs using the klme approach, *SPE Journal*, pp. 239–247, 2006.

Mclaughlin, D., An integrated approach to hydrologic data assimilation: interpolation, smoothing, and filtering, *Advances in Water Resources*, **25**, 1275–1286, 2002.

Nævdal, G., L. M. Johnsen, S. I. Aanonsen, and E. H. Vefring, Reservoir monitoring and continuous model updating using ensemble Kalman filter, *SPE Journal*, **10**(1), 66–74, 2005.

Pham, D. T., J. Verron, and M. C. Roubaud, A singular evolutive extended kalman filter for data assimilation in oceanography, *Journal of Marine System*, **16**, 323–340, 1998.

Verlaan, M. and A. W. Heemink, Tidal flow forecasting using reduced-rank square root filters, *Stochastic Hydrol. Hydraul.*, **11**, 349–368, 1997.

Wen, X.-H. and W. Chen, Real-time reservoir model updating using ensemble Kalman filter, SPE-92991, in *Proceedings of the 2005 SPE Reservoir Simulation Symposium*, pp. 1–14, 2005.

Whitaker, J. and T. Hamill, Ensemble assimilation without perturbed observation, *Monthly Weather Review*, **130**, 1913–1924, 2002.

Xiu, D. and G. E. Karniadakis, Modeling uncertainty in flow simulations via generalized polynomial chaos, *Journal of Computational Physics*, **187**, 137–167, 2003.

Yeh, T. C., M. Jin, and S. Hanna, An iterative stochastic inverse method: Conditional effective transmissivity and hydraulic head fields, *Water Resour. Res.*, **32**(1), 85–92, 1996.

Zhang, D., *Stochastic Methods for Flow in Porous Media: Coping with uncertainties*, Academic Press, San Diego, Calif, 2002.

Zhang, D. and Z. Lu, An efficient, higher-order perturbation approach for the flow in randomly heterogeneous porous media via karhunen-loeve decomposition, *J. Comput. Physics*, **194**(2), 773–794, 2003.

Dongxiao Zhang is the Miller Chair Professor in Petroleum and Geological Engineering at the U. of Oklahoma, where he has been a faculty member since 2004. From 1996 to 2003, he was a technical staff member and a team leader at the Los Alamos Natl. Laboratory. His research interests include stochastic uncertainty quantification of reservoir simulations, pore-scale processes, and geological carbon sequestration. Zhang holds MS and PhD degrees in hydrology from the U. of Arizona. **Zhiming Lu** is a technical staff member in the Earth and Environmental Sciences Div. at the Los Alamos Natl. Laboratory. His research interests include quantifying uncertainty for flow and transport in randomly heterogeneous porous media and inverse modeling. He holds a BS degree in applied mathematics from Zhejiang U., China, an MS degree in applied earth sciences from Stanford U., and a PhD degree in hydrology from the U. of Arizona. **Yan Chen** is a PhD student at the U. of Oklahoma majoring in petroleum and geological engineering. Her research interests include stochastic uncertainty quantification, automatic history matching, and reservoir optimization. She holds a BS degree in geological engineering from Nanjing U., China.