Comput Geosci (2010) 14:721–744
DOI 10.1007/s10596-010-9183-5

ORIGINAL PAPER

A stochastic collocation based Kalman filter for data assimilation

Lingzao Zeng · Dongxiao Zhang

Received: 15 January 2009 / Accepted: 1 March 2010 / Published online: 23 March 2010

Abstract In this paper, a stochastic collocation-based Kalman filter (SCKF) is developed to estimate the hydraulic conductivity from direct and indirect measurements. It combines the advantages of the ensemble Kalman filter (EnKF) for dynamic data assimilation and the polynomial chaos expansion (PCE) for efficient uncertainty quantification. In this approach, the random log hydraulic conductivity field is first parameterized by the Karhunen–Loeve (KL) expansion and the hydraulic pressure is expressed by the PCE. The coefficients of PCE are solved with a collocation technique. Realizations are constructed by choosing collocation point sets in the random space. The stochastic collocation method is non-intrusive in that such realizations are solved forward in time via an existing deterministic solver independently as in the Monte Carlo method. The needed entries of the state covariance matrix are approximated with the coefficients of PCE, which can be recovered from the collocation results. The system states are updated by updating the PCE coefficients. A 2D heterogeneous flow example is used to demonstrate the applicability of the SCKF with respect to different factors, such as initial guess, variance, correlation length, and the number of observations. The results are compared with those from the EnKF method. It is shown that the SCKF is computationally more efficient than the EnKF under certain conditions. Each approach has its own advantages and limitations. The performance of the SCKF decreases with larger variance, smaller correlation ratio, and fewer observations. Hence, the choice between the two methods is problem dependent. As a non-intrusive method, the SCKF can be easily extended to multiphase flow problems.

Keywords Data assimilation · Stochastic collocation · Ensemble Kalman filter · Uncertainty quantification

L. Zeng · D. Zhang
Sonny Astani Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, CA, USA

D. Zhang (✉)
Department of Energy and Resources Engineering, College of Engineering, Peking University, Beijing 100871, People's Republic of China
e-mail: dxz@pku.edu.cn

1 Introduction

Geologic formations are intrinsically deterministic. However, owing to the incomplete knowledge of the medium properties, such as the hydraulic conductivity and porosity, these parameters are usually treated as random space functions, and the equations describing flow and transport in the media become stochastic [1–3]. Large efforts have been made to estimate the parameters of the formations from all available observations. Owing to the high cost associated with direct measurements of formation parameters, the observations often include a limited number of direct measurements and a certain amount of indirect measurements. Estimating the parameters from the indirect measurements is a challenging inverse problem.

The Kalman filter is widely used as a sequential data assimilation method [4]. It is optimal if all the probability distributions involved are Gaussian, i.e., the system is linear, and all the random variables are normally

distributed. Different methods have been proposed to apply Kalman filter to nonlinear problems. These approaches include extended Kalman filter (EKF) [5], ensemble Kalman filter (EnKF) [6], and their variants. EKF is based on the first-order linearization of the system. It becomes rather time consuming when dealing with large-scale problems. The EnKF is essentially a Monte Carlo method. The information of the state covariance is represented by an ensemble of realizations. Owing to its conceptual simplicity, ease in implementation, and relatively lower computational cost compared to other approaches, the EnKF has been widely used in different fields such as meteorology, oceanography, hydrology, and reservoir engineering [6–13]. However, the size of the ensemble is crucial for the performance of the EnKF. Owing to the slow convergence with the ensemble size N_e, a large ensemble size is required to get accurate estimations of the system and an even larger size for the estimation of the associated uncertainty. On the other hand, since it is time consuming to run each simulation for large-scale problems, one can only afford a small ensemble size. Some methods have been proposed to reduce the sampling errors in the EnKF with small-sized ensembles. Ensemble square root filter [14] uses different Kalman gains to update the ensemble mean and the perturbations separately. Double ensemble Kalman filter [15] divides the ensemble into two parts, each of which is updated using the Kalman gain calculated from the other. Furthermore, some ad hoc techniques, such as localizations and inflations, have been proposed to handle the spurious correlations approximated by the small sized ensemble [16]. These approaches are found to give improved results with relatively small ensemble sizes. However, extra efforts are needed.

Recently, there are increasing interests in solving inverse problems via the stochastic spectral method. As one of the most popular stochastic spectral methods, the polynomial chaos expansion (PCE) method, pioneered by Ghanem and Spanos in the field of stochastic mechanics [17], provides a powerful tool in uncertainty quantification. In this method, the random process of interest is represented by the polynomial chaos basis. The expansion coefficients are solved via the Galerkin technique. This PCE method allows high-order approximations of random input variables. Optimal convergence can be achieved by choosing the proper basis, known as the generalized polynomial chaos (gPC) [18]. A Bayesian approach to a transient diffusion problem based on PCE was proposed in [19]. The PCE was used to accelerate the Bayesian inference without solving the stochastic differential equation. A PCE-based EnKF was developed in data assimilation for multiphase flow problem [20], where the inputs were random variables and the statistics of the states were represented by PCE terms. In the above two approaches, the Galerkin technique was employed; hence, one has to solve coupled equations for the PCE coefficients. It becomes difficult when the governing equations take complicated nonlinear forms. A dimension-reduced Kalman filter based on the Karhunen–Loeve-based moment equation (KLME) method for reservoir data assimilation was developed in [21]. The forward problem was solved by the KLME method. The estimations of the hydraulic conductivity field were sequentially updated using updated KL expansion coefficients. In this approach, the equations are not coupled but recursive since the high-order equations depend on the lower-order ones. This method cannot be easily extended to multiphase (nonlinear) problems, since the KLME results in new types of equations at high order. In all the spectral approaches listed above, new codes need to be developed to deal with the corresponding new equations.

To alleviate this difficulty, collocation methods such as the probabilistic collocation method (PCM) [22–24] and the stochastic collocation method [25, 26] have been developed for uncertainty quantification. A comparative study of different collocation methods for flow in porous media can be found in [27]. In these methods, after choosing collocation point sets in the random space, one only needs to solve the corresponding deterministic governing equation repeatedly. In this sense, the stochastic collocation methods are non-intrusive as in the traditional Monte Carlo method. It is, however, found that the former are more efficient than the latter under certain conditions. The research in inverse problems via collocation methods just started very recently. A method for the stochastic inverse heat conduction was proposed in [28], where the stochastic inverse problem was transformed to a deterministic optimization problem via a sparse grid collocation method. However, it still requires developing new codes to solve the resulting sensitivity equations, what may be difficult for complex systems. Furthermore, it uses the observations in the entire history (not real time) and is thus very demanding for data storage in geological problems. These are typical disadvantages of gradient-based methods in the inverse problem.

In this study, a stochastic collocation-based Kalman filter (SCKF) is developed to sequentially update the formation conductivity field from all available observations. The covariance matrix is approximated by the coefficients of PCE, which are obtained from the stochastic collocation results. This paper is organized as follows. In Section 2, the mathematical formulations

are presented briefly. The implementation of the SCKF is discussed in Section 3. Then, illustrative examples are given in Section 4 to show the applicability of the SCKF. Some discussion is given in Section 5 before the paper in concluded.

2 Mathematical formulations

2.1 Governing equation

We consider transient water flow in saturated geologic formations satisfying the following governing equation:

$$\nabla\left[K_{\mathrm{s}}(\mathbf{x})\nabla h(\mathbf{x},t)\right]+g(\mathbf{x},t)=S_{\mathrm{s}}\frac{\partial h(\mathbf{x},t)}{\partial t}, \tag{1}$$

subject to the initial and boundary conditions:

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

$$h\left(\mathbf{x},t\right)=H\left(\mathbf{x},t\right),\mathbf{x}\in\Phi_D, \tag{3}$$

$$K_{\mathrm{s}}\left(\mathbf{x}\right)\nabla h\left(\mathbf{x},t\right)\cdot n\left(\mathbf{x}\right)=-Q\left(\mathbf{x},t\right),x\in\Phi_N, \tag{4}$$

where $g(\mathbf{x},t)$ is the source/sink term, $h(\mathbf{x},t)$ is the pressure head, $H_0(\mathbf{x})$ is the initial head in the domain D, $H(\mathbf{x},t)$ is the prescribed head on Dirichlet boundary segments Φ_D, $K_{\mathrm{s}}(\mathbf{x})$ is the hydraulic conductivity, $Q(\mathbf{x},t)$ is the prescribed flux across Neumann boundary segments Φ_N, $n(\mathbf{x})=(n_1,n_2,\ldots n_d)^T$ is an outward vector normal to the boundary Φ_N, and S_{s} is the specific storage. In this study, the conductivity $K_{\mathrm{s}}(\mathbf{x})$ is considered as a random space function with lognormal distribution. We usually work with the log transformed hydraulic conductivity $Y=\ln K_{\mathrm{s}}$. We treat the specific storage S_{s} as a deterministic constant.

Since $K_{\mathrm{s}}(\mathbf{x})$ is a random function, the above flow equations become stochastic partial differential equations, which can be solved by different methods. In this study, a stochastic collocation method is used. This method is based on the stochastic spectral expansions of random processes, which are formulated in the following sections.

2.2 Karhunen–Loeve expansion

Let $Y\left(\mathbf{x},\omega\right)=\ln\left[K_{\mathrm{s}}\left(\mathbf{x},\omega\right)\right]$ be a Gaussian stochastic process, where $\mathbf{x}\in D$ and $\omega\in\Omega$ (a probability space). Since the covariance function $C_Y\left(\mathbf{x},\mathbf{y}\right)=\left\langle Y'\left(\mathbf{x},\omega\right)Y'\left(\mathbf{y},\omega\right)\right\rangle$ is bounded, symmetric, and positive definite, it can be decomposed into

$$C_Y\left(\mathbf{x},\mathbf{y}\right)=\sum_{i=1}^{\infty}\lambda_i f_i\left(\mathbf{x}\right)f_i\left(\mathbf{y}\right), \tag{5}$$

where λ_i and $f_i(x)$ are eigenvalues and eigenfunctions, respectively. The stochastic process $Y(\mathbf{x},\omega)$ can be expanded with Karhunen–Loeve (KL) expansion as:

$$Y\left(\mathbf{x},\omega\right)=\bar{Y}\left(\mathbf{x}\right)+\sum_{i=1}^{\infty}\varepsilon_i\left(\omega\right)\sqrt{\lambda_i}f_i\left(\mathbf{x}\right), \tag{6}$$

where $\bar{Y}(\mathbf{x})$ is the mean component and ε_i are independent Gaussian random variables with unit variance and zero mean. It has been shown that the KL expansion is of mean square convergence when the underlying process is Gaussian. In practice, the expansion is usually truncated up to the first M terms. Increased number of terms is needed to sufficiently approximate the random field with the decrease of the correlation scale relative to the domain size (correlation ratio). For the correlation function with some special forms such as the separable exponential form used in this paper, the KL expansion can be obtained analytically. Usually, one has to solve the Fredholm equation to get the eigenvalues and eigenfunctions numerically [17]. This problem can be transformed to eigen-decomposition of the covariance matrix. It is very time consuming when the models are large. Two methods can be used to reduce the computational burden. One is to use interpolation based on coarse nodes to approximate the covariance and the corresponding eigenfunctions. If the covariance is smooth enough, this method can effectively approximate the first few KL expansion terms. Another new approach is kernel method, which is to use Monte Carlo realizations to approximate the eigenfunctions and eigenvalues [29]. In this method, one has to check the convergence of the covariance represented by a limited number of realizations.

2.3 Polynomial chaos expansion

The polynomial chaos expansion is more general than the KL expansion. It can be used to represent random processes without the prior knowledge of the

covariance function. With this expansion, the random process of interest can be expressed as

$$y(\mathbf{x},\omega) = a_0(\mathbf{x}) + \sum_{i_1=1}^{\infty} a_{i_1}(\mathbf{x})\,\Gamma_1\left(\xi_{i_1}(\omega)\right) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1} a_{i_1 i_2}(\mathbf{x})\,\Gamma_2\left(\xi_{i_1}(\omega),\xi_{i_2}(\omega)\right) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1}\sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3}(\mathbf{x})\,\Gamma_3\left(\xi_{i_1}(\omega),\xi_{i_2}(\omega),\xi_{i_3}(\omega)\right)+\ldots, \tag{7}$$

where the coefficient functions $a_0(\mathbf{x})$ and $a_{i_1 i_2,\ldots,i_d}(\mathbf{x})$ are deterministic and unknown. $\Gamma_d\left(\xi_{i_1},\xi_{i_2},\ldots,\xi_{i_d}\right)$ are multi-dimensional Hermite polynomials of order d

$$\Gamma_d\left(\xi_{i_1},\ldots,\xi_{i_d}\right) = (-1)^d e^{\frac{1}{2}\xi^T\xi}\frac{\partial^d}{\partial\xi_{i_1}\ldots\partial\xi_{i_d}}\left[e^{-\frac{1}{2}\xi^T\xi}\right], \tag{8}$$

where ξ is a vector denoting $\left(\xi_{i_1},\ldots,\xi_{i_d}\right)^T$. Hermite polynomials are optimal basis for Gaussian processes with exponential convergence rate. A more general discussion about random variables with different distributions can be found in [18].

In practice, Eq. 7 is usually truncated by a finite number of terms and can be rewritten as follows

$$y(\mathbf{x},t,\omega) = \sum_{j=0}^{Q} c_j(\mathbf{x},t)\,\Psi_j(\xi). \tag{9}$$

There is a one-to-one correspondence between the terms in Eqs. 7 and 9. The total number of terms $(Q+1)$ can be determined by the random dimensionality M and the order of the polynomial chaos expansion d,

$$Q+1 = \frac{(M+d)!}{M!d!}. \tag{10}$$

2.4 Stroud-2-based stochastic collocation method

The stochastic collocation methods are based on the theory of multidimensional integration. In these methods, after parameterizing the random inputs with random variables, the governing equations are solved at the given point sets in the random space. Among different collocation methods, the cubature rule of degree 2, also called Stroud-2, requires the minimal collocation sets. Since the random dimensionality in our problem is large ($M \geq 100$), the Stroud-2 points are employed. For M random dimensional problems, $M+1$ collocation point sets are needed. For multidimensional integration with Gaussian weights, the collocation point sets $\xi_k = \left[\xi_{k,1},\xi_{k,2},\ldots,\xi_{k,M}\right]$, $k = 0,1,\ldots,M$, are defined as [30]

$$\xi_{k,2r-1} = \sqrt{2\cos\frac{2rk\pi}{M+1}},\ \xi_{k,2r} = \sqrt{2\sin\frac{2rk\pi}{M+1}},\ r = 1,2,\ldots,\left[M/2\right], \tag{11}$$

where $[M/2]$ is the greatest integer not exceeding $M/2$, if M is odd, $\xi_{k,M} = (-1)^k$. Collocation points for random inputs with other distributions can also be found in [30]. Each collocation set is of equal weight. The collocation points given in Eq. 11 form an integration formula of degree 2, i.e.

$$\int_{\Omega} G(\xi)\,\rho(\xi)d\xi \approx \frac{1}{M+1}\sum_{i=0}^{M} G(\xi_i), \tag{12}$$

where Ω is M dimensional space and ρ is multivariable Gaussian probability density function. If we represent a random process in PCE form as shown in Eq. 9, the coefficient $c_j(\mathbf{x},t)$ can be obtained by making projections onto each basis as follows

$$c_j(\mathbf{x},t) = \frac{\left\langle y(\mathbf{x},t,\omega)\Psi_j(\xi)\right\rangle}{\left\langle\Psi_j^2\right\rangle} = \frac{\int_{\Omega} y(\mathbf{x},t,\omega)\Psi_j(\xi)\rho(\xi)d\xi}{\left\langle\Psi_j^2\right\rangle}. \tag{13}$$

According to the Stroud-2 cubature rule, Eq. 13 can be approximated by the following expression

$$c_j(\mathbf{x},t) \approx \frac{1}{(M+1)\left\langle\Psi_j^2\right\rangle}\sum_{i=0}^{M} y(\mathbf{x},t,\xi_i)\,\Psi_j(\xi_i), \tag{14}$$

where $y(\mathbf{x},t,\xi_i)$ is the collocation result given the ith collocation set, and $\Psi_j(\xi_i)$ is the jth PCE basis given the ith collocation set. Equation 14 can be seen as the post-processing step of the stochastic collocation method. Note that the Stroud-2 rule is exact for integrations of polynomials of degree at most two. Hence, Eq. 14 is accurate as long as $\Psi_j(\xi)$ is up to the first order. Therefore, if we recover PCE terms from the results of the Stroud-2-based collocation method, the approximation is only up to the first order and Q equals M. Although this accuracy is relatively low, this collocation-based method is still effective in some problems where the random dimensionality is large.

It should be noted that the SCKF is not limited to Stroud-2 collocation rule. It can be employed with PCE up to any order, if the computational effort is affordable. In the implementation of the stochastic collocation method, at the first step, one has to decide

the order of PCE according to the nonlinearity of the problem. Since for random variables with certain distributions, the orthogonal polynomial basis Ψ_j is fixed. With the chosen order and corresponding collocation points ξ_i, $\Psi_j(\xi_i)$ is thus fixed during the entire procedure. Therefore, it can be calculated once and saved for later use.

2.5 Ensemble Kalman filter

The ensemble Kalman filter (EnKF) is a Monte Carlo method. It is easy to implement and similar to the Kalman filter. In the Kalman filter, the covariance is explicitly computed and propagated in time, while in the EnKF, the covariance is calculated from the ensemble. The basic formulas are listed below.

In this single phase flow problem, a joint vector is defined as

$$\mathbf{s} = \left[\mathbf{Y}^T \quad \mathbf{h}^T \right]^T, \tag{15}$$

where the state vector $\mathbf{Y}^T$ is for the log conductivity, and $\mathbf{h}^T$ is for the pressure head. In the implementation of the EnKF, realizations of state vector are collected in a matrix to form an ensemble

$$\mathbf{S} = \left[\mathbf{s}_1 \quad \mathbf{s}_2 \quad \dots \quad \mathbf{s}_{N_e} \right], \tag{16}$$

where N_e is the ensemble size. The forecast state $\mathbf{s}^f(i)$ can be obtained by running each ensemble member with any existing simulator. In the analysis step with time index i, each ensemble member is updated via

$$\mathbf{s}_j^a(i) = \mathbf{s}_j^f(i) + \mathbf{P}(i)\,\mathbf{H}^T \left[\mathbf{H}\mathbf{P}(i)\,\mathbf{H} + \mathbf{R}(i)\right]^{-1} \left(\mathbf{d}_{\text{obs},j}(i) - \mathbf{H}\mathbf{s}_j^f(i)\right),\ j = 1, 2, \dots, N_e, \tag{17}$$

where j is the member index, $\mathbf{d}_{\text{obs},j}(i)$ is the perturbed observation, and $\mathbf{H}$ is the observation operator. $\mathbf{R}(i)$ is the covariance matrix of the observation errors, and $\mathbf{P}(i)$ is the covariance matrix of the forecasted states, which can be calculated from the ensemble.

3 Kalman filter in PCE basis space

In this section, we discuss the combination of the Kalman filter and the PCE. Since the state $\mathbf{s}$ is a random process, it can be expanded using PCE as

$$\mathbf{s} = \sum_{j=0}^{Q} \mathbf{c}_j \Psi_j(\xi). \tag{18}$$

$\mathbf{c}_j$ is defined as

$$\mathbf{c}_j = \left[\mathbf{c}_{Y,j}^T \quad \mathbf{c}_{h,j}^T \right]^T,\ j = 0, 1, \dots Q, \tag{19}$$

where $\mathbf{c}_{Y,j}^T$ and $\mathbf{c}_{h,j}^T$ are the PCE coefficient vectors of the log conductivity and pressure head, respectively. Because the Stroud-2 collocation method is used in this paper, both the log conductivity and the pressure head are approximated to first order. Since Q equals M, only M is used in all the following formulas. The covariance can be expressed with non-zeroth order PCE coefficients as

$$\mathbf{P} = \left\langle \left(\sum_{j=1}^{M} \mathbf{c}_j \Psi_j \right) \left(\sum_{j=1}^{M} \mathbf{c}_j \Psi_j \right)^T \right\rangle = \sum_{j=1}^{M} \mathbf{c}_j \left(\mathbf{c}_j\right)^T \left\langle \Psi_j^2 \right\rangle. \tag{20}$$

The analysis step at the time indexed by i can be written in the PCE form as

$$\sum_{j=0}^{M} \mathbf{c}_j^a(i)\,\Psi_j = \sum_{j=0}^{M} \mathbf{c}_j^f(i)\,\Psi_j + \mathbf{P}(i)\,\mathbf{H}^T \left[\mathbf{H}\mathbf{P}(i)\,\mathbf{H}^T + \mathbf{R}(i)\right]^{-1} \times \sum_{j=0}^{M} \left[\mathbf{d}_{\text{obs}}(i)\,\delta_{0j} - \mathbf{H}\mathbf{c}_j^f(i)\right] \Psi_j, \tag{21}$$

where $\mathbf{d}_{\text{obs}}(i)$ is the observation. Since the measurement errors are independent of the system state, the PCE terms of observations are expressed with the Kronecker delta. Multiplying by Ψ_j and taking expectation, Eq. 21 yields

$$\mathbf{c}_j^a(i) = \mathbf{c}_j^f(i) + \mathbf{K}(i) \left[\mathbf{d}_{\text{obs}}(i)\,\delta_{0j} - \mathbf{H}\mathbf{c}_j^f(i)\right],\ j = 0, 1, \cdots, M, \tag{22}$$

where $\mathbf{K}(i)$ is the Kalman gain. Therefore, each PCE coefficient of the log conductivity and the pressure head is updated separately. Here, the log conductivity field Y is initially parameterized by the KL expansion, which is expressed by the first-order of Gaussian random variables. It is interesting to note that, if PCE with higher order is used in the problem, higher-order PCE coefficients of the log conductivity Y will be produced after the analysis, and the conditional field will become non-Gaussian. Once the PCE coefficients are updated, corresponding collocation realizations can be obtained via

$$\mathbf{s}^a(\xi_i) = \sum_{j=0}^{M} \mathbf{c}_j^a \Psi_j(\xi_i),\ i = 0, 1, \dots, M, \tag{23}$$

where $\Psi_j(\xi_i)$ is the jth PCE term at the ith collocation set. Once the updating step is finished, with the new conductivity realizations, the forecast step moves to the next time when the observations become available. The algorithm of stochastic collocation-based Kalman filter

can be summarized in Fig. 1. It is of interest to note that besides the statistical moments obtained directly with Eq. 23, more (conditional) realizations can be generated with new random vectors ξ, on the basis of which probability density functions of the state vector may be obtained.

In the standard EnKF, the initial realizations are usually generated based on the random sampling. Some strategies in the initial sampling were discussed in [31]. It is interesting to note that, since the SCKF is implemented at the designed collocation point sets, it is very similar to the EnKF with deterministic sampling methods. In fact, it can be shown that, with the Stroud-2 sampling method, the EnKF with non-perturbed observations is equivalent to the SCKF with the first-order PCE. The proof is given in Appendix.

When the PCE order is higher than one, we can solve the problem with the PCM [23, 32, 33]. For each uncertain parameter, the collocation points are selected from the roots of the next higher-order orthogonal polynomial. After solving the forecasting problem via PCM, we can update the PCE coefficients via Eq. 22. It should be noted that, in the analysis step, since each PCM realization is no longer equally weighted, we cannot update each realization directly. In this paper, the nonlinearity of the single phase problem is not strong, the Gaussian assumption is not strongly violated, and collocation method based on Stroud-2 rule is able to capture the dynamics of the system. Although updating the PCE coefficients is equivalent to updating the Stroud-2 realizations, implementing the analysis step in the PCE framework shows the order of accuracy more clearly.

In the SCKF, the first step is to parameterize the uncertain inputs with a set of independent random variables. For the Gaussian random fields, which can be completely characterized by the first two moments, the KL expansion is a convenient tool for parameterization. Parameterization of non-Gaussian random fields is still an active research area. Some approaches have been proposed based on different assumptions. It is common that the prior statistics are only the correlation and the marginal distributions, which cannot sufficiently characterize a non-Gaussian field. In that case, one possible approach is to use polynomial transformations of the Gaussian process (PCE) to match the one-point

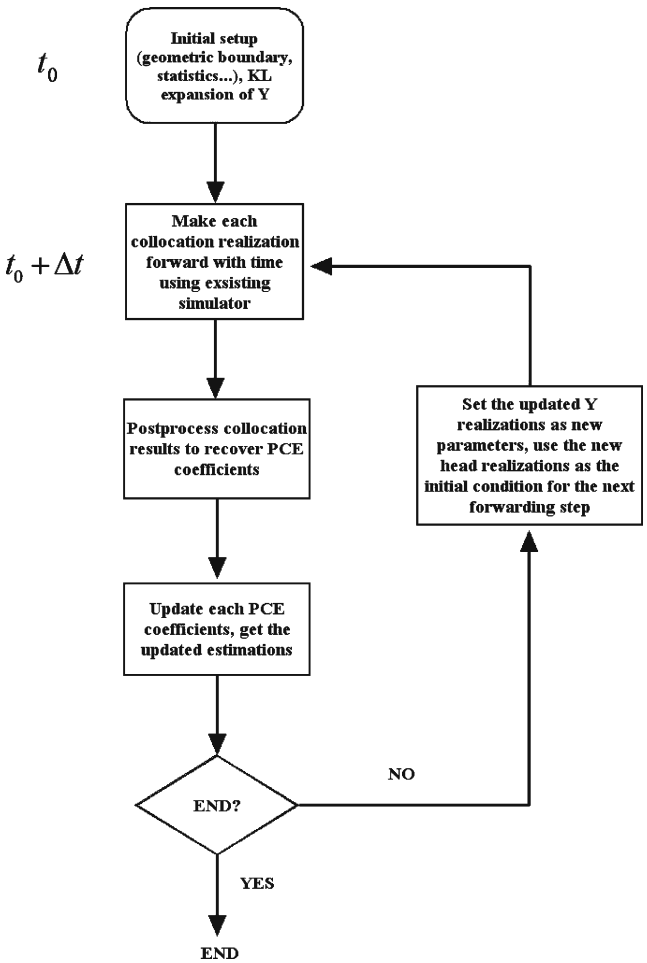

Fig. 1 Flow chart of stochastic collocation based Kalman filter

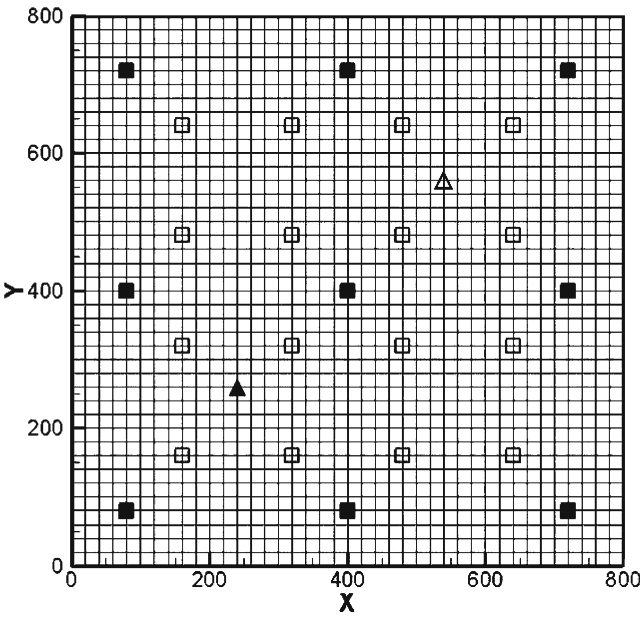

Fig. 2 The flow domain and the observation locations for log hydraulic conductivity (nine *filled squares*) and pressure head (all the 25 *squares*). The *filled triangle* and *empty triangle* are the pumping and injection well, respectively

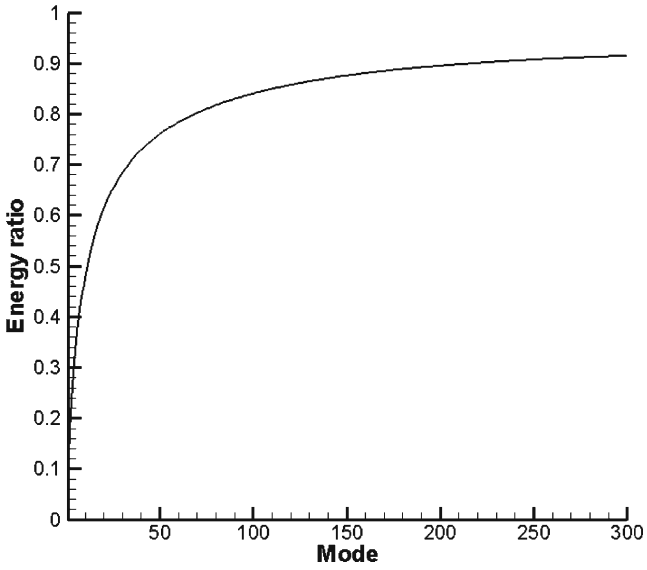

Fig. 3 The ratio of energy captured

marginal probability density functions, and use the KL expansion of the underlying Gaussian process to match the target correlation function [34]. With the PCE-based parameterization of the non-Gaussian random fields, we can implement the SCKF as described in this paper. If samples of the non-Gaussian field are assumed to be available, another approach is to employ the KL expansion as a dimension reduction tool to parameterize the field with a set of uncorrelated random variables [35]. The distributions of random variables in the KL expansion are no longer Gaussian and can be estimated from the samples. Based on the specific marginal distribution of each random variable, the orthogonal generalized polynomial chaos basis can be constructed. In this approach, the random variables in the KL expansion are assumed to be independent. This approximation can simplify the problem. On the other hand, model errors will be introduced since the random variables are essentially uncorrelated (statistically dependent). Hence, the information of joint statistics among these variables will be lost. In [35], the authors discussed some correcting methods to alleviate this problem. Our ongoing investigations show that, combined with independent component analysis, the KL expansion can parameterize the non-Gaussian fields with a set of independent random variables. Then, the forward problem can be solved by collocation methods, and the coefficients of gPC can be updated similarly.

We should also keep in mind that both the SCKF and EnKF are variants of the Kalman filter, in which only the first two moments are used. They are both suboptimal for nonlinear problems. If the nonlinearity is strong, more sophisticated methods should be employed, however, with the cost of larger computational efforts. A generalization of the ensemble Kalman filter for the non-Gaussian channelized field has been proposed in [36] recently, where the higher-order statistics were used to update the states.

4 Illustrative examples

In this section, in order to demonstrate the applicability of the SCKF to estimate the hydraulic conductivity by assimilating the measurements of the pressure head and the hydraulic conductivity, a 2D model of transient saturated flow is used. The results are compared with those of the EnKF method.

In the implementations of both the SCKF and the EnKF, since the same model (MODFLOW) is used for solving both the forecast model and the reference, we assume that the system is free of model errors. The

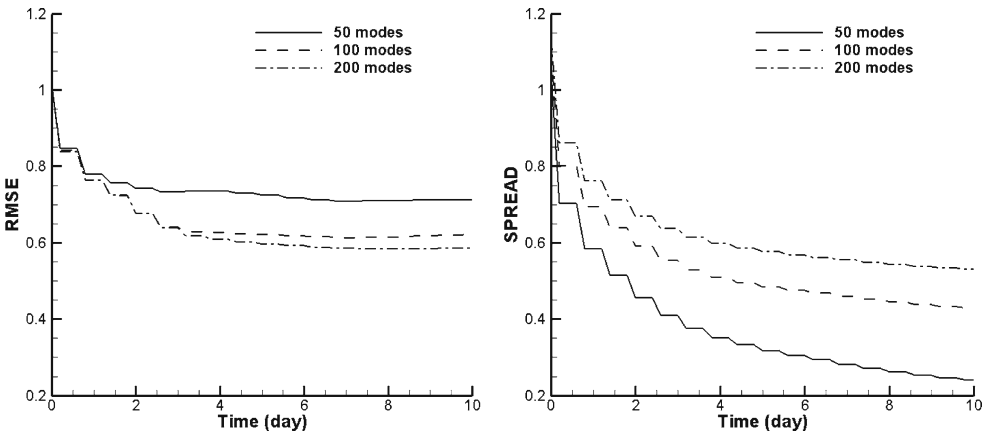

Fig. 4 The RMSE and SPREAD for the SCKF with different number of modes

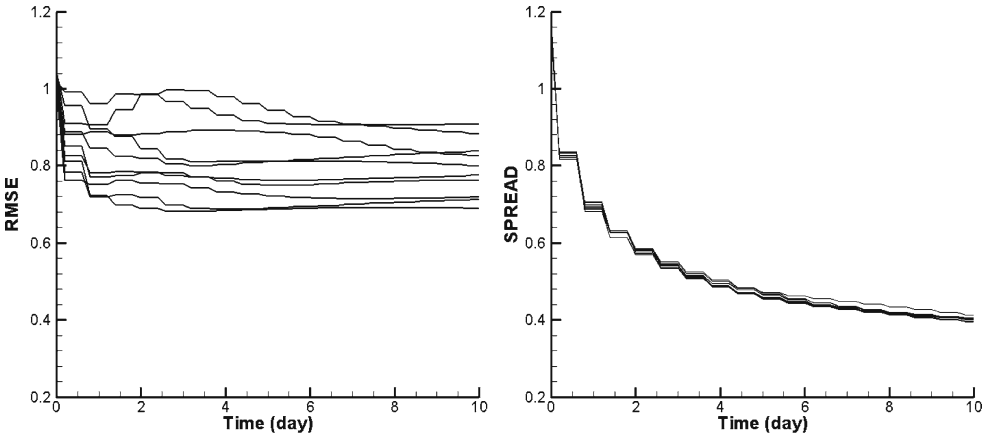

Fig. 5 RMSE and SPREAD for ten groups of EnKF with 100 realizations

domain is a square of size $L_x = L_y = 800m$, which is uniformly discretized into 40×40 elements, as shown in Fig. 2. A pumping well and an injection well are placed at ($240m$ and $260m$) and ($540m$ and $560m$), respectively, with a constant volumetric flow rate of 150 m^3/day during the entire assimilation time. The two lateral sides are no-flow boundaries, while the left and the right are Dirichlet boundaries with prescribed pressure heads of $202m$ and $198m$, respectively. The constant storage coefficient is assumed to be 0.0001. The log hydraulic conductivity field is treated as spatially correlated Gaussian random field with zero mean and unit variance. The correlation of the unconditional hydraulic conductivity field is assumed to be in a separable exponential form,

$$C_Y(\mathbf{x}_1, \mathbf{x}_2) = C_Y(x_1, y_1; x_2, y_2) = \sigma^2 \exp\left[-\frac{|x_1 - x_2|}{\lambda_x} - \frac{|y_1 - y_2|}{\lambda_y}\right], \quad (24)$$

where σ^2 is the variance and λ_x and λ_y are the correlation length in the x and y directions, respectively.

In this example, an unconditional realization of the log hydraulic conductivity field with given statistics ($\sigma^2 = 1.0$, $\lambda_x = 200m$ and $\lambda_y = 100m$) is generated by the KL decomposition. This field is then considered as the true field, called the reference field. A forward transient simulation is conducted using the reference hydraulic conductivity field. Ten days is chosen as the duration of the total assimilation time, which is equally subdivided into 50 time intervals with a size of 0.2 day. As shown in Fig. 2, observations are obtained at 25 locations. Nine measurements of the log hydraulic conductivity field are taken at the filled squares, and 25 measurements of the pressure head are obtained at all the squares. At $t = 0.2$ day, both the pressure head and conductivity are measured. After that, only pressure heads are measured at every 0.6 day up to day 10. The measurements of the hydraulic conductivity are assumed to be perfect, and the measurement errors of pressure head are assumed to follow a Gaussian

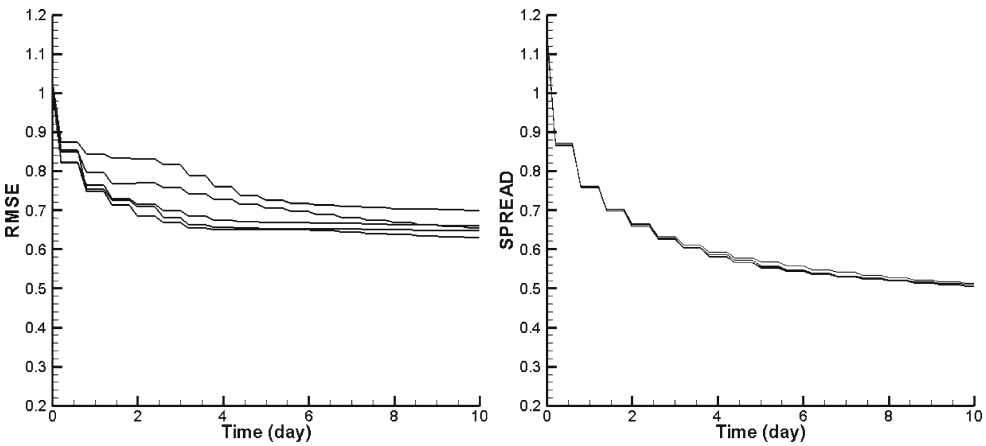

Fig. 6 RMSE and SPREAD for five groups of EnKF with 200 realizations

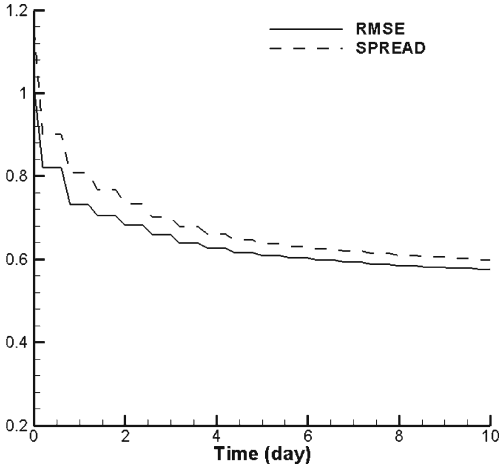

Fig. 7 RMSE and SPREAD for the EnKF with 1,000 realizations

distribution with zeros mean and standard deviation $\sigma = 0.05m$. For both the EnKF and SCKF, the initial pressure head is assumed to be known without uncertainty. To reduce the condition number of the matrix $[\mathbf{H}\mathbf{P}^f(i)\mathbf{H}^T + \mathbf{R}(i)]$ in Eq. 21, a relaxation term is added to the diagonal terms, as in the KLME-based Kalman filter [21]. This relaxation term improves the stability of the SCKF. There is no standard method to decide the relaxation value. Our numerical experiments show that the performance of the SCKF is sensitive to the choice of relaxation term when the number of modes is small, for example, being 50. With larger number of modes, the SCKF performs well as long as the relaxation term remains within a reasonable range. For the sake of comparison, in the following discussions, the same relaxation term 0.3 is used in both the SCKF and the EnKF.

5 Results and discussions

Using the KL expansion, we parameterize the log conductivity field with a set of finite number (modes) of independent standard Gaussian variables. Therefore, it is important to know how much energy is kept by the random modes. For the statistics with correlation length $\lambda_x = 200m$ and $\lambda_y = 100m$, the fraction of energy captured by the eigenvalues versus the number of KL modes used is shown in Fig. 3. It is shown that about 85% energy of the field is captured using the first

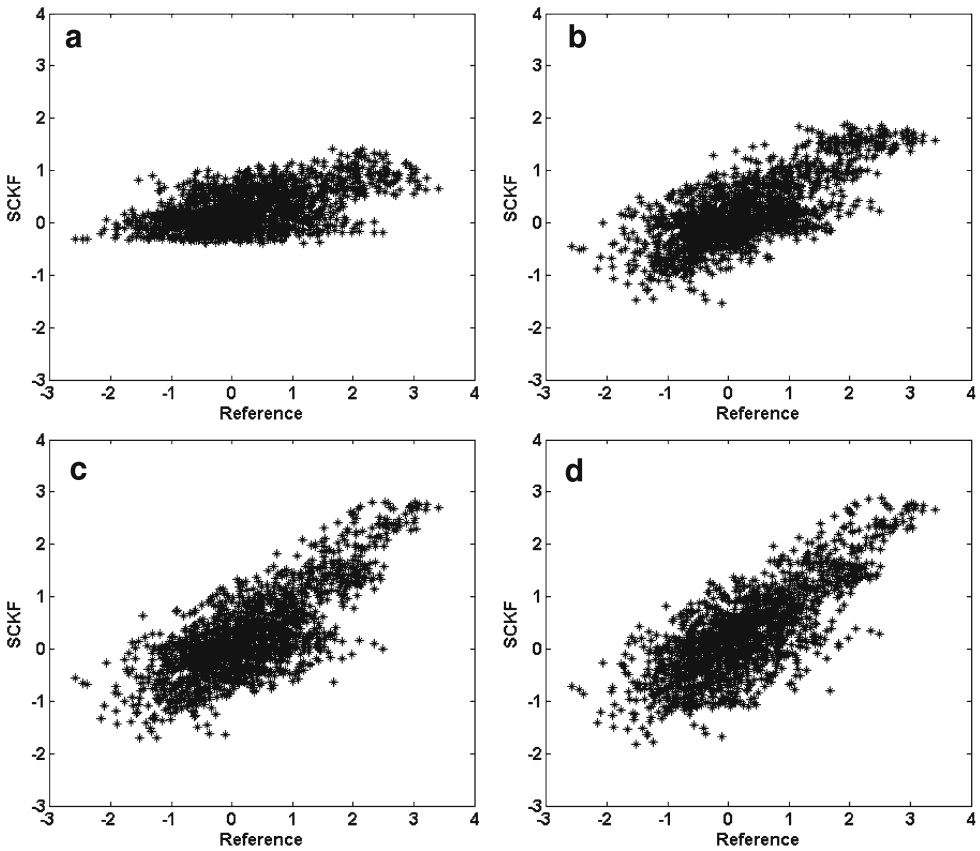

Fig. 8 Comparison between the estimated ln K_s from the SCKF with 100 modes and the reference field at different times: **a** 0.2, **b** 2.0, **c** 5.0, and **d** 10.0 day

100 modes, while 90% of energy is kept using the first 200 modes. After that, the fraction increases very slowly with the increase in the mode index.

To measure the performance of the Kalman filter, two quantities are commonly used. The root mean square error (RMSE) stands for the mean deviation of the estimated mean field from the reference field,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[E\,(Y\,(x_i)) - Y^t\,(x_i)]^2}, \tag{25}$$

where operator E is the expectation on the ensemble and $E(Y(x_i))$ is the estimated mean, $Y^t(x_i)$ is the reference values, and N is the number of grid nodes. Another measure of the performance is the ensemble spread, which is the square root of the averaged variance of the ensemble, defined as

$$\mathrm{SPREAD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathrm{VAR}\,(x_i)}. \tag{26}$$

In this study, we compare the performance of the SCKF and EnKF by contrasting their respective RMSE and SPREAD. The RMSE is used to measure the mean estimation, while the SPREAD is used to measure the uncertainty in the estimation. Here, we show that

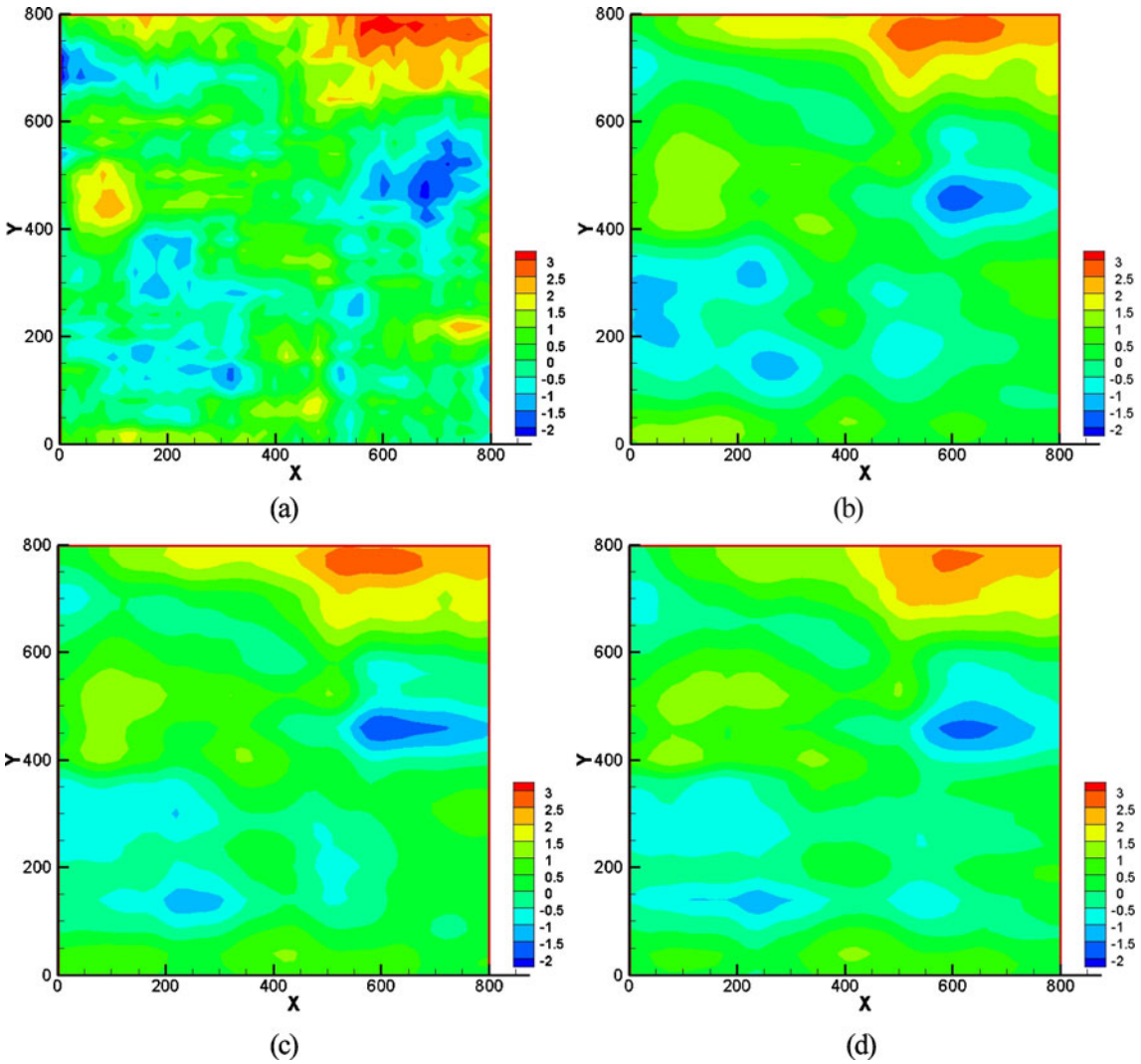

Fig. 9 The contours of ln K_s: **a** reference field, the estimated mean from **b** SCKF with 100 modes, **c** SCKF with 200 modes, and **d** EnKF with 1,000 realizations

the match between the RMSE and SPREAD provides another measure of performance. Equation 26 can be also written as

$$\text{SPREAD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\{E\left[Y(x_i)^2\right] - E\left[Y(x_i)\right]^2\right\}}. \quad (27)$$

Hence, the difference between the squares is

$$\begin{aligned}\text{RMSE}^2 - \text{SPREAD}^2 &= \frac{1}{N}\sum_{i=1}^{N}\left\{E\left[Y(x_i)\right]^2 - 2E\left[Y(x_i)\right]Y^t(x_i) + Y^t(x_i)^2 \right.\\ &\qquad \left. - E\left[Y(x_i)^2\right] + E\left[Y(x_i)\right]^2\right\}\\ &= \frac{2}{N}\sum_{i=1}^{N} E\left[Y(x_i)\right]\left\{E\left[Y(x_i)\right] - Y^t(x_i)\right\}\\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\left\{Y^t(x_i)^2 - E\left[Y(x_i)^2\right]\right\}. \quad (28)\end{aligned}$$

It is shown from Eq. 28 that if

(a) $E[Y(x)^2] \rightarrow Y^t(x)^2$, in an grid-average form, and
(b) $E[Y(x)] \rightarrow Y^t(x)$ in a weighted grid-average form,

the difference will tend to be zero. Therefore, the match between RMSE and SPREAD is a measure involved with the both mean and uncertainty estimations.

In real-world applications, since the reference field (the true field) is not available, the RMSE cannot be calculated. Furthermore, the RMSE and SPREAD may not be proper for non-Gaussian field. A more straightforward way is to see the match between the estimated pressure and the reference. The estimations are given by rerunning the simulations from the initial time with updated ensemble members (in the EnKF) or collocation realizations (in the SCKF). In order to show

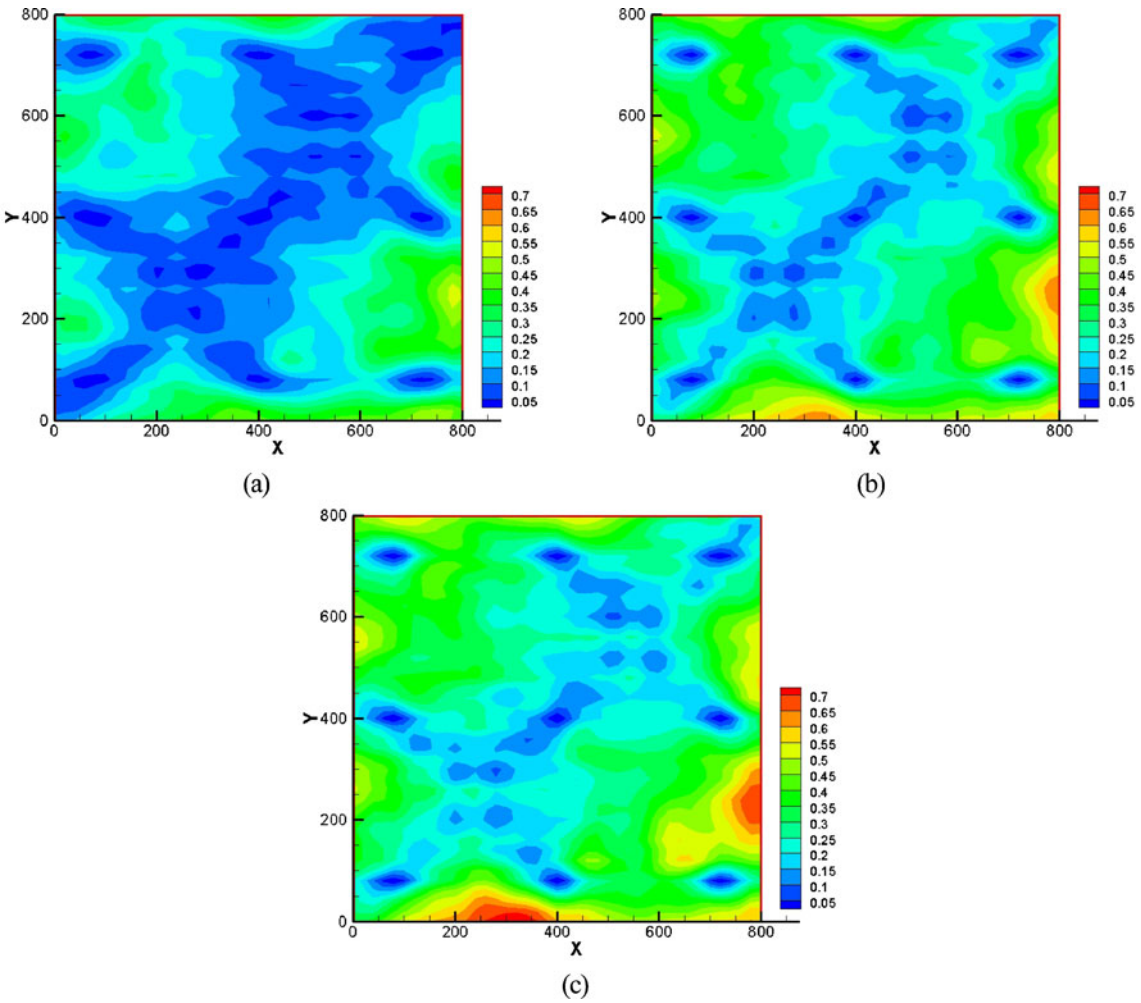

Fig. 10 The estimated ln K_s variance from **a** SCKF with 100 modes, **b** SCKF with 200 modes, and **c** EnKF with 1,000 realizations

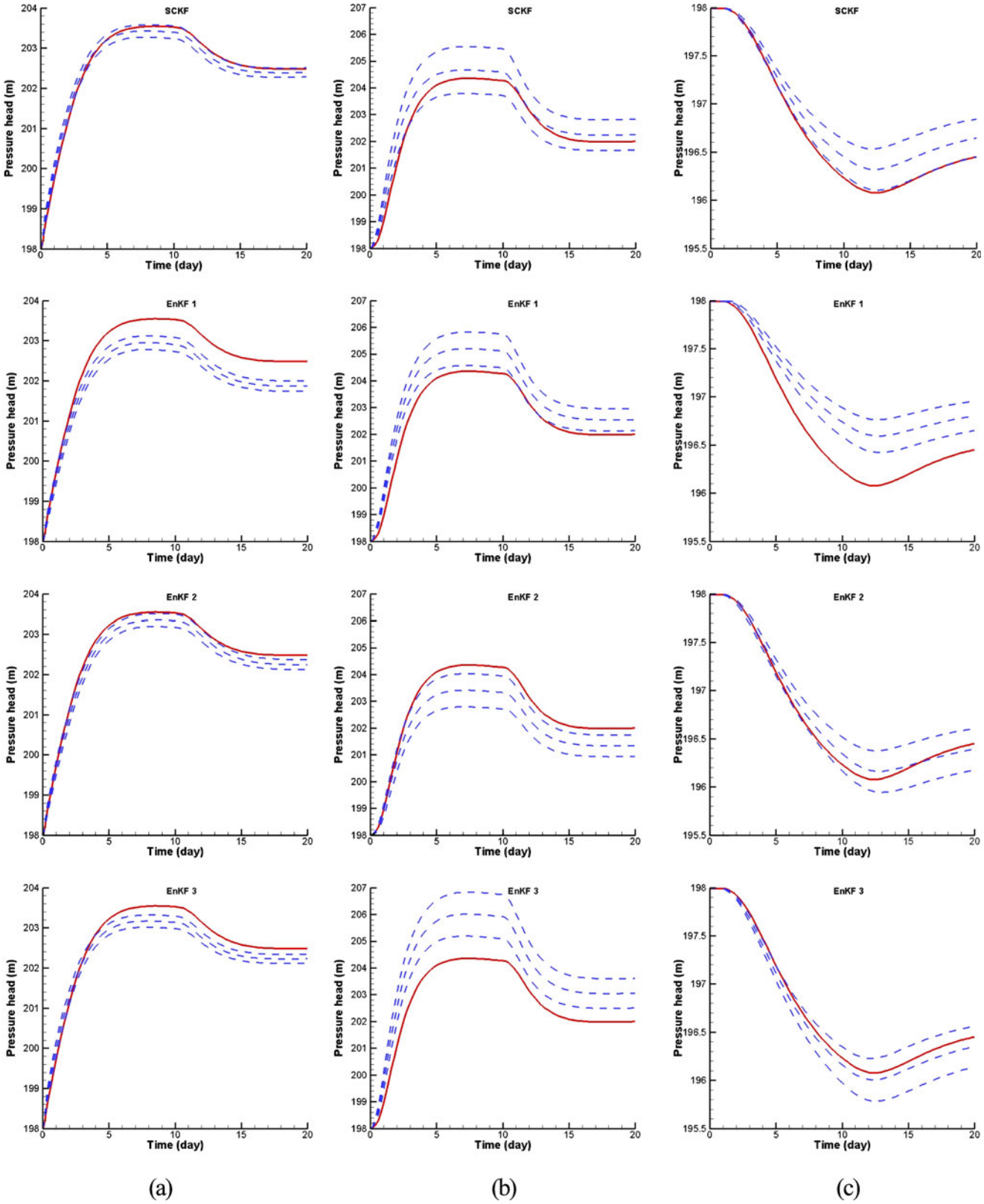

Fig. 11 Mean and confidence interval estimations of pressure calculated from the SCKF with 100 modes and different runs of the EnKF with 100 realizations, for the reference (*red lines*), for the estimations (*blue dashed lines*). From *left* to *right*, figures in column **a**, **b**, and **c** are for the location (80m and 720m), (320m and 120m), and (720m and 80m), respectively. In each column, the *top figure* is for the SCKF, the remaining three figures are for different groups of the EnKF

the predictions under different well-operating scenario, from day 10, the rates of the two wells are changed from 150 to 100. Accordingly, the flow behaviors will be significantly changed from days 10 to 20. It should also be noted that, since the inverse problems are usually ill-posed, the solutions with accuracy to a certain degree may not be unique. Similar pressure head behaviors can be produced by different conductivity fields. Even the match between the estimations and observations is not an adequate measure either.

In the SCKF, the forecast and analysis are implemented in the PCE basis space. Each state is decomposed into $(Q+1)$ PCE components; hence, $(Q+1)$ collocation realizations are required. In the EnKF, the forecast and analysis are implemented in the ensemble space. Each state has N_e Monte Carlo realizations. In both of the SCKF and the EnKF, the main computational efforts are used to solve the flow equation. The SCKF with random dimensionality M and PCE order d requires solving the flow equation $(M+d)!/(M!d!)$ times. The EnKF with ensemble size N_e requires solving the flow equation N_e times. The challenge exists for both the SCKF and the EnKF in large-scale problems. Due to the computational burden, the ensemble size or the number of collocation realizations is usually several orders of magnitude smaller than the problem dimension, i.e., the grid number in this study. Both the EnKF and the SCKF use reduced rank approximation of the system covariance matrix. The updates are also restricted to the subspace spanned by the forecast ensemble members or PCE coefficients. Furthermore, the sampling errors in the EnKF with limited ensemble size may be dominant. As introduced in Section 1, in the EnKF, many methods have been proposed to reduce the necessary ensemble size requirement. In the SCKF, the computational efforts are determined by the random dimensionality and the PCE order. The KL expansion is used as a dimension reduction tool to parameterize the random field. The random dimensionality is then reduced from the number of grids to the number of random variables in the KL expansion, i.e., M. For large-scale problems where the total number of grids may be of order 10^5, an M of the order 10^{2-3} is usually sufficient for parameterization. For strongly nonlinear problems, higher PCE order is necessary. It seems that in a strongly nonlinear problem with a large random dimensionality, the number of required collocation realizations is so huge that the SCKF may become

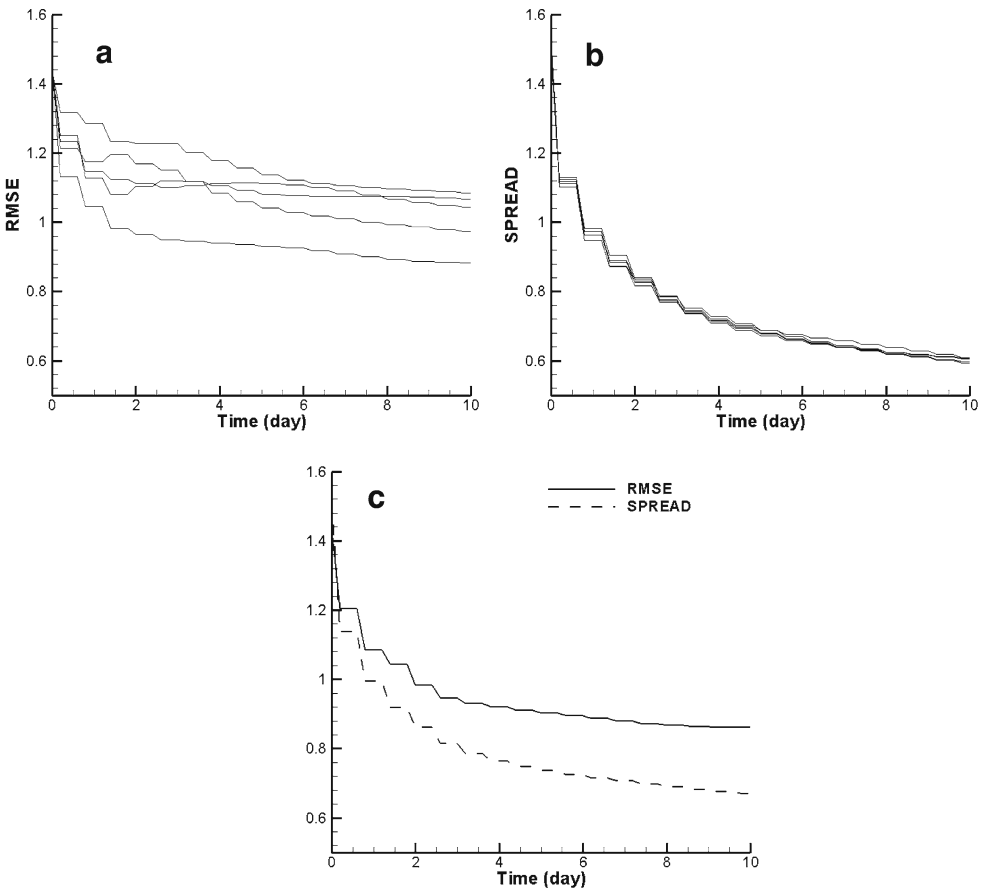

Fig. 12 With the reference variance $\sigma^2 = 2$: **a** RMSE for EnKF with 200 realizations, **b** SPREAD for EnKF with 200 realizations, and **c** RMSE and SPREAD for SCKF with 200 modes

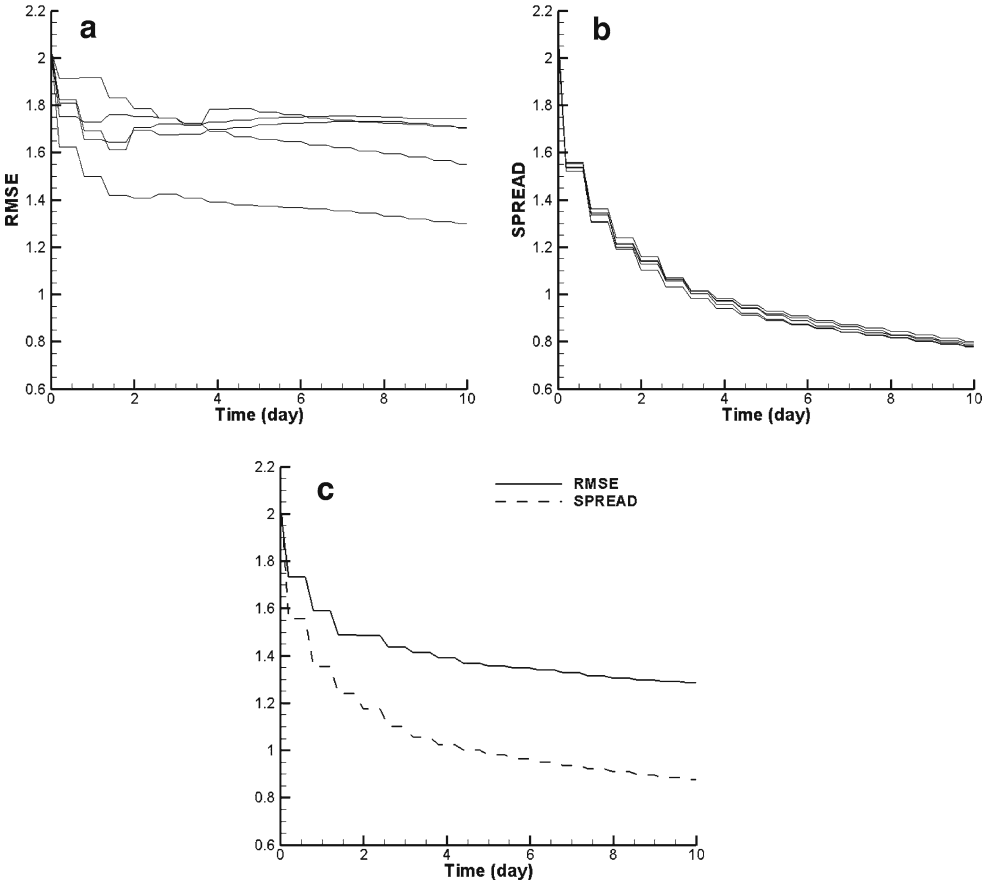

Fig. 13 With the reference variance $\sigma^2 = 4$: **a** RMSE for EnKF with 200 realizations, **b** SPREAD for EnKF with 200 realizations, and **c** RMSE and SPREAD for SCKF with 200 modes

impractical. This is the so-called curse of dimensionality. However, since not all the PCE terms make significant contributions in the system response, one possible way to reduce the computational burden in PCE-based methods is to use the leading PCE terms. For example, in the recent work on stochastic analysis of unsaturated flow with probabilistic collocation method [37], a case with the random dimensionality $M = 150$ and PCE order $d = 2$ was studied. The number of full PCE terms was $(150 + 2)!/(150!2!) = 11{,}476$, which makes the PCM computationally expensive. However, it has been shown that the cross PCE terms can be neglected, and the number of the remained PCE terms is reduced to $(1 + 2\ M)$, i.e., only 301. This leading term approximation makes the second order PCM still more efficient than Monte Carlo simulation in that nonlinear problem.

In this study, the stochastic collocation method based on the Stroud-2 rule is employed. M modes requires solving the corresponding deterministic equations for $M + 1$ time. Therefore, the computational efficiency of the two approaches can be compared by just looking at the number of modes in the SCKF and the number of realizations in the EnKF.

5.1 Initial statistics

Owing to the fact that the statistics of the conductivity field are commonly known with an incomplete knowledge, the SCKF is initialized with slightly different

Table 1 With the reference variance $\sigma^2 = 2$, the final RMSE, SPREAD, and the difference of squares for EnKF with 200 realizations and SCKF with 200 modes

	EnKF					SCKF
RMSE	1.037	0.995	0.933	0.907	0.954	0.861
SPREAD	0.654	0.648	0.655	0.644	0.644	0.670
R2 – S2	0.647	0.570	0.441	0.408	0.495	0.292

statistics. The mean of the log hydraulic conductivity is still zeros, while the variance is set as $\sigma^2 = 1.4$. The correlation lengths are specified as $\lambda_x = 220m$ and $\lambda_y = 120m$. The RMSE and SPREAD of SCKF with different number of modes are shown in Fig. 4. It is shown that the RMSE decreases with time, which means that the estimated mean tends to the reference value when more observations are incorporated. The SCKF with 200 modes gives the lowest RMSE. The SCKF with 100 modes gives a similar RMSE compared to SCKF with 200 modes, which means that the performance of the SCKF is not improved much beyond 100 modes in terms of RMSE. However, the SPREAD of the SCKF with 200 modes matches with the RMSE much better than with 100 modes, indicating that the ability to estimate the uncertainty still improves beyond 100 modes.

For comparison, 1,000 unconditional realizations of the log hydraulic conductivity field are generated by the KL expansion by keeping a larger number (400) of modes. The initial ensemble is divided into one, five, and ten groups to perform the EnKF simulations. The RMSE and SPREAD of EnKF with different number of realizations are shown in Figs. 5, 6, and 7.

If the ensemble size is small, the EnKF is far from convergence and is therefore realization dependent. The RMSE shows a large variation among different groups for the EnKF with 100 realizations (Fig. 5). The variation can be reduced using more realizations, which indicates a better convergence, as shown in Fig. 6 for the EnKF with 200 realizations. The EnKF with 1,000 realizations gives the lowest RMSE, while the similar value can be obtained by the SCKF with 200 modes. It is seen from Figs. 5 and 6 that with the same number of realizations, the SPREAD declines similarly for different EnKF groups but systematically underestimates the RMSE. Using more realizations, the SPREAD shows a better match with the RMSE. In Fig. 7, the EnKF with 1,000 realizations shows a very good match between the RMSE and the SPREAD. It is worthwhile noting that a similar match is given by the SCKF with 200 modes, as shown in Fig. 4. Since the SCKF with 200 modes requires 201 solutions of the flow equation while the EnKF with 1,000 realizations requires 1,000 solutions, the computational burden is greatly reduced in the SCKF.

To give a dynamic illustration of the data assimilation process, the reference field and the mean estimated log hydraulic conductivity field from the SCKF with 100 modes at different times are compared in Fig. 8. For a good estimate, all points should locate near the diagonal line. It is shown that, with the sequentially incorporated observations, the estimated field tends closer to the reference field.

The contours of the reference field and the mean of log hydraulic conductivity field estimated from the SCKF with 100 and 200 modes and from the EnKF with 1,000 realizations are plotted in Fig. 9. It is shown that all these three filters can identify the main patterns of the reference field. We also notice that the patterns along the main flow direction between the two wells (two triangles) are estimated better. The estimated variances are plotted in Fig. 10. The variances are the lowest at the conductivity measurement location. We can regard the EnKF with 1,000 realizations as a reference. It is shown that, although the SCKF with 100 modes provides a good mean estimation (Fig. 9), it underestimates the uncertainty (Fig. 10a). The SCKF with 200 modes gives a similar estimation of the variance field compared to that estimated by the EnKF with 1,000 realizations. Therefore, 200 modes should be used if the purpose is to estimate the uncertainty.

By rerunning the simulations with updated ensemble members or collocation realizations from the initial time, we can obtain the statistic moments (e.g., mean and variance) of the pressure, based on which the confidence intervals can then be constructed. In the EnKF, the mean and standard deviation can be calculated directly from the ensemble. In the SCKF, the mean estimations and the standard deviations are given by the zeroth and higher-order PCE coefficients, respectively. The confidence intervals of the pressure are shown in Fig. 11. From left to right, each column is for the pressure at the measurement location ($80m$ and $720m$), ($320m$ and $120m$), and ($720m$ and $80m$), respectively. The top row is for the SCKF with 100 modes. The remaining three rows are for different groups of EnKF with 100 realizations. The central dashed blues lines are

Table 2 With the reference variance $\sigma^2 = 4$, the final RMSE, SPREAD, and the difference of squares for EnKF with 200 realizations and SCKF with 200 modes

	EnKF					SCKF
RMSE	1.561	1.668	1.445	1.449	1.461	1.287
SPREAD	0.842	0.860	0.870	0.853	0.841	0.875
R2 – S2	1.728	2.043	1.331	1.372	1.427	0.891

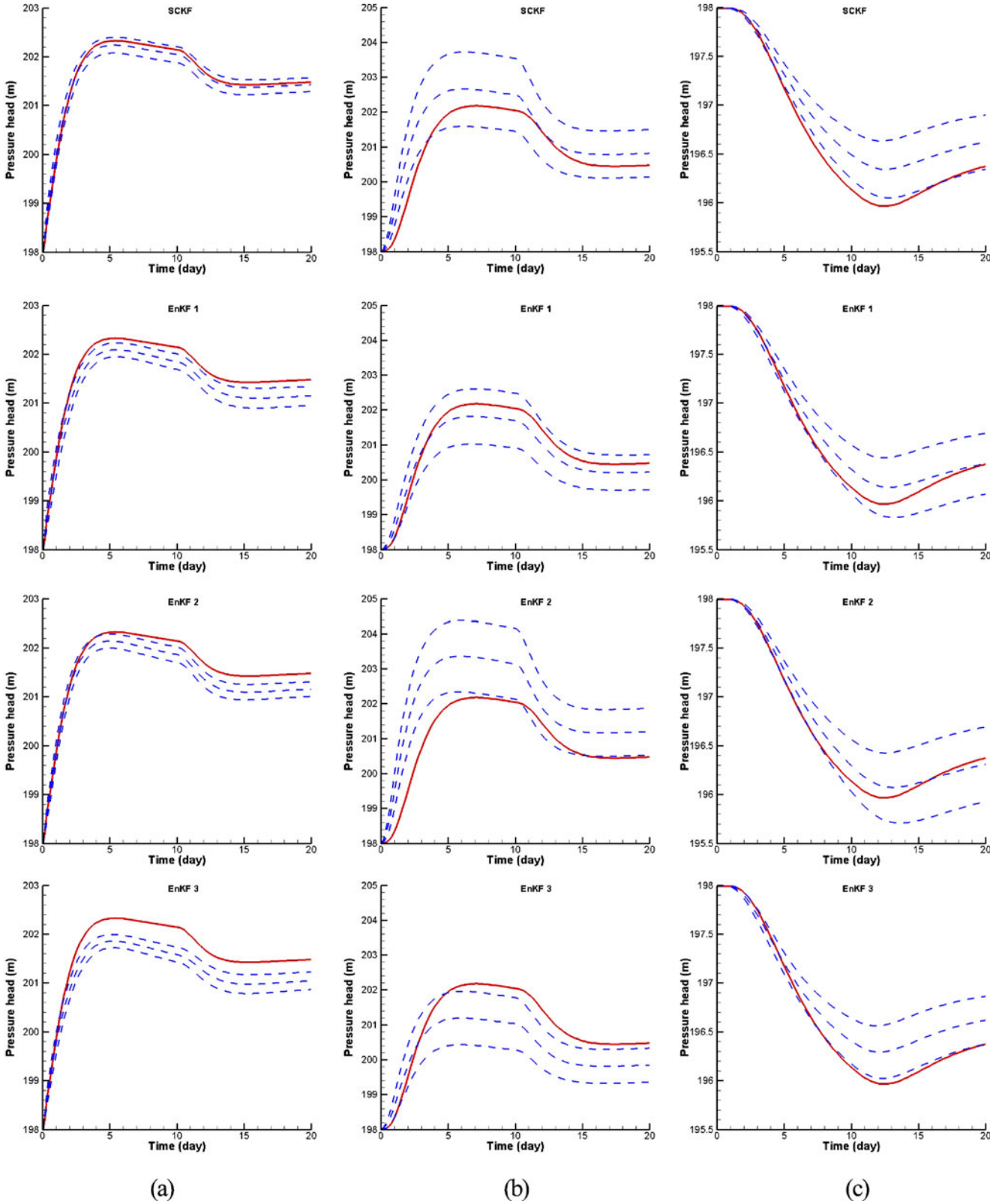

Fig. 14 With the reference variance $\sigma^2 = 2$, mean and confidence interval estimations of pressure calculated from the SCKF with 200 modes and different runs of the EnKF with 200 realizations, for the reference (*red lines*), for the estimations (*blue dashed lines*). From *left* to *right*, figures in column **a**, **b**, and **c** are for the location (80m and 720m), (320m and 120m) and (720m and 80m), respectively. In each column, the *top figure* is for the SCKF, the remaining three figures are for different groups of the EnKF

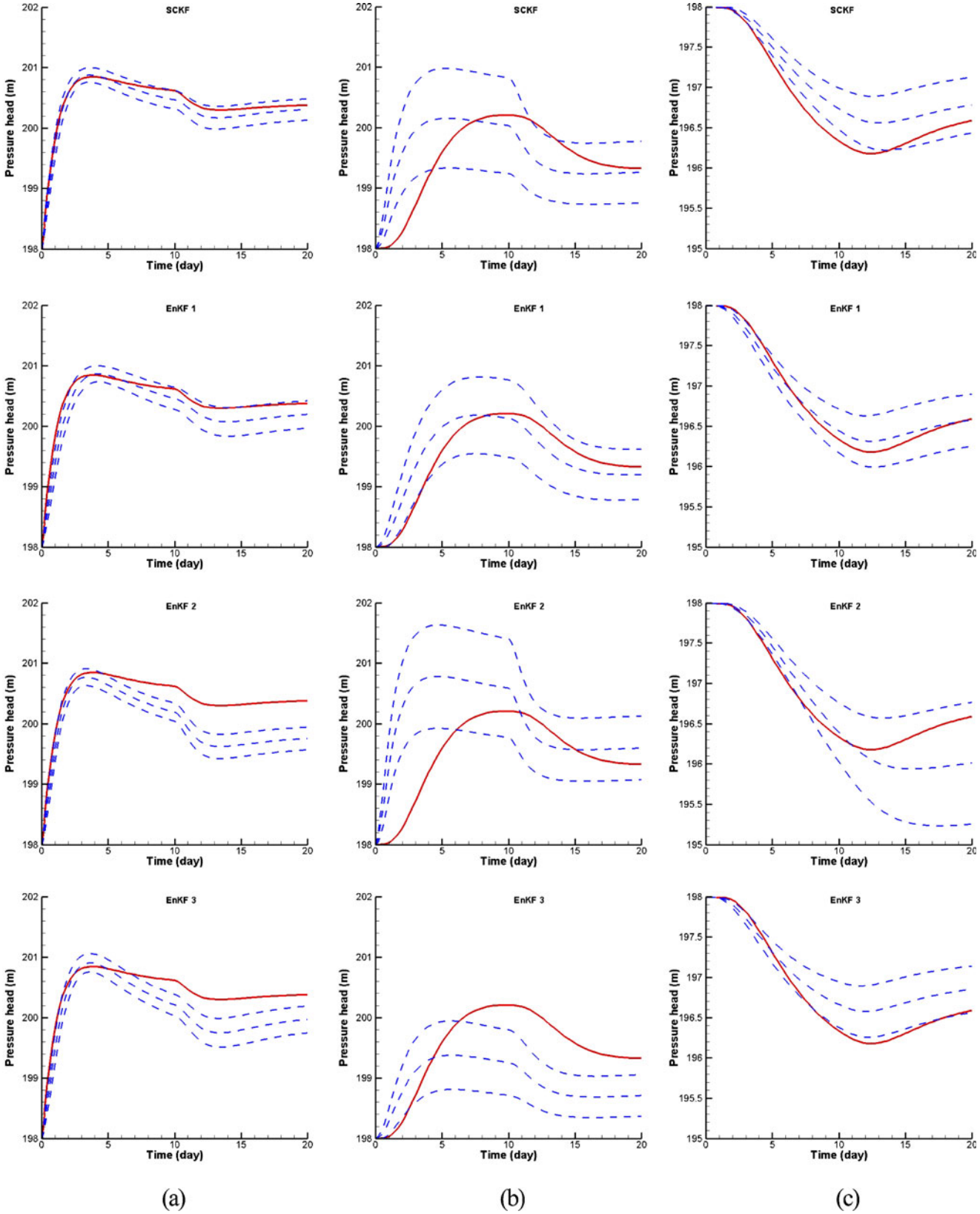

Fig. 15 With the reference variance $\sigma^2 = 4$, mean and confidence interval estimations of pressure calculated from the SCKF with 200 modes and different runs of the EnKF with 200 realizations, for the reference (*red lines*), for the estimations (*blue dashed lines*). From *left* to *right*, figures in column **a**, **b**, and **c** are for the location (80*m* and 720*m*), (320*m* and 120*m*), and (720*m* and 80*m*), respectively. In each column, the *top figure* is for the SCKF, the remaining three figures are for different groups of the EnKF

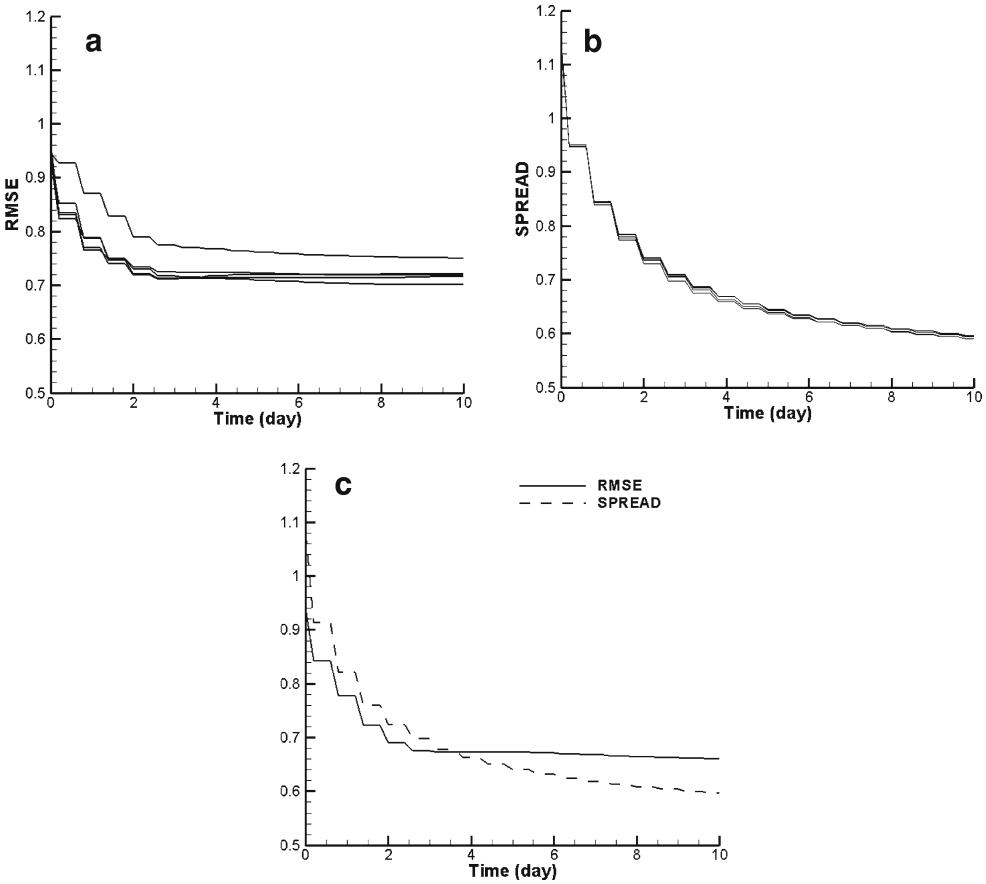

Fig. 16 With the reference correlation lengths $\lambda_x = 80m$ and $\lambda_y = 80m$: **a** RMSE for EnKF with 200 realizations, **b** SPREAD for EnKF with 200 realizations, and **c** RMSE and SPREAD for SCKF with 200 modes

for mean estimations. For good mean estimations, the central blue lines should be close to red lines (reference). The top and bottom dashed blue lines are mean plus and minus one standard deviation, respectively. The interval between the two lines represents the uncertainty to some degree of accuracy. For the Gaussian distribution, the interval is of 68.3% confidence. Although, in general, it is known that the assimilation is needed whenever significant flow behavior changes occur, such as adding new wells in the reservoir, well-rate changing, and other operations, in this specific case, it is shown that both the SCKF and the EnKF are able to predict the trends of pressure heads from days 10 to 20. It is also shown that, compared to the EnKF, the SCKF generally gives estimations with better matches to the reference. Many of the estimations from the EnKFs still show relatively large deviations from the reference. Therefore, in terms of pressure estimation, the SCKF performs better than the EnKF.

5.2 Large variance

In order to test the performance of the SCKF in the presence of large log conductivity variances, two cases are studied in this subsection. The references are with variance $\sigma^2 = 2$ and $\sigma^2 = 4$, corresponding to the coefficient of variation of 253% and 732% for the hydraulic conductivity, respectively. The initial variance are set as $\sigma^2 = 2.5$ and $\sigma^2 = 4.8$, respectively. All the other parameters are the same as those in the previous case. For the two cases, the RMSE and SPREAD of the EnKF with 200 realizations and the SCKF with 200 modes are shown in Figs. 12 and 13, respectively. The final RMSE, SPREAD, and the difference of the

Table 3 With the reference correlation lengths $\lambda_x = 80m$, $\lambda_y = 80m$, the final RMSE, SPREAD, and the difference of squares for EnKF with 200 realizations and SCKF with 200 modes

	EnKF					SCKF
RMSE	0.719	0.701	0.717	0.750	0.721	0.660
SPREAD	0.596	0.590	0.594	0.591	0.597	0.596
R2 - S2	0.162	0.143	0.161	0.213	0.163	0.080

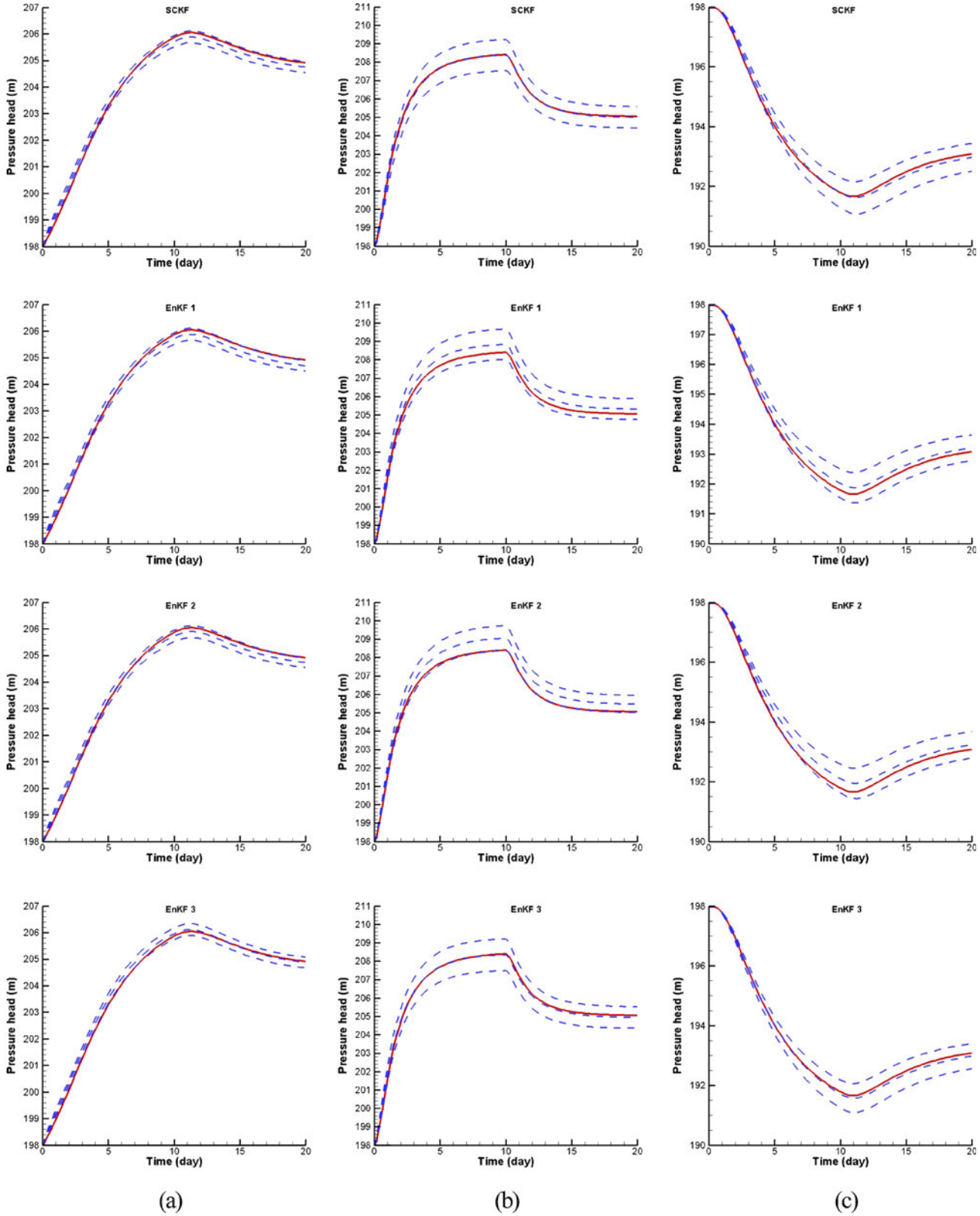

Fig. 17 With the reference correlation lengths $\lambda_x = 80m$ and $\lambda_y = 80m$, mean and confidence interval estimations of pressure calculated from the SCKF with 200 modes and different runs of the EnKF with 200 realizations, for the reference (*red lines*), for the estimations (*blue dashed lines*). From *left* to *right*, figures in column **a**, **b**, and **c** are for the location ($80m$ and $720m$), ($320m$ and $120m$), and ($720m$ and $80m$), respectively. In each column, the *top figure* is for the SCKF, the remaining three figures are for different groups of the EnKF

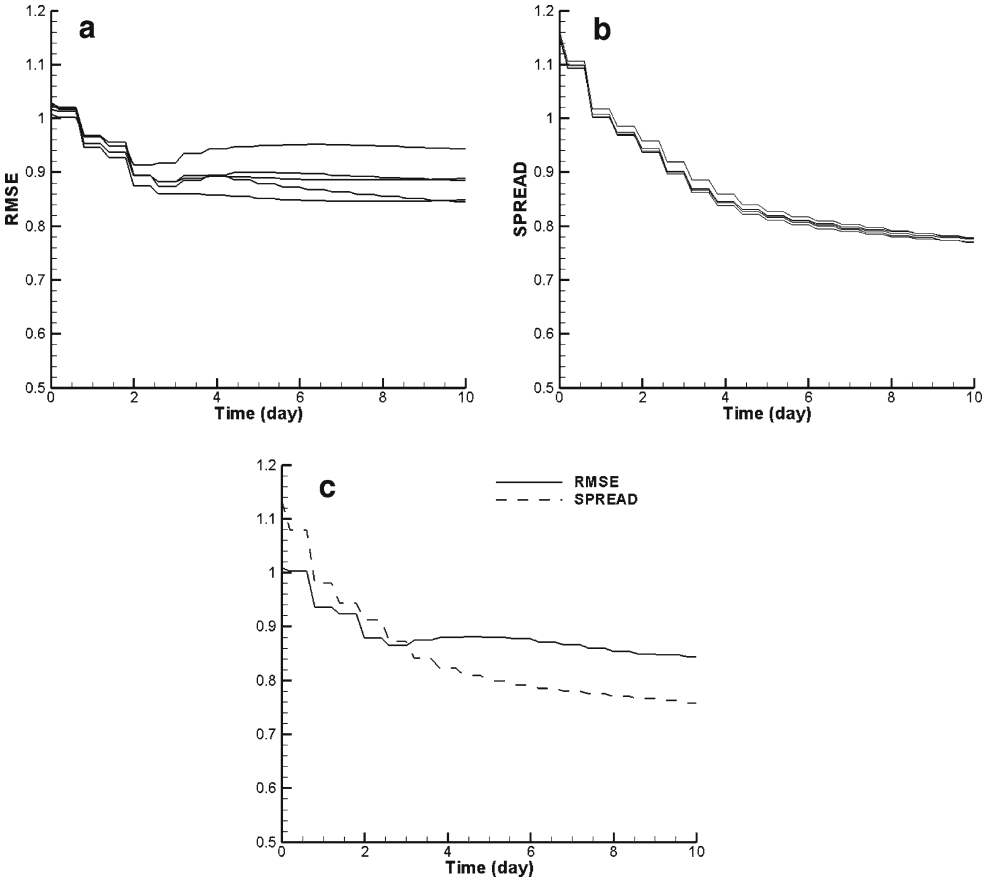

Fig. 18 With observations at nine *filled squares*: **a** RMSE for EnKF with 200 realizations, **b** SPREAD for EnKF with 200 realizations, and **c** RMSE and SPREAD for SCKF with 200 modes

squares are shown in Tables 1 and 2. It is shown that the RMSE of SCKF is lower than those given by the EnKF. It is seen that, in both methods, the SPREAD underestimates the RMSE. However, compared to the EnKF, the SCKF still provides a better match between the RMSE and the SPREAD.

The estimated pressure heads from day 0 to 20 with $\sigma^2 = 2$ and $\sigma^2 = 4$ are plotted in Figs. 14 and 15, respectively. It is shown that, for $\sigma^2 = 2$, compared to the EnKF, the SCKF estimates the pressure with comparable or even better accuracy. For $\sigma^2 = 4$, all the estimations of pressure heads given by the EnKF and SCKF show larger deviations from the reference due to the larger prior uncertainty. At location (a), the SCKF estimates the pressure better than do the EnKFs. While at the other two locations, the SCKF performs similarly to the EnKF. Therefore, in terms of pressure estimation, the superiority of the SCKF over the EnKF decreases with the increase of variance. It is noted that in Fig. 15c, the mean estimation given by the second EnKF match with the observations well during the assimilation period (from day 0 to 10). However, it starts to deviate from the reference in the prediction period (from day 10 to 20). It is a demonstration of non-uniqueness of solutions in the inverse problems.

5.3 Correlation length

In the implementation of SCKF, the random log hydraulic conductivity field is parameterized by the KL

Table 4 With observations at nine filled squares, the final RMSE, SPREAD, and the difference of squares for EnKF with 200 realizations and SCKF with 200 modes

	EnKF					SCKF
RMSE	0.943	0.885	0.888	0.848	0.845	0.843
SPREAD	0.779	0.771	0.770	0.776	0.777	0.758
R2 – S2	0.282	0.189	0.196	0.117	0.110	0.136

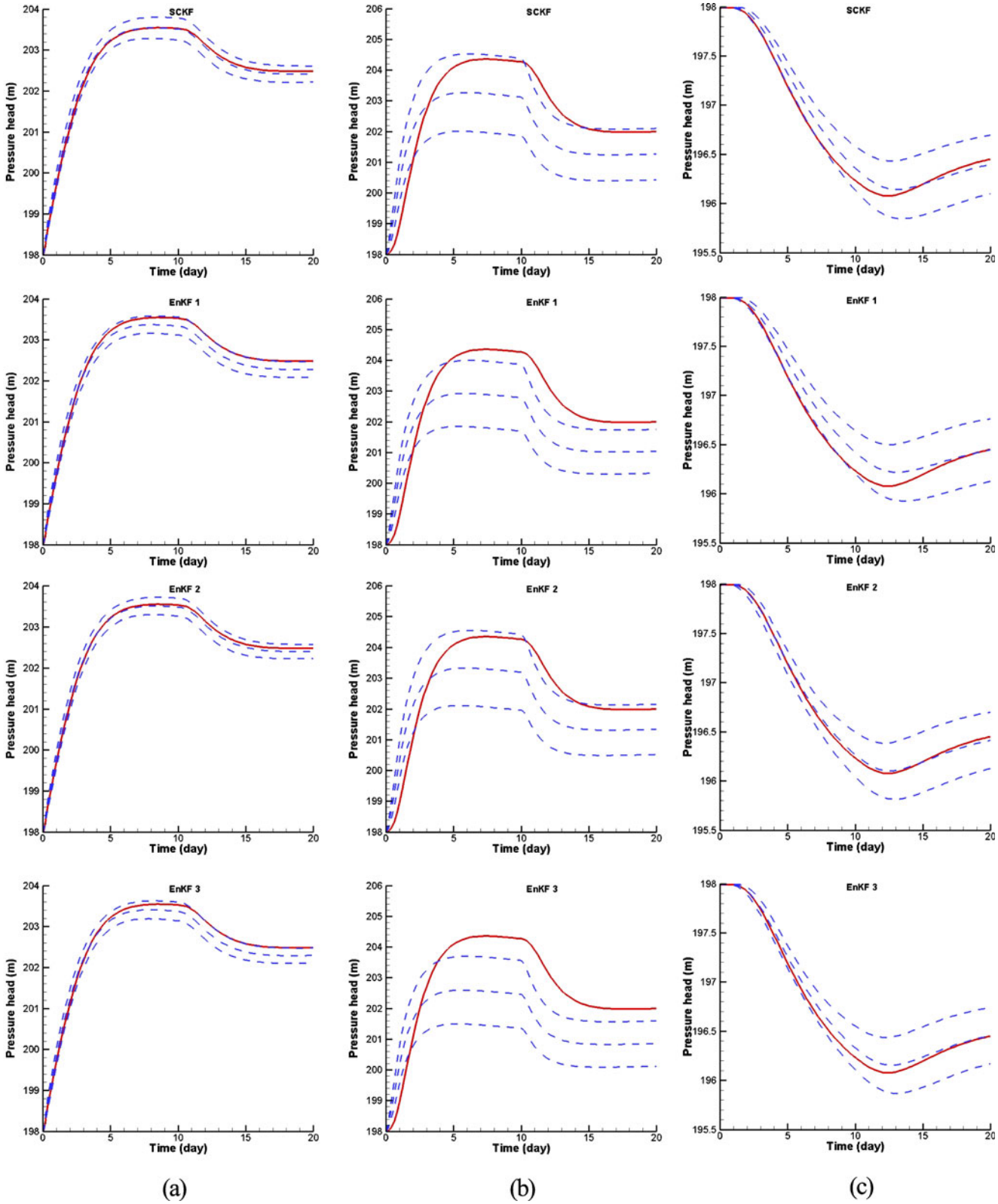

Fig. 19 With observations at nine *filled squares*, mean and confidence interval estimations of pressure calculated from the SCKF with 200 modes and different runs of the EnKF with 200 realizations, for the reference (*red lines*), for the estimations (*blue dashed lines*). From *left* to *right*, figures in column **a**, **b**, and **c** are for the location (80m and 720m), (320m and 120m), and (720m and 80m), respectively. In each column, the *top figure* is for the SCKF, the remaining three figures are for different group of the EnKF

expansion with independent standard Gaussian random variables. Larger number of expansion terms is required to describe the random field with the decrease of the correlation ratio (correlation length to the domain size). In this subsection, the correlation lengths of the reference field are set as $\lambda_x = 80m$ and $\lambda_y = 80m$. The initial setup for both the SCKF and the EnKF is $\lambda_x = 100m$ and $\lambda_y = 100m$. All the other parameters are chosen as the same as those in Section 5.1. The RMSE and SPREAD for the EnKF with 200 realizations and the SCKF with 200 modes are shown in Fig. 16.

In this case, by retaining 200 modes, only 81% of energy is kept in the KL expansion. Because of this, the RMSE performance of the SCKF with 200 modes is not so good as that in Section 5.1. The final RMSE, SPREAD, and the difference of the squares are shown in Table 3. It is shown that the SCKF gives a lower RMSE than do most of the EnKF. The pressures estimated by the two methods are plotted in Fig. 17. It is noted that, although the random field is less sufficiently parameterized, the pressure heads estimated by the SCKF are as good as or slightly better than the EnKFs.

In the single-phase flow problem studied in this paper, for the case with correlation ratio 1:4 and 1:8 in two directions, 200 modes in KL expansion seem to be enough to parameterize the random field. The SCKF performs better than the EnKF with the similar computational efforts. With the decrease in correlation ratio, the performance of SCKF with the fixed number of KL modes will decrease. As shown in this case with ratio 1:10 in both directions, although the SCKF still performances slightly better than do most of the EnKFs, the superiority decreases.

5.4 Number of observations

In order to test the performance of the SCKF with fewer observations, we assume that at $t = 0.2$ day, all the observations of the hydraulic conductivity and the hydraulic pressure head are only measured at the nine filled squares. After that, the hydraulic pressure heads are measured at the nine filled squares at every 0.6 day up to day 10. All the other parameters are the same as those in Section 5.1. The RMSE and SPREAD for the EnKF with 200 realizations and the SCKF with 200 modes are plotted in Fig. 18. In this case, owing to the fact that less information is provided in the observations, most RMSEs of the EnKF and the SCKF exhibit oscillations. The final RMSE, SPREAD, and the difference of the squares are shown in Table 4. It can be shown that, although the SCKF still gives a slightly lower RMSE value than do most of the EnKF, this superiority is not obvious. In terms of the match between the RMSE and the SPREAD, the EnKF and the SCKF also perform similarly. The estimated pressures given by the SCKF and EnKF are plotted in Fig. 19. It is also shown that, although the SCKF still estimates the pressure with comparable or slightly better accuracy compared to the EnKF, the superiority decreases.

6 Conclusions

In this study, a stochastic collocation-based Kalman filter is developed to estimate the hydraulic conductivity field from hydrologic observations. The conductivity field is treated as a random field and parameterized by the Karhunen–Loeve expansion. The pressure head field is approximated by the polynomial chaos expansion. With given collocation point sets, each realization is forwarded in time via deterministic solver independently, similar to the implementation of the EnKF. The coefficients of the polynomial chaos expansion are obtained in the post-processing step of the stochastic collocation method. Once the observations are available, both the conductivity field and the pressure head field are updated through updating the PCE coefficients. Compared to other stochastic methods such as Karhunen–Loeve-based moment equation method and stochastic Galerkin PCE methods, the stochastic collocation method is non-intrusive in that it results in the same governing equations and only requires repetitive runs of existing deterministic solver.

A 2D single-phase flow example is used to demonstrate the applicability of the SCKF. The results are compared with those of the EnKF. With respect to the performance, different factors, including the initial guess, the variance, the correlation length, and the number of observations, are discussed. It is shown that the SCKF is more efficient than the EnKF under certain conditions. For the correlation ratio 1:4 and 1:8 in two directions, the SCKF gives satisfactory estimations of the hydraulic conductivity field even when the spatial variance of log conductivity is as large as 4.0 (i.e., the coefficient of variation of hydraulic conductivity being 732%). It is also shown that the superiority of the SCKF over the EnKF decreases with the larger variance, the shorter correlation ratio, and the fewer number of observations. Therefore, there should be a critical point on which the choice between the two methods can be decided. This critical point is, however, problem dependent. To improve the ability of the SCKF in strongly nonlinear problems with high random dimensionality, studies about using the leading PCE approximation, localizations, and inflation need to be investigated in future work.

Since the SCKF is a non-intrusive method, it can be directly utilized in complex or nonlinear problems such as multiphase flow problems. The forward modeling of multiphase flow with a stochastic collocation method has been reported by our group recently [32]. It should be noted that, although the Kalman filter can only be utilized in nonlinear problems where the basic Gaussian assumption is not strongly violated, the stochastic collocation method itself is able to describe the uncertainties with arbitrary distributions, as long as the approximation order is high enough. Although the stochastic collocation method in this study is based on the Stroud-2 rule, which is up to the first order of PCE, it can be extended to the higher order in a similar manner.

Acknowledgements This work is partially funded by National Science Foundation through grant DMS-0801425, National Natural Science Foundation of China through grant 50688901, and the Chinese National Basic Research Program through grant 2006CB705800.

Appendix

In this appendix, we try to demonstrate the equivalence between the SCKF with accuracy up to first order PCE and the EnKF with non-perturbed observations using the sampling method based on Stroud-2 rule.

Let $\mathbf{s}(\xi_i)$, $i = 0, 1, \ldots M$ be the realizations corresponding to the Stoud-2 point set ξ_i, i.e.,

$$\mathbf{s}\left(\xi_i\right) = \sum_{j=0}^{M} \mathbf{c}_j \Psi_j\left(\xi_i\right), i = 0, 1, \ldots M. \tag{29}$$

Since every realization is with equal weight, we can update each realization directly,

$$\mathbf{s}^a\left(\xi_i\right) = \mathbf{s}^f\left(\xi_i\right) + \mathbf{K}\left[\mathbf{d}_{\text{obs}} - \mathbf{H}\mathbf{s}^f\left(\xi_i\right)\right], i = 0, 1, \ldots, M. \tag{30}$$

In SCKF, let $\mathbf{c}_j$ and $j = 0, 1, \ldots, M$ be the PCE terms up to first order, the updating step is

$$\mathbf{c}_j^a = c_j^f + \mathbf{K}\left[\mathbf{d}_{\text{obs}}\delta_{0j} - \mathbf{H}\mathbf{c}_j^a\right], \; j = 0, 1, \ldots, M, \tag{31}$$

where δ_{0j} is Kronecker delta. We have to prove that Eq. 30 is equivalent to Eq. 31.

(a) Multiplying $\Psi_j\left(\xi_i\right)$ to the both sides of Eq. 31 and taking summation with respect to j, we have

$$\sum_{j=0}^{M} \mathbf{c}_j^a \Psi_j\left(\xi_i\right) = \sum_{j=0}^{M} \mathbf{c}_j^f \Psi_j\left(\xi_i\right) + \mathbf{K}\left[\mathbf{d}_{\text{obs}} - \mathbf{H}\sum_{j=0}^{M} \mathbf{c}_j^f \Psi_j\left(\xi_i\right)\right], \quad i = 0, 1, , M. \tag{32}$$

According to Eqs. 29 and 32 yields Eq. 30.

(b) Multiplying to the both sides of Eq. 30 and taking summation with respect to i, we have

$$\frac{1}{(M+1)}\sum_{j=0}^{M} \mathbf{s}^a(\xi_i)\Psi_j(\xi_i) = \frac{1}{(M+1)}\sum_{j=0}^{M} \mathbf{s}^f(\xi_i)\Psi_j(\xi_i) + \ldots \frac{1}{(M+1)}\left[\sum_{j=0}^{M} \mathbf{d}_{\text{obs}}\Psi_j(\xi_i) - \mathbf{H}\sum_{j=0}^{M} \mathbf{s}^f(\xi_i)\Psi_j(\xi_i)\right], \quad j = 0, 1, \ldots, M. \tag{33}$$

Since ξ_i is Stoud-2 point set and Ψ_j is up to the first order, we have

$$\frac{1}{(M+1)}\sum_{j=0}^{M} \mathbf{s}(\xi_i)\Psi_j(\xi_i) = \langle \mathbf{s}\Psi_j \rangle = \mathbf{c}_i, \tag{34}$$

where $\langle\rangle$ is the expectation operator. Since $\mathbf{d}_{\text{obs}}$ are independent to the deviations of system states, we have

$$\frac{1}{(M+1)}\sum_{j=0}^{M} \mathbf{d}_{\text{obs}}\Psi_j(\xi_i) = \langle \mathbf{d}_{\text{obs}}\Psi_j \rangle = \mathbf{d}_{\text{obs}}\langle \Psi_j \rangle = \mathbf{d}_{\text{obs}}\delta_{0,j} \tag{35}$$

According to Eqs. 34 and 35, Eq. 33 yields Eq. 31. This completes the proof.

References

1. Dagan, G.: Flow and Transport in Porous Formations. Springer, New York (1989)
2. Gelhar, L.W.: Stochastic Subsurface Hydrology. Prentice-Hall, Englewood Cliffs (1993)
3. Zhang, D.X.: Stochastic Methods for Flow in Porous Media: Coping with Uncertainties. Academic, San Diego (2002)
4. Gelb, A.: Applied Optimal Estimation. MIT Press, Cambridge (1974)
5. Ljung, L.: Asymptotic-behavior of the extended kalman filter as a parameter estimator for linear-systems. IEEE Trans. Autom. Control. **24**, 36–50 (1979)
6. Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using monte-carlo methods to forecast error statistics. J. Geophys. Res. **99**, 10143–10162 (1994)

7. Pham, D.T., Verron, J., Roubaud, M.C.: A singular evolutive extended Kalman filter for data assimilation in oceanography. J. Mar. Syst. **16**, 323–340 (1998)
8. McLaughlin, D.: An integrated approach to hydrologic data assimilation: interpolation, smoothing, and filtering. Adv. Water Resour. **25**, 1275–1286 (2002)
9. Chen, Y., Zhang, D.X.: Data assimilation for transient flow in geologic formations via ensemble Kalman filter. Adv. Water Resour. **29**, 1107–1122 (2006)
10. Naevdal, G., Johnsen, L.M., Aanonsen, S.I., Vefring, E.H.: Reservoir monitoring and continuous model updating using ensemble Kalman filter. SPE J. **10**, 66–74 (2005)
11. Liu, N., Oliver, D.S.: Ensemble Kalman filter for automatic history matching of geologic facies. J. Pet. Sci. Eng. **47**, 147–161 (2005)
12. Gu, Y.Q., Oliver, D.S.: History matching of the PUNQ-S3 reservoir model using the ensemble Kalman filter. SPE J. **10**, 217–224 (2005)
13. Zafari, M., Reynolds, A.C.: Assessing the uncertainty in reservoir description and performance predictions with the ensemble Kalman filter. SPE J. **12**, 382–391 (2007)
14. Whitaker, J.S., Hamill, T.M.: Ensemble data assimilation without perturbed observations. Mon. Weather Rev. **130**, 1913–1924 (2002)
15. Houtekamer, P.L., Mitchell, H.L.: Data assimilation using an ensemble Kalman filter technique. Mon. Weather Rev. **126**, 796–811 (1998)
16. Evensen, G.: The ensemble Kalman filter for combined state and parameter estimation. IEEE Control. Syst. Mag. **29**, 83–104 (2009)
17. Ghanem, R., Spanos, P.: Stochastic Finite Element. A spectral approach. Springer, New York (1991)
18. Xiu, D.B., Karniadakis, G.E.: The Wiener–Askey polynomial chaos for stochastic differential equations. Siam. J. Sci. Comput. **24**, 619–644 (2002)
19. Marzouk, Y.M., Najm, H.N., Rahn, L.A.: Stochastic spectral methods for efficient Bayesian solution of inverse problems. J. Comput. Phys. **224**, 560–586 (2007)
20. Saad, G.A.: Stochastic Data Assimilation with Application to Multi-Phase Flow and Health Monitoring Problems. PhD thesis, University of Southern California, Los Angeles (2007)
21. Zhang, D.X., Lu, Z.M., Chen, Y.: Dynamic reservoir data assimilation with an efficient, dimension-reduced Kalman filter. SPE J. **12**, 108–117 (2007)
22. Tatang, M.A., Pan, W.W., Prinn, R.G., McRae, G.J.: An efficient method for parametric uncertainty analysis of numerical geophysical models. J. Geophys. Res. **102**, 21925–21932 (1997)
23. Li, H., Zhang, D.X.: Probabilistic collocation method for flow in porous media: comparisons with other stochastic methods. Water Resour. Res. **43**, W9409 (2007)
24. Shi, L.S., Yang, J.H., Zhang, D.X., Li, H.: Probabilistic collocation method for unconfined flow in heterogeneous media. J. Hydrol. **365**, 4–10 (2009)
25. Xiu, D.B., Hesthaven, J.S.: High-order collocation methods for differential equations with random inputs. Siam. J. Sci. Comput. **27**, 1118–1139 (2005)
26. Babuska, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. Siam J. Numer. Anal. **45**, 1005–1034 (2007)
27. Chang, H.B., Zhang, D.X.: A comparative study of stochastic collocation methods for flow in spatially correlated random fields. Commun. Comput. Phys. **6**, 509–535 (2009)
28. Zabaras, N., Ganapathysubramanian, B.: A scalable framework for the solution of stochastic inverse problems using a sparse grid collocation approach. J. Comput. Phys. **227**, 4697–4735 (2008)
29. Sarma, P., Durlofsky, L.J., Aziz, K.: Kernel principal component analysis for efficient, differentiable parameterization of multipoint geostatistics. Math. Geosci. **40**, 3–32 (2008)
30. Xiu, D.B.: Numerical integration formulas of degree two. Appl. Numer. Math. **58**, 1515–1520 (2008)
31. Chen, Y., Oliver, D.S.: Improved initial sampling for the ensemble Kalman filter. Computat. Geosci. **13**, 13–26 (2009)
32. Li, H., Zhang, D.X.: Efficient and accurate quantification of uncertainty for multiphase flow with probabilistic collocation method. SPE J. **14**, 665–679 (2009)
33. Sarma, P., Durlofsky, L. J., Aziz, K.: Efficient closed-loop production optimization under uncertainty. SPE paper 94241 (2005)
34. Sakamoto, S., Ghanem, R.: Polynomial chaos decomposition for simulation of non-Gaussian non-stationary stochastic processes. ASCE J. Eng. Mech. **128**, 190–201 (2002)
35. Wan, X., Karniadakis, G.E.: Solving elliptic problems with non-Gaussian spatially-dependent random coefficients. Comput. Meth. Appl. Mech. Eng. **198**, 1985–1995 (2009)
36. Sarma, P., Chen, W.H.: Generalization of the ensemble Kalman filter using kernels for non-gaussian random fields. SPE paper 119177 (2009)
37. Li, W.X., Lu, Z.M., Zhang, D.X.: Stochastic analysis of unsaturated flow with probabilistic collocation method. Water Resour. Res. **45**, W08425 (2009)