



# Ensemble Smoother with Enhanced Initial Samples for Inverse Modeling of Subsurface Flow Problems

Qinzhuo Liao<sup>1</sup>

Received: 22 December 2020 / Accepted: 18 November 2021 / Published online: 21 March 2023  
© King Fahd University of Petroleum & Minerals 2023

## Abstract

Ensemble-based data assimilation methods have been extensively investigated for inverse problems of fluid flow in porous media. However, when the permeability field is characterized by fine-scale gridblocks, the problem can be ill-posed and result in non-unique solutions. To address this issue, the principal component analysis with truncation was presented, but it may lead to biased estimation. In this paper, we propose to keep all eigenfunctions without truncation and add an additional sorting step after principal component analysis: sorting the initial samples according to the dimensional variability and assigning the dimensions with large variances to the leading eigenfunctions. The estimation is expected to be more accurate as the subspace spanned by the ensemble favors the dominant components. The proposed method is tested for multiple synthetic flow and transport cases. The results show that it provides more accurate estimation of the permeability fields and generates better history matching and prediction results for the production data (by 10–15%) than the results from the standard ensemble smoother, with the same computational cost. This sorting approach can be readily extended to the ensemble Kalman filter as well, for inverse modeling and estimating reservoir properties.

**Keywords** Inverse problems · Data assimilation · Ensemble smoother · Enhanced initial samples

## Abbreviations

$C$	Covariance	$\lambda$	Eigenvalue
$f$	Eigenfunction	$\xi$	Independent random variable
$h$	Hydraulic head, m	$\eta$	Correlation length, ft
$k$	Absolute permeability, mD	$\sigma^2$	Variance
$k_{r\alpha}$	Relative permeability for $\alpha$ phase fluid	$\mu_\alpha$	Viscosity of $\alpha$ phase fluid, Pa·s
$N$	Normal distribution	$\rho_\alpha$	Density of $\alpha$ phase fluid, kg/m <sup>3</sup>
$p_c$	Capillary pressure, psia	$\phi$	Porosity
$p_\alpha$	Pressure of $\alpha$ phase fluid, psia	RMSE	Root mean square error
$q_\alpha$	Source/sink term, kg/s	BHP	Bottom hole pressure, psia or bar
$S_\alpha$	Saturation of the $\alpha$ phase fluid	OPR	Oil production rate, bbl/day or m <sup>3</sup> /day
$S_{or}$	Residual oil saturation	WPR	Water production rate, bbl/day or m <sup>3</sup> /day
$S_{wc}$	Irreducible water saturation	GOR	Gas–oil ratio
$t$	Time, day	WCT	Water cut
$u_\alpha$	Velocity of $\alpha$ phase fluid, m/s	FOPT	Field oil production total, bbl/day or m <sup>3</sup> /day
$x$	Location in space, m	FGPT	Field gas production total, bbl/day or m <sup>3</sup> /day
$Y$	Log-permeability, mD	FWPT	Field water production total, bbl/day or m <sup>3</sup> /day

✉ Qinzhuo Liao  
liaoqz@gmail.com

<sup>1</sup> College of Petroleum Engineering, China University of Petroleum (Beijing), Beijing 102249, China

## 1 Introduction

A solid understanding of geological formation properties is fundamental for analyzing geophysical systems [1–3]. These

properties are usually heterogeneous and contain substantial uncertainties due to limited knowledge and large scales in space [4–6]. Data assimilation of indirect measurements plays an important role in characterization of formation properties and uncertainty reduction [7]. The most common properties are spatially correlated parameters, such as conductivity/permeability and porosity. These parameter values are usually updated to match the available data, such as hydraulic head, pressure and fluid flow rate [8–10].

Ensemble-based data assimilation methods, such as the ensemble Kalman filter (EnKF), have attracted a great deal of attention and have been successfully applied in geophysical problems recently. The EnKF, which is basically a Monte Carlo-type Kalman filter, has been widely used for data assimilation of nonlinear problems in oceanic [11], atmospheric [12], hydrological [13–17] modeling and petroleum industry [18–23]. However, as a sequential data assimilation technique that updates model parameters and states simultaneously at each step, the EnKF may suffer from the inconsistency between updated parameters and states in strongly nonlinear problems [24]. Although this issue can be alleviated by an iterative or restarting approach [19, 25], additional computational cost is required. In this case, computing a global update by simultaneously assimilating all observations is preferred, which leads to the ensemble smoother (ES) [26–28]. It has been shown that the ES obtains results similar to those from the EnKF, but in a more efficient and convenient way, for some reservoir history matching problems [29].

In the above ensemble-based methods, the quality of initial samples is important for a successful application of data assimilation. Considering that the ensemble size (i.e., the number of samples in the ensemble) is limited to computational cost or storage space, it is usually in order of hundreds and much smaller than the number of parameters, which is usually in order of thousands to millions. That is, the initial samples essentially construct a subspace whose dimension or degree of freedom is one less than the ensemble size, if the samples are linearly independent. When the number of observations is larger than the ensemble size, especially when the measurements are in time series [30], the system becomes an overdetermined problem, i.e., a larger number of equations than the number of unknowns. In this situation, the subspace spanned by the initial samples becomes more crucial to the performance of data assimilation than that in the conventional cases.

Several approaches have been proposed and implemented to alleviate this issue. Firstly, the ensemble size may be increased by sampling additional model states, hence increasing the dimension of the subspace spanned by the ensemble, albeit at a cost of extra computational burden. The second choice is to find an ensemble that possibly best represents the covariance matrix for a given ensemble size, which can be achieved by principle component analysis and is usually implemented by proper orthogonal decomposition, e.g., eigenvalue decomposition or singular value decomposition. Specifically, only the first few dominant eigenfunctions are retained (by truncation) and used to generate the initial samples. Such an idea has been applied to the ensemble-based data assimilation for geophysical models [31–34]. However, as pointed out in [35, 36], these approaches may lead to too smooth realizations and biased estimation.

In this paper, we propose a simple yet effective approach, the sorted ensemble smoother, to address the above issue by enhancing the initial samples for improved accuracy in parameter estimation. We sort the samples according to the variance in each dimension and assign the dimensions with larger variances to the dominant components. In this way, the dominant components are more likely to be updated/corrected while all eigenfunctions are retained without truncation. We remark that an iterative scheme of ES to assimilate the measurements multiple times can be adopted to achieve satisfactory data match in nonlinear problems. This paper is organized as follows: The methodology is introduced in Sect. 2. The numerical examples are given in Sect. 3. Some discussions are presented in Sect. 4. Finally, the conclusions are provided in Sect. 5.

## 2 Methodology

### 2.1 Governing Equations

The steady-state, single-phase groundwater flow in the aquifer satisfies the following equation [37]:

$$\nabla \cdot [-k(\mathbf{x})\nabla h(\mathbf{x})] = q(\mathbf{x}) \quad (1)$$

where  $\mathbf{x}$  is the location,  $k(\mathbf{x})$  is the hydraulic conductivity,  $h(\mathbf{x})$  is the hydraulic head, and  $q(\mathbf{x})$  is the source/sink term. In this study, the hydraulic head and the source/sink term are assimilated to estimate the conductivity, which is treated as a random field in space.

The oil/water two-phase immiscible flow model can be expressed by the following continuity equation as [38]:

$$\frac{\partial(\phi\rho_\alpha S_\alpha)}{\partial t} = -\nabla \cdot (\rho_\alpha \mathbf{u}_\alpha) + q_\alpha, \quad \alpha = w, o \tag{2}$$

where  $\phi$  is the porosity, and each phase has its own density  $\rho_\alpha$ , saturation  $S_\alpha$ , phase velocities  $\mathbf{u}_\alpha$ , and source/sink term  $q_\alpha$ . Darcy’s law for multiphase flow is:

$$\mathbf{u}_\alpha = -\frac{\mathbf{k}k_{r\alpha}}{\mu_\alpha}(\nabla p_\alpha - \rho_\alpha g \nabla z), \quad \alpha = w, o \tag{3}$$

where  $\mathbf{k}$  is the absolute permeability;  $g$  is the gravitational acceleration;  $z$  is the depth; and  $k_{r\alpha}$ ,  $\mu_\alpha$ , and  $p_\alpha$  are the relative permeability, viscosity, and pressure for phase  $\alpha$ , respectively. Equations (1) and (2) are usually coupled with:

$$S_w + S_o = 1, \quad p_c(S_w) = p_o - p_w \tag{4}$$

where  $p_c$  is the capillary pressure; which is a function of  $S_w$ . In this study, the absolute permeability is considered as a random field and updated using the observations of pressure and flow rate at wells.

In geostatistics, the log-conductivity or log-permeability  $Y = \ln k$  is usually treated as a second-order stationary Gaussian random field with a mean of  $\bar{Y}$  and a covariance of [38]

$$C_{\ln k}(\mathbf{x}, \mathbf{x}') = \sigma_{\ln k}^2 \exp \left[ -\sqrt{\left(\frac{x-x'}{\eta_x}\right)^2 + \left(\frac{y-y'}{\eta_y}\right)^2 + \left(\frac{z-z'}{\eta_z}\right)^2} \right] \tag{5}$$

where  $\sigma_{\ln k}^2$  is the variance and  $\eta$  is the correlation length.

### 2.2 Ensemble Smoother

If we represent the solution of the flow and transport equation as the following forward simulation:

$$\mathbf{d} = g(\mathbf{m}) + \boldsymbol{\varepsilon} \tag{6}$$

where  $\mathbf{d}$  is a  $N_d \times 1$  vector for the measurements,  $g(\cdot)$  is the system model,  $\mathbf{m}$  is a  $N_m \times 1$  vector for the uncertain parameters,  $\boldsymbol{\varepsilon}$  is a  $N_d \times 1$  vector for the measurement errors. Here we use  $N_d$  to denote the total number of measurements and  $N_m$  the number of parameters. After the forecast stage by forward simulations, the unknown model parameters are updated using the measurements/observations as:

$$\mathbf{m}_j^u = \mathbf{m}_j^f + \mathbf{C}_{\mathbf{MD}}^f (\mathbf{C}_{\mathbf{DD}}^f + \mathbf{C}_{\mathbf{D}})^{-1} \times (\mathbf{d}_j - g(\mathbf{m}_j^f)), \quad j = 1, \dots, N_e \tag{7}$$

where  $\mathbf{m}_j^u$  is the  $j$ -th updated ensemble member,  $\mathbf{m}_j^f$  is the  $j$ -th prior ensemble member,  $u$  indicates “update,”  $f$  represents “forecast,”  $j$  is the ensemble member index, and  $N_e$  is the

ensemble size.  $\mathbf{C}_{\mathbf{MD}}^f$  is the  $N_m \times N_d$  cross-covariance matrix between the prior parameters  $\mathbf{M}^f = [\mathbf{m}_1^f, \dots, \mathbf{m}_{N_e}^f]$ , and the predicted data  $\mathbf{D}^f = [g(\mathbf{m}_1^f), \dots, g(\mathbf{m}_{N_e}^f)]$ ;  $\mathbf{C}_{\mathbf{DD}}^f$  is the  $N_d \times N_d$  covariance matrix of predicted data  $\mathbf{D}^f$ ;  $\mathbf{C}_{\mathbf{D}}$  is the  $N_d \times N_d$  covariance matrix of measurement errors;  $\mathbf{d}_j = \mathbf{d} + \boldsymbol{\varepsilon}_j$  is the  $j$ -th realization of the perturbed measurements; and  $\boldsymbol{\varepsilon}_j$  is a random realization of the measurement errors. The ES procedure can be repeated for nonlinear models, which is usually referred to as iterative ES.

### 2.3 Principle Component Analysis

The initial samples in the ES should be generated such that best represents the statistics of the log-conductivity including mean and covariance. Conceptually, the realizations in the initial ensemble should be linear combinations of the principle components of the covariance matrix. Such an algorithm is known as Karhunen–Loève expansion [39]:

$$Y(\mathbf{x}; \boldsymbol{\xi}) = \bar{Y}(\mathbf{x}) + \sum_{n=1}^{\infty} \sqrt{\lambda_n} f_n(\mathbf{x}) \xi_n \tag{8}$$

where  $\bar{Y}(\mathbf{x})$  is the mean log-conductivity,  $\{\xi_n\}$  are independent random variables normally distributed as  $\xi_n \sim N(0, 1)$ , and  $\{\lambda_n, f_n(\mathbf{x})\}$  are the eigenvalues and eigenfunctions from solving the Fredholm integral equation of the second kind:

$$\int C_{\ln k}(\mathbf{x}, \mathbf{x}') f_n(\mathbf{x}) d\mathbf{x} = \lambda_n f_n(\mathbf{x}') \tag{9}$$

and the covariance matrix can be decomposed as:

$$C_{\ln k}(\mathbf{x}, \mathbf{x}') = \sum_{n=1}^{\infty} \lambda_n f_n(\mathbf{x}) f_n(\mathbf{x}') \tag{10}$$

Although the theorem provides a series of infinite eigenfunctions, the number is finite in real applications (e.g., no more than the number of gridblocks for a random field in space). Since  $\bar{Y}(\mathbf{x})$  and  $\{\lambda_n, f_n(\mathbf{x})\}$  are fixed for given statistical properties, there is a one-to-one mapping relation between  $Y(\mathbf{x}; \boldsymbol{\xi})$  and  $\{\xi_n\}$ . Therefore, we just need to update the random variables  $\{\xi_n\}$  for updating  $Y(\mathbf{x}; \boldsymbol{\xi})$ ; thus, the parameters to be updated are  $\mathbf{m} = [\xi_1, \dots, \xi_{N_m}]^T$ .

### 2.4 Sorted Ensemble Smoother

Assume the ensemble size is  $N_e$ , the dimension of the subspace spanned by the ensemble is hence  $N_e - 1$ . We may keep the first  $N_e - 1$  dominant components (i.e., leading eigenfunctions) of the covariance. By doing so, the subspace spanned by the ensemble minimizes the total mean square error in decomposition and optimally compacts the energy.

Similar ideas such as the discrete cosine transform [40] also play a similar role for re-parameterization. However, these approaches lead to too smooth realizations and biased estimation [35, 36].

One of the reasons why the random sampling in the standard ES may generate poor initial samples is that the sample variances in the dominant components could be very small. For example, consider a simple mathematical model  $y = 10x_1 + x_2$ ,  $x_i \sim N(0, 1)$ ,  $i = 1, 2$ . Apparently,  $x_1$  is the dominant parameter compared to  $x_2$ . Assume we use only two samples, which consists a 1D subspace, i.e., a straight line. If the line is parallel to the  $x_1$ -axis, it means the update or adjustment is for  $x_1$  only, which is equivalent to keeping only the first (dominant) component by truncation. However, if the line is orthogonal to the  $x_1$ -axis, it means there is no update or adjustment for  $x_1$ , which is equivalent to keeping only the second (less-important) component. Clearly, the second case is what we are attempting to avoid. The differences in these two cases can be detected from the sample variances in each dimension of the components: in the first case, the variance for  $x_1$  is larger than that for  $x_2$ , and vice versa in the second case.

If we would like to enhance the initial samples without too smooth realizations, we may keep all components without

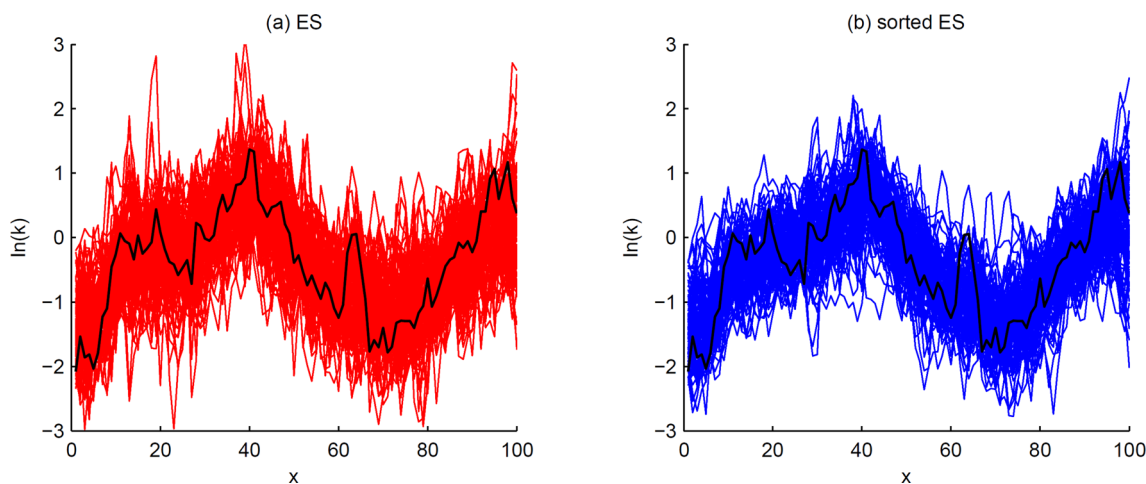
truncation, while making the sample variances in the dominant components larger than the sample variances in the less important components. Based on this idea, we propose a simple and efficient way by sorting the variances of the samples in each dimension of the components and then adjusting the samples by assigning the dimensions with larger variances to the dominant components. For example, in the above mathematical model, assume we randomly generated two sample points ( $x_1 = 0$ ,  $x_2 = -1$ ) and ( $x_1 = 1$ ,  $x_2 = 2$ ). That is,  $x_1 = 0, 1$  for the first dimension and  $x_2 = -1, 2$  for the second dimension. Since the variance for  $x_2$  is larger than the variance for  $x_1$ , considering that  $x_1$  is the dominant component, we will assign the random variables  $-1$  and  $2$  to the first dimension for  $x_1$  and  $0$  and  $1$  to the second dimension for  $x_2$ . Hence, the new sample points become ( $x_1 = -1$ ,  $x_2 = 0$ ) and ( $x_1 = 2$ ,  $x_2 = 1$ ). By this means, the dominant components are more likely to be updated than the less-important components. Then, the standard ES approach can be performed after the above sorting process. It should be noted that for strongly nonlinear problems, we may use the iterative ES, and the sorting process is only implemented once for the initial samples. Complete scheme of the iterative sorted ES algorithm is presented in Algorithm 1. The sorted ES without iteration can be performed easily by setting the iteration counter to one.

---

#### Algorithm 1. Iterative sorted ensemble smoother

---

- 1: Perform principle component analysis and solve equation (9) for  $\{\lambda_n, f_n(\mathbf{x})\}, n = 1, \dots, N_m$ .
  - 2: Randomly generate initial samples  $\mathbf{M}^f = [\mathbf{m}_1^f, \dots, \mathbf{m}_{N_e}^f]$  from the prior distribution, where  $\mathbf{m}$  is a  $N_m \times 1$  vector.
  - 3: Compute the sample variances in each dimension from 1 to  $N_m$  and sort the sample variances in a decreasing order.
  - 4: Assign the dimensions with larger variances to the dominant components, resulting in sorted initial samples  $\hat{\mathbf{M}}^f$ , where the first row has largest variance, the second row has the second largest variance, etc.
  - 5: Compute initial log-conductivity samples using equation (8) and  $\hat{\mathbf{M}}^f$ .
  - 6: **For** iteration counter  $t = 1, \dots, N_t$ 
    - 7: Run forward simulations using equation (6) as forecast stage.
    - 8: Adjust the parameters using equation (7) as update stage.
  - 9: **End for**
-



**Fig. 1** One hundred ensembles of log-conductivity in 1D linear example: **a** the red lines represent the ES results; **b** the blue lines represent the sorted ES results. The black line shows the true reference. Readers are referred to the online version for the color lines

### 3 Numerical Examples

In this section, the proposed method was tested in flow and transport problems. We start with relatively simple models because we would like to compare the performance of the ES and the sorted ES statistically, which requires running the forward model many times. First, we consider a 1D linear model, in which the conductivity is observed in some locations directly. The second example is a 1D single-phase flow model, in which the hydraulic head is assimilated to infer the conductivity field. The third example is a 2D multiphase flow model, in which the pressure and flow rate are observed in injection and production wells.

We remark that we use multiple ensembles (each contains a certain number of ensemble members) in this study, in order to analyze the performances statistically. This does not mean that we need to use multiple ensembles in real applications, where one ensemble is used in the traditional way.

The following terms are defined to quantify accuracy. Denote  $\tilde{y}(n)$  the true model parameter (i.e.,  $\ln(k)$ ) at grid-block  $n, n = 1, \dots, N_b$ , where  $N_b$  is the total number of blocks.  $y_{ij}(n)$  is the estimated parameter in the  $i$ -th ensemble and the  $j$ -th ensemble member. Hence, the mean (over samples in each ensemble)  $\ln(k)$  for ensemble  $i$  is:

$$\bar{y}_i(n) = \frac{1}{N_e} \sum_{j=1}^{N_e} y_{ij}(n) \tag{11}$$

The error of mean  $\ln(k)$  for ensemble  $i$  is:

$$\text{mean error} = \sqrt{\frac{1}{N_b} \sum_{n=1}^{N_b} [\bar{y}_i(n) - \tilde{y}(n)]^2} \tag{12}$$

The spread of mean  $\ln(k)$  for ensemble  $i$  is:

$$\text{spread} = \sqrt{\frac{1}{N_b} \frac{1}{N_e} \sum_{n=1}^{N_b} \sum_{j=1}^{N_e} [y_{ij}(n) - \bar{y}_i(n)]^2} \tag{13}$$

The root mean square error (RMSE) of  $\ln(k)$  for ensemble  $i$  is:

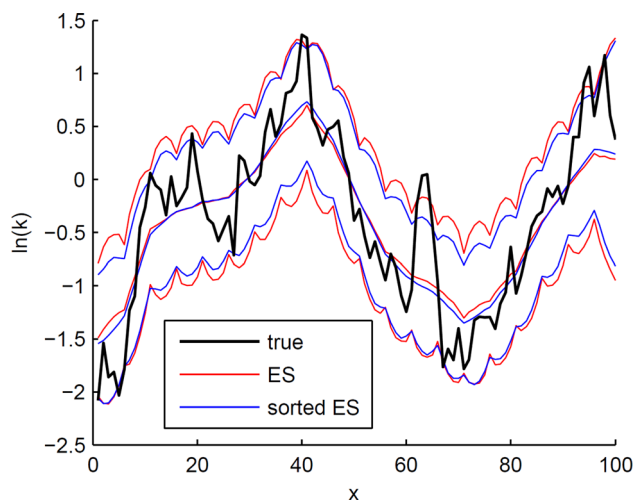
$$\text{RMSE} = \sqrt{\frac{1}{N_b} \frac{1}{N_e} \sum_{n=1}^{N_b} \sum_{j=1}^{N_e} [y_{ij}(n) - \tilde{y}(n)]^2} \tag{14}$$

And the relation of the above three is:  $\text{mean error}^2 + \text{spread}^2 = \text{RMSE}^2$ .

#### 3.1 1D Linear Model

Consider a linear problem in which the log-conductivity  $\ln(k)$  in a 1D space of  $N_b = 100$  gridblocks is treated as a Gaussian random field with a mean of 0 and a covariance of  $C_{\ln k} = \sigma_{\ln k}^2 \exp(-|x_1 - x_2|/\eta)$ , where  $\sigma_{\ln k}^2 = 1.0$  and the correlation length is  $\eta = 20$  gridblocks. The true reference is randomly generated from the above statistics as shown in Fig. 1 (in black). We set the ensemble size to be  $N_e = 10$  and measure the log-conductivity at  $N_d = 20$  equispaced locations  $x = 1, 6, 11, \dots, 96$  with standard error of 0.01 and infer the log-conductivity in the whole space. Clearly, it is a linear problem since we observed the parameters directly. Therefore, the ES or the sorted ES is performed only once without iteration.

Figure 1a shows 100 ensembles after data assimilation in the ES. Note that the degree of freedom is  $N_e - 1 = 9$ , which is less than the number of observation  $N_d = 20$ . Thus, it is

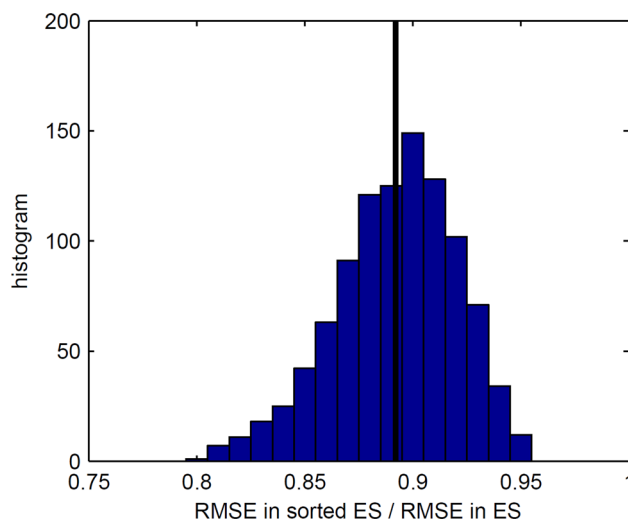


**Fig. 2** Log-conductivity in 1D linear example. The black line shows the true reference. The red lines show the P10, P50 and P90 (over 10,000 ensembles) of the mean estimation in ES. The blue lines show the P10, P50 and P90 of the mean estimation in sorted ES

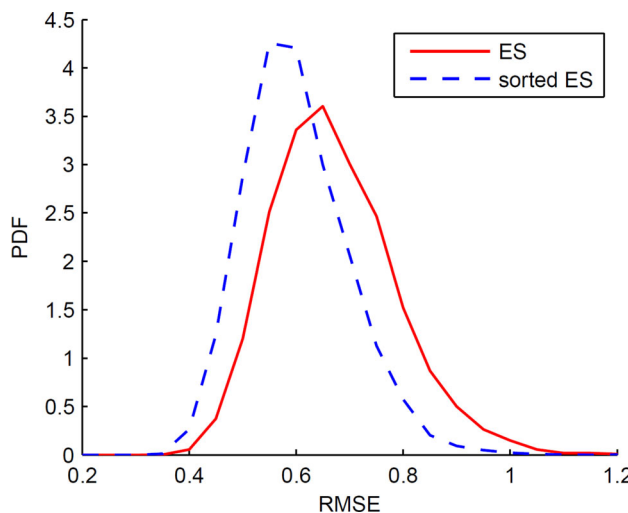
an overdetermined problem and the solution is in a least-square sense. Since the observation error is relatively small as 0.01, the 10 realizations in each ensemble are very close to each other (overlapped in the figure). Figure 1b shows 100 ensembles after data assimilation in the sorted ES. The deviation becomes smaller compared to the results from the ES, e.g., the overshoots near  $x = 18$  and  $x = 38$  in the ES disappear in the sorted ES.

We can see that different initial samples result in different results. To compare the performance of the two methods statistically, we generate  $N_s = 10,000$  ensembles, each of which contains 10 samples. Figure 2a presents the percentiles (over 10,000 ensembles) of the mean (of 10 samples in each ensemble) log-conductivity using the ES (in red) and the sorted ES (in blue). It can be seen that the P50 from the sorted ES has a smaller deviation than that from the ES. For example, the value from the sorted ES is closer to the true reference than the value from the ES near  $x = 40$ . It can also be observed that the confidence interval (represented by P10 and P90) in the sorted ES is narrower than that in the ES, indicating the former is likely to provide a more accurate result than the latter.

Figure 3 shows the distribution of RMSE for the 10,000 ensembles. It is clear that the RMSE of the sorted ES has a lower expectation and a smaller variance than the RMSE of the ES. Table 1 compares the mean error, spread and RMSE from the two approaches. The error ranges denoted by ‘ $\pm$ ’ are the standard deviations from 10,000 ensembles. In this case, the spread is almost negligible, because the number of unknown is smaller than the number of equations. Actually, although there are 100 parameters (one at each block), there are only 9 degree of freedom in the subspace formed by 10



**Fig. 4** Histogram of the ratio of the RMSE in sorted ES to the RMSE in ES, using 1,000 true references, each uses 10,000 ensembles. The black vertical line shows the mean of the ratio

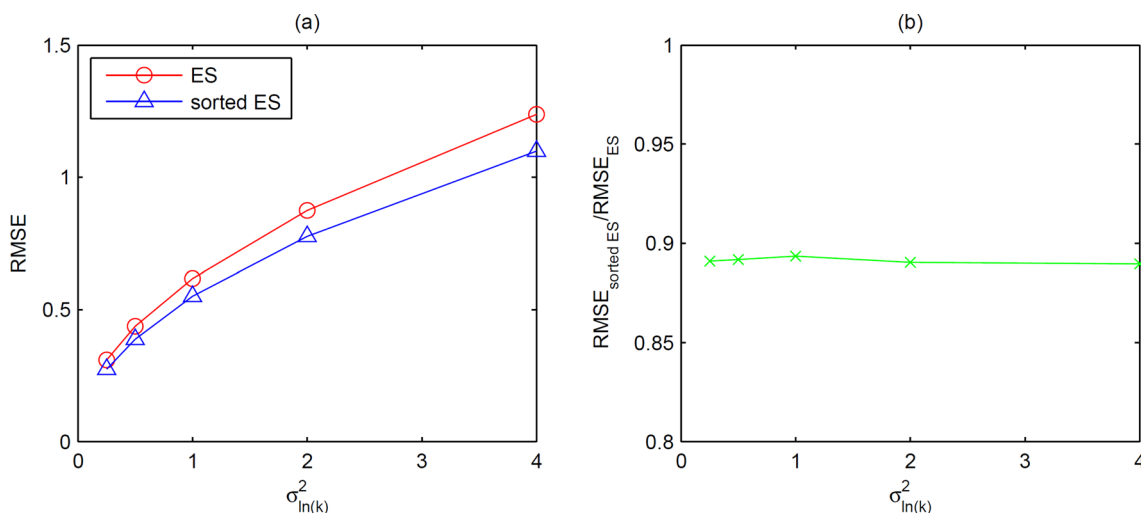


**Fig. 3** PDF of the RMSE in 1D linear example using 10,000 ensembles. The red solid line shows the ES results. The blue dashed line shows the sorted ES results

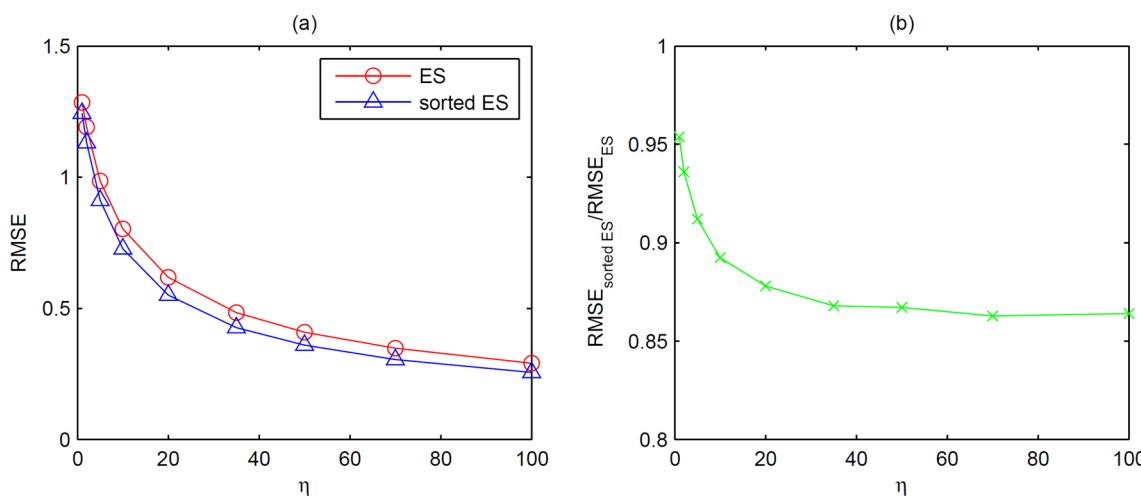
**Table 1** 1D linear test

	Mean error	Spread	RMSE
ES	$0.671 \pm 0.116$	$1.3e-4 \pm 3.7e-5$	$0.671 \pm 0.116$
Sorted ES	$0.600 \pm 0.096$	$9.1e-5 \pm 2.3e-5$	$0.600 \pm 0.096$

samples. Hence, the ES approaches tend to find the best linear solution in the least-square sense. To quantify the improvement in ES using the sorted process, we compute the ratio of the RMSE expectation (over 10,000 ensembles) in sorted ES to the RMSE expectation in ES. The value is  $0.600/0.671 =$



**Fig. 5** Effect of the variance of log-conductivity field: **a** RMSE expectations in ES and sorted ES; **b** the ratio of the RMSE expectation in sorted ES to the RMSE expectation in ES



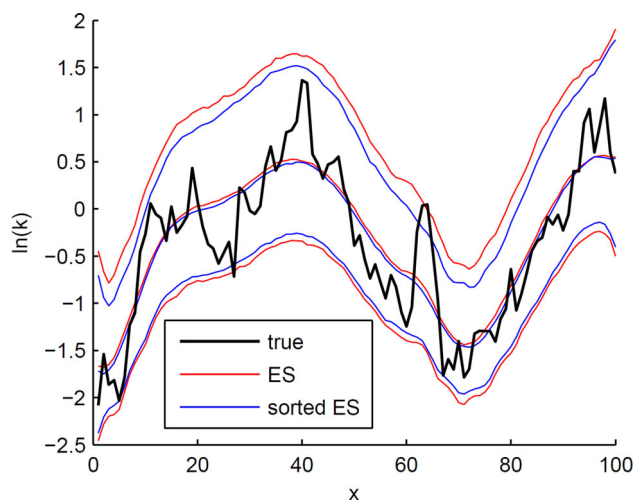
**Fig. 6** Effect of the correlation length of log-conductivity field: **a** RMSE in ES and sorted ES; **b** the ratio of the RMSE in sorted ES to the RMSE in ES

0.894, which is less than 1, indicating that the accuracy is improved by 10.6%.

Considering that the results may vary if the true reference is changed, we randomly generated 1,000 true references; each uses 10,000 ensembles. Figure 4 shows the histogram of the RMSE ratio in these 1,000 tests. It can be observed that in all tests, the sorted ES is expected to be more accurate than the ES. The black vertical line depicts the mean of the ratio, which equals to 0.892.

Finally, we performed another two tests on the sensitivity of the RMSE ratio to the statistics of the log-conductivity field, including the variance  $\sigma_{\ln k}^2$  and correlation length  $\eta$  as in Eq. (5). Figure 5 shows the effect of the variance of log-conductivity field. We can see from Fig. 5a that both RMSEs (in ES and sorted ES) increase as the variance increases, which is reasonable because a larger uncertainty in the prior

makes the inverse problem more challenging. Figure 5b reveals the ratio of the RMSE in sorted ES to the RMSE in ES. This value is not sensitive to the variance since the total variance does not change the ratio of eigenvalues in the decomposition. Figure 6 shows the effect of the correlation length of log-conductivity field. We can see Fig. 6a that both RMSEs (in ES and sorted ES) are reduced as the correlation length grows. This is because a larger correlation length leads to larger eigenvalues in the dominant components, and hence, it is easier for the initial samples, which forms reduced-order subspace, to represent the statistics of the prior. Figure 6b shows that the ratio of the RMSE in sorted ES to the RMSE in ES is also reduced as the correlation length increases. In the extreme case when the correlation length approaches 0 (i.e., uncorrelated), all the eigenvalues become the same; therefore, the sorting process does not have any effect. On the



**Fig. 7** Log-conductivity in 1D single-phase flow example. The black line shows the true reference. The red lines show the P10, P50 and P90 (over 10,000 ensembles) of the mean estimation in ES. The blue lines show the P10, P50 and P90 of the mean estimation in sorted ES

other hand, when the correlation length is significantly large, only the first eigenvalue becomes dominant, and thus, the improvement in the sorted ES compared to the ES will be stable.

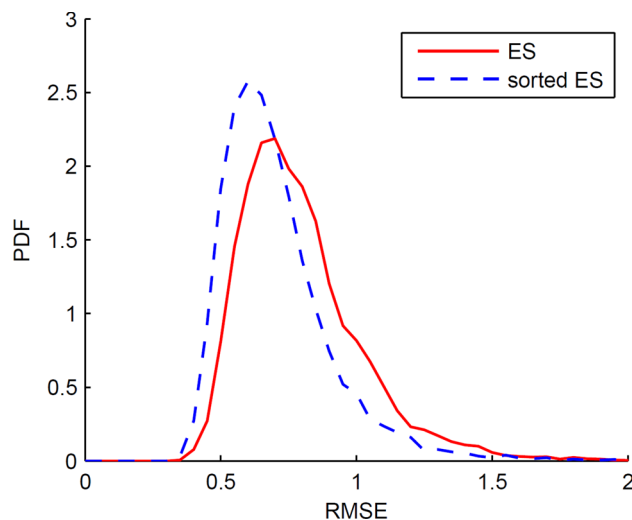
### 3.2 1D Single-phase Flow Model

Let us consider a 1D space with the same statistical properties as above. We use the same true reference as shown in Fig. 1 and also observe the hydraulic heads at the same locations  $x = 1, 6, 11, \dots, 96$  with standard error of 0.01 and estimate the log-conductivity in the whole space. Now we assume the groundwater flow has following boundary conditions (a constant flux at the left end and a constant head at the right end):

$$k \frac{dh}{dx} \Big|_{x=0} = -q_0, \quad h|_{x=L} = 0 \tag{15}$$

where  $q_0 = 1.0$ . This is a nonlinear problem, and thus, 10 iterations in both the ES and the sorted ES were performed. We still use 10,000 ensembles with 10 members in each ensemble.

Figure 7 illustrates the updated log-conductivity after the last iteration. It is observed that the sorted ES matches the true reference better than the ES. Figure 8 shows the distribution of the RMSE. Again, we see that the error of the sorted ES is likely to be less than the ES error. These findings are validated in Table 2, where the mean error, spread and RMSE are compared in details. In this case, the ratio of RMSE expectation in the sorted ES to the RMSE expectation in the ES is  $0.707/0.796 = 0.888$ , which is close to that in the 1D linear tests. Note that the differences of RMSE between



**Fig. 8** PDF of the RMSE in 1D single-phase flow example using 10,000 ensembles. The red solid line shows the ES results. The blue dashed line shows the sorted ES results

**Table 2** 1D single-phase flow test

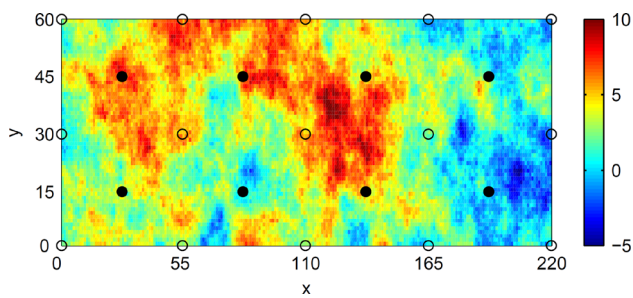
	Mean error	Spread	RMSE
ES	$0.795 \pm 0.228$	$0.037 \pm 0.004$	$0.796 \pm 0.228$
Sorted ES	$0.706 \pm 0.230$	$0.035 \pm 0.004$	$0.707 \pm 0.230$

the two methods are due to the differences in mean errors, while the spreads in the two methods are very close, indicating the sorting process honors the variability/uncertainty in the ensemble.

### 3.3 2D Multiphase Flow Model

As our last example, we test the proposed method using a multiphase flow model in a 2D heterogeneous formation: the third layer in the SPE10 benchmark problem. The domain contains  $60 \times 220$  cells, and each size is  $20 \times 10$  ft. The true log-permeability (in mD) has a large variability from  $-5$ – $10$  as shown in Fig. 9. We assume the porosity is 0.2, the relative permeabilities follow quadratic functions as  $k_{rw}(S_w) = (S^*)^2$  and  $k_{ro}(S_w) = (1 - S^*)^2$ , where  $S^* = (S - S_{wc}) / (1 - S_{wc} - S_{or})$ ,  $S_{wc} = 0.2$  and  $S_{or} = 0.2$ . The initial pressure is 6,000 psi, and the initial saturation is 0.2. The log-permeability is taken as a random field with a mean of 3.0, a variance of  $\sigma_{\ln k}^2 = 6.0$  and correlation lengths as  $\eta_x = \eta_y = 300$  ft. To infer the permeability, flow is induced from 15 injection wells to 8 production wells (Fig. 9). The injection wells are controlled by constant rates of 200 bbl/day and the production wells by constant pressures at 4,000 psi. The bottom hole pressure (BHP) at the injection wells, as well as the oil production rate (OPR) and water





**Fig. 9** True log-permeability  $\ln(k)$  and well configuration for the 2D example. The hollow circles indicate injection wells and the solid circles indicate production wells

production rate (WPR) at the production wells, is observed monthly up to 3 years. In the ensemble smoother approaches, five iterations are performed and 1,000 ensembles with 50 members in each ensemble are used.

Figure 10 shows three ensembles (randomly chosen out of 1,000 ensembles) of the mean log-permeability (each ensemble contains 50 samples) from the ES and the sorted ES after the last iteration. Both methods capture the high permeability zones at the top left and in the center, as well as the low permeability zone at the right. The results from the sorted ES are closer to each other and to the true reference, too.

Figure 11 shows the confidence interval using percentiles. We can see that the P50 of both methods is similar, whereas the P10/P90 of the sorted ES is greater/less than that of the ES, indicating the sorted ES is more stable. Figure 12 shows the detailed results along two lines (i.e.,  $y = 30$  and  $x = 110$ ). It is clear that the sorted ES provides a smaller confidence interval than the ES. The ratio of RMSE expectation in the sorted ES to the RMSE expectation in the ES is  $2.26/2.66 = 0.85$ , i.e., the improvement is about 15%. Figure 13 reveals the injection and production data. The updated results from the ES and sorted ES are closer to the true observations than the prior results. In addition, the sorted ES improves matches compared to the ES with a smaller confidence interval, which is consistent with the results in permeability.

### 3.4 3D PUNQ-S3 Benchmark Case

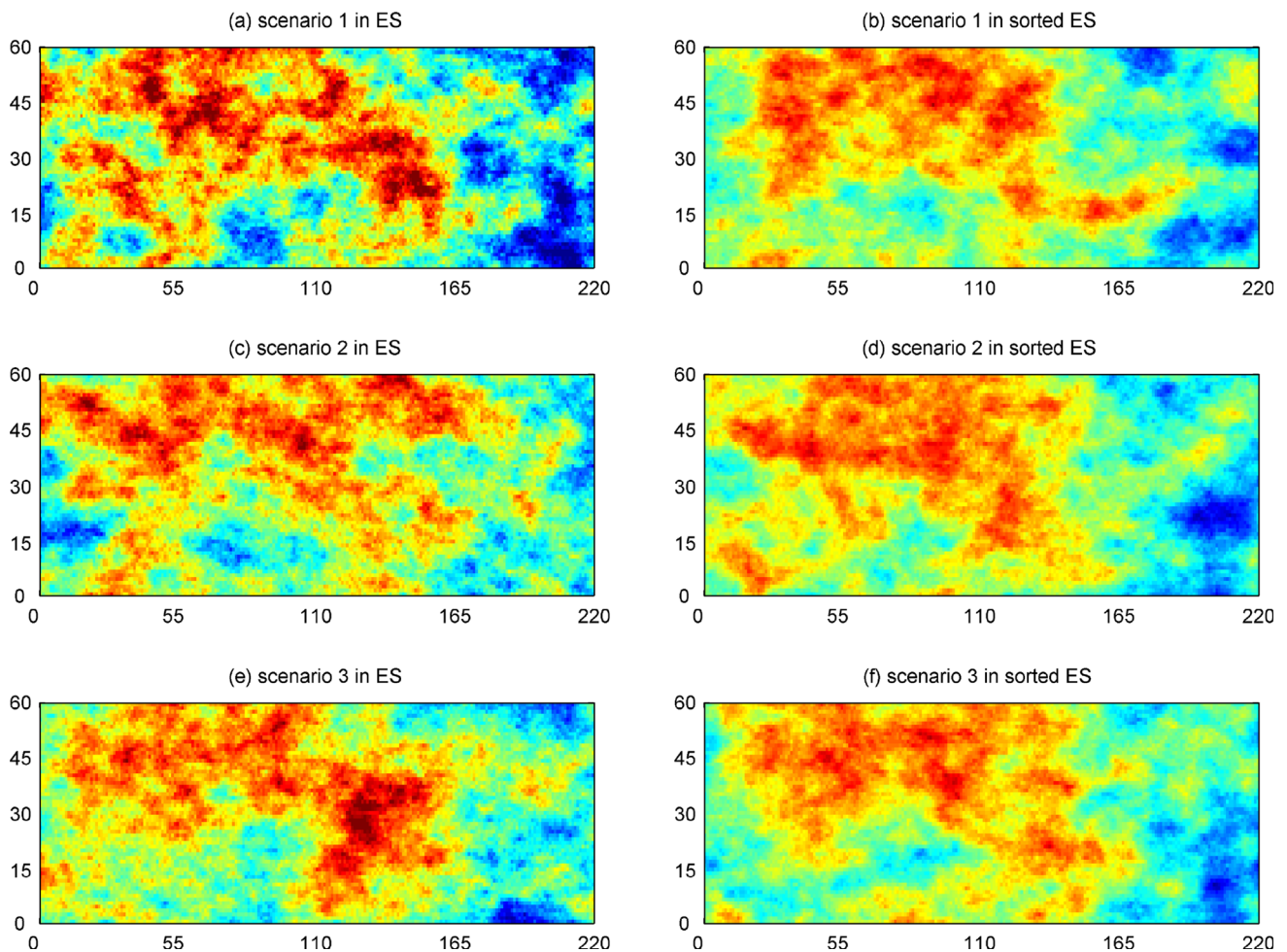
The PUNQ-S3 case study is a benchmark reservoir model based on a real field example set up by the Production forecasting with Uncertainty Quantification (PUNQ) project and widely used for history matching and prediction [41, 42]. The model contains  $19 \times 28 \times 5$  gridblocks, the dimensions of which in the x- and y-directions are  $180 \times 180 \text{ m}^2$ . The field contains 6 production wells, but no injection wells, as it links to a strong aquifer. The true porosity and permeability data are given and used to produce the true observation. A full description of this case study can be found on the PUNQ-S3

webpage [43], following which we follow randomly generate 40 realizations of the static state as in [42], including porosity, horizontal permeability and vertical permeability using a sequential Gaussian simulation/co-simulation module in SGeMS (Stanford Geostatistical Modeling Software) with Gaussian random fields for each layer. The project uses the production history of the first 8 years, including 1 year of well testing, 3 years of field shut-in and 4 years of actual production. Starting from year 5, the oil production rate is capped at  $150 \text{ m}^3/\text{day}$ , the minimum bottom hole pressure is 120 bar, and all wells are shut-in at the first 2 weeks in each year. The production data, including well bottom hole pressure (WBHP), well oil production rate (WOPR), well gas–oil ratio (WGOR) and well water cut (WWCT), are assimilated once a year. Then, the model is simulated up to 16 years, i.e., the last 8 years' forecasting results are used for comparison.

Figure 14 shows the history matching (from year 0–8) and prediction (from year 9–16) results for well PRO-11 generated from the prior (first row), and the final updated porosity and permeability. The thin curves indicate the ensemble forecasts, and the thick black curve indicates the reference. We can see that the ES (second row) roughly matches the production history, but fails to predict the forecast period for WBHP, WOPR and WWCT, due to poor estimation in porosity and permeability. However, the sorted ES matches the reference much better and reduces the uncertainty associated with the prior estimation. Similar conclusion is drawn from Fig. 15, where the results of field oil production total (FOPT), field gas production total (FGPT) and field water production total (FWPT) are illustrated. We remark that the FOPT and FGPT have relatively small variability, due to the fact the well is essentially controlled by the oil rate (caped at  $150 \text{ m}^3/\text{day}$ ), while the FWPT clearly shows the improvement using the sorted ES over the traditional ES.

## 4 Discussions

In this study, we aim to test the proposed method statistically with a large number of ensembles. Hence, we start with linear and single-phase flow models in relatively small spatial domains to make it affordable to run the forward model many times. Consider that the spatial domain could be very large with a great number of parameters (e.g., up to millions) in real field while the ensemble size is often much fewer (e.g., in hundreds), and the number of observations could be also very large especially in time series analysis. This is actually a severely overdetermined problem, i.e., greater number of observations than the degree of freedom in parameter space. Therefore, in the 1D examples, we set the ensemble size to be 10, which is smaller than the number of observations as 20. In the 2D case, we observe the BHP in 15 injection wells and OPR and WPR in 8 production wells monthly up



**Fig. 10** Three scenarios of the mean (of 50 samples) log-permeability  $\ln(k)$  after matching the history: the left column is from the ES; the right column is from the sorted ES

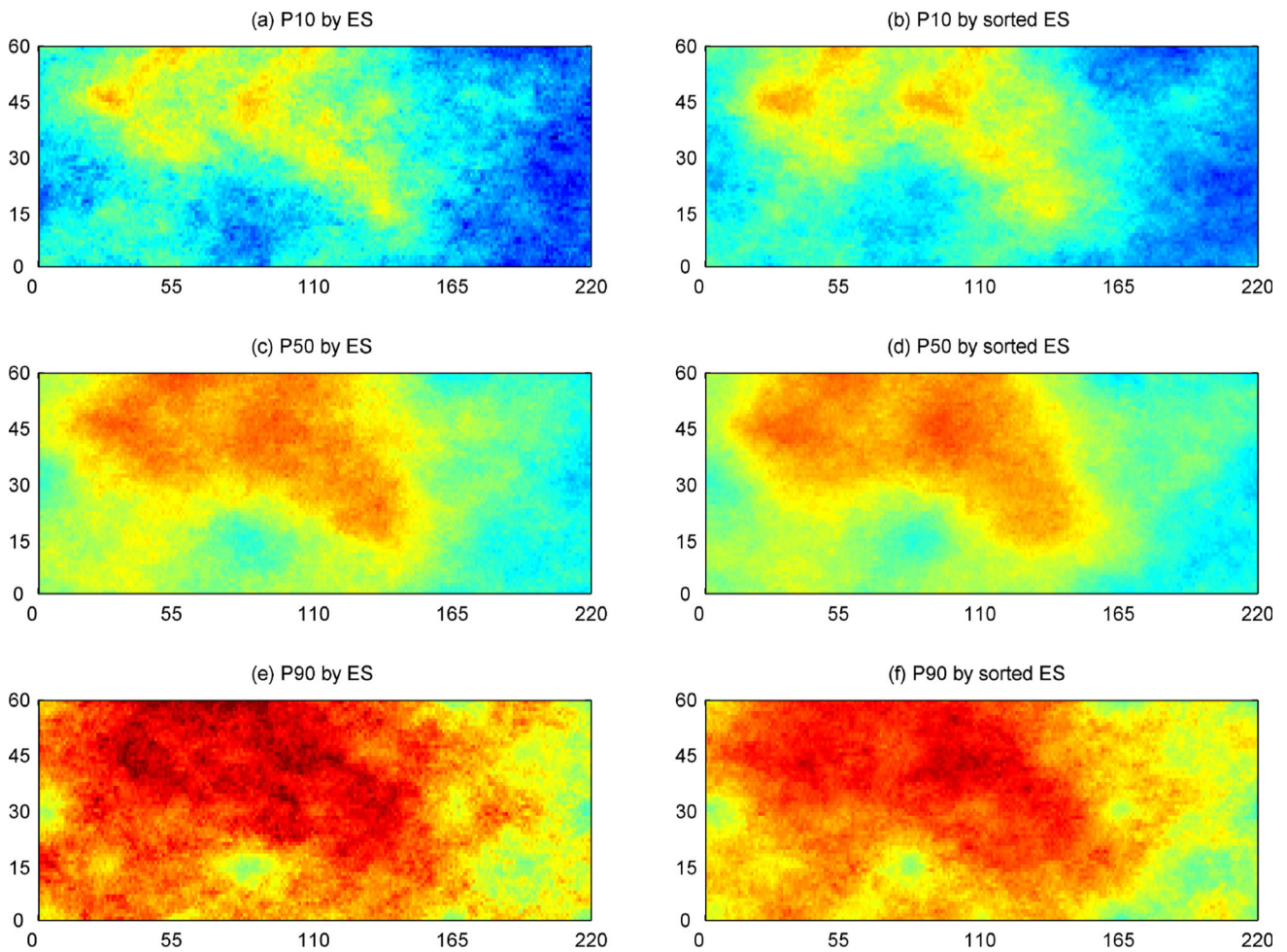
to 36 months. Therefore, the total number of observations is  $(15 + 2 \times 8) \times 36 = 1116$ , which is much larger than the ensemble size of 50. Although we can use more samples to improve the matching results, it is usually unaffordable to set an ensemble size close or greater than to the number of observations.

In the above tests, the improvements in the sorted process are not significant possibly because although the ensemble in the standard ES may be spanned in the unimportant components, this does not always happen, and the sorted process just improves these ensembles. However, the improvement is stable for different input variances, since the variance does not change the ratio of the eigenvalues. Moreover, the sorted process is more promising with larger correlation lengths, where the eigenvalues decay faster. The improvements in the 1D (about 10%) and 2D (about 15%) tests suggest that the improvement could be more obvious for complicated models. Note that the computational cost in the sorting process is negligible, indicating the improvement is still favorable.

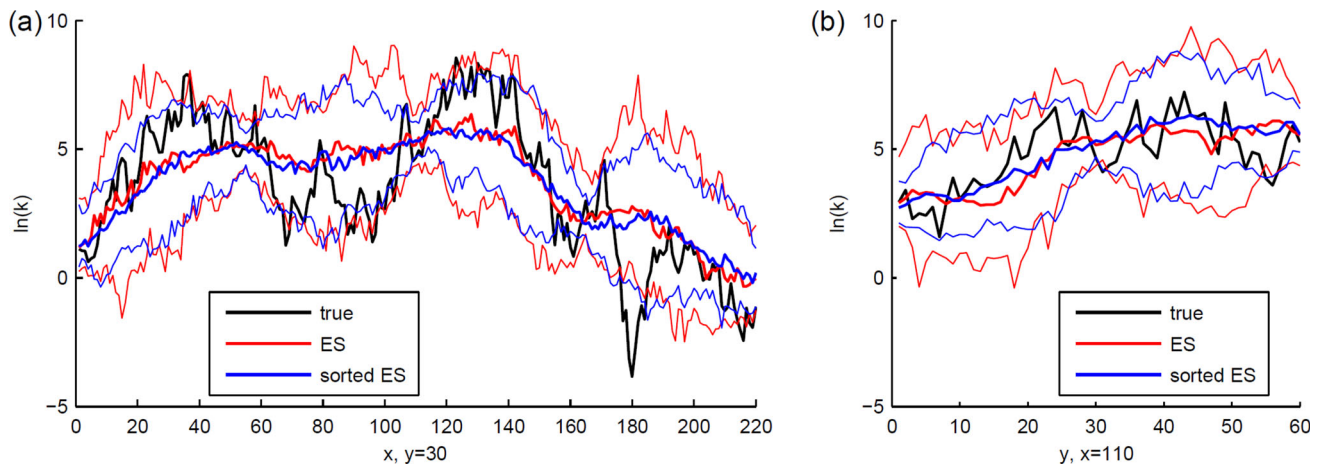
Note that the dominant components as we presented here do not necessarily have the largest impact in the observations. This is because the dominant components from the principle component analysis only account for the prior information but not the model or the likelihood. Actually, sensitivity analysis is required to determine the most important components, but with additional cost in forward simulations. If such a sensitivity analysis is affordable, the sorted process can be applied as well, in which the samples with larger variances should be assigned to more sensitive components.

## 5 Conclusions

Application of the ensemble smoother (ES) to groundwater model calibration and data assimilation has been investigated in recent years. In this work, we proposed a sorted ES by adding a sorting process in initial sampling to improve accuracy. Specifically, we perform principle component analysis

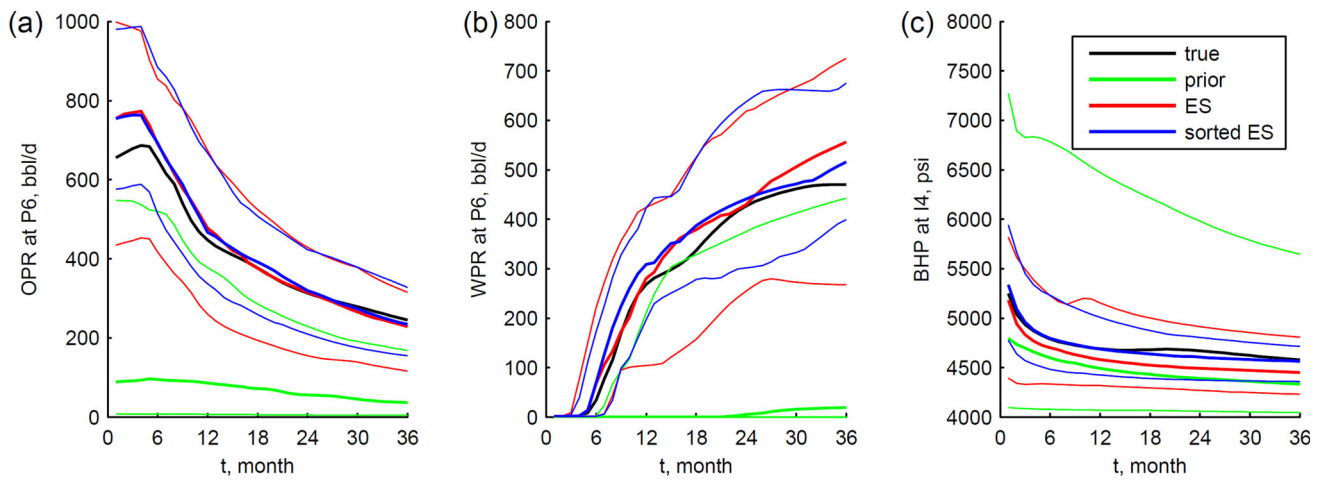


**Fig. 11** Percentiles of the mean (of 50 samples) log-permeability  $\ln(k)$  after matching the history: the left column is from the ES; the right column is from the sorted ES



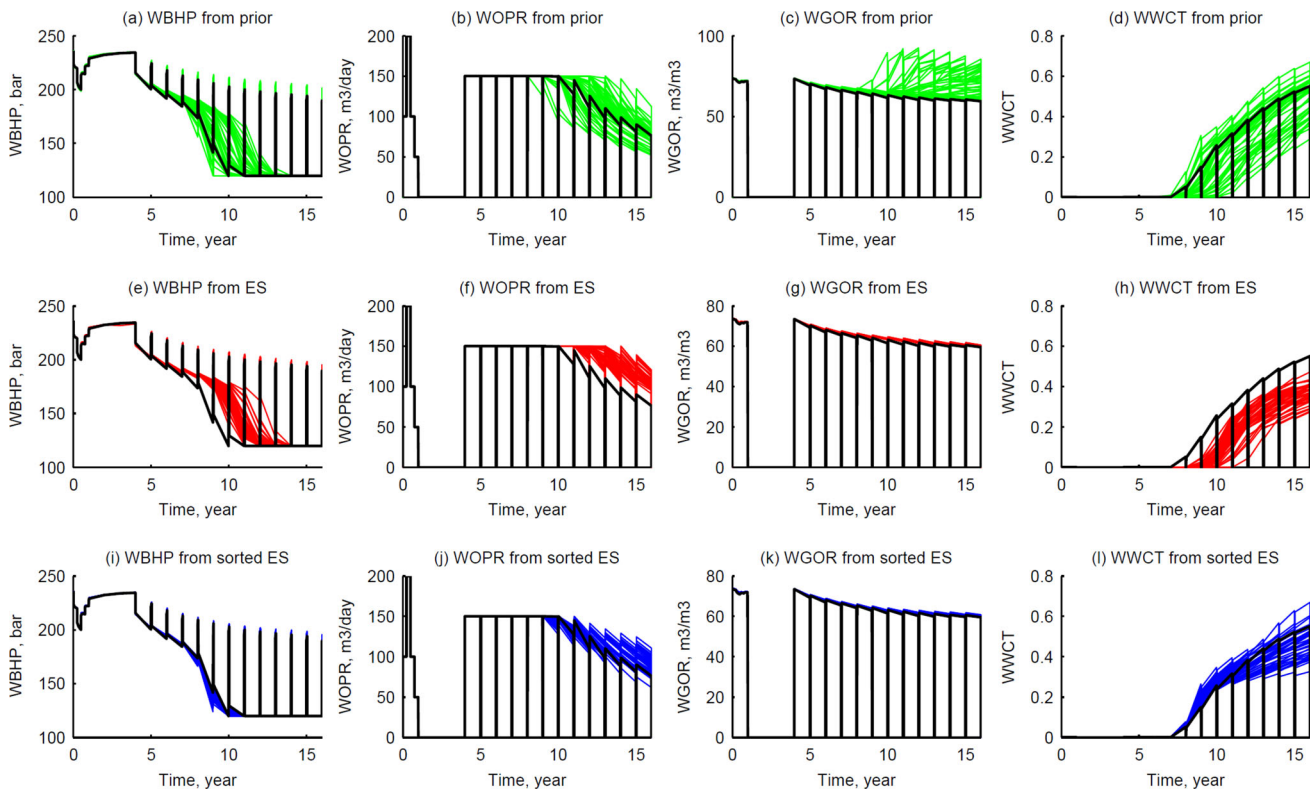
**Fig. 12** Percentiles of the mean log-permeability  $\ln(k)$  after matching the history along: **a** horizontal line  $y = 30$ ; **b** vertical line  $x = 110$ . The black line shows the true reference. The red lines show the P10, P50

and P90 (over 1,000 ensembles) of the mean estimation in ES. The blue lines show the P10, P50 and P90 of the mean estimation in sorted ES



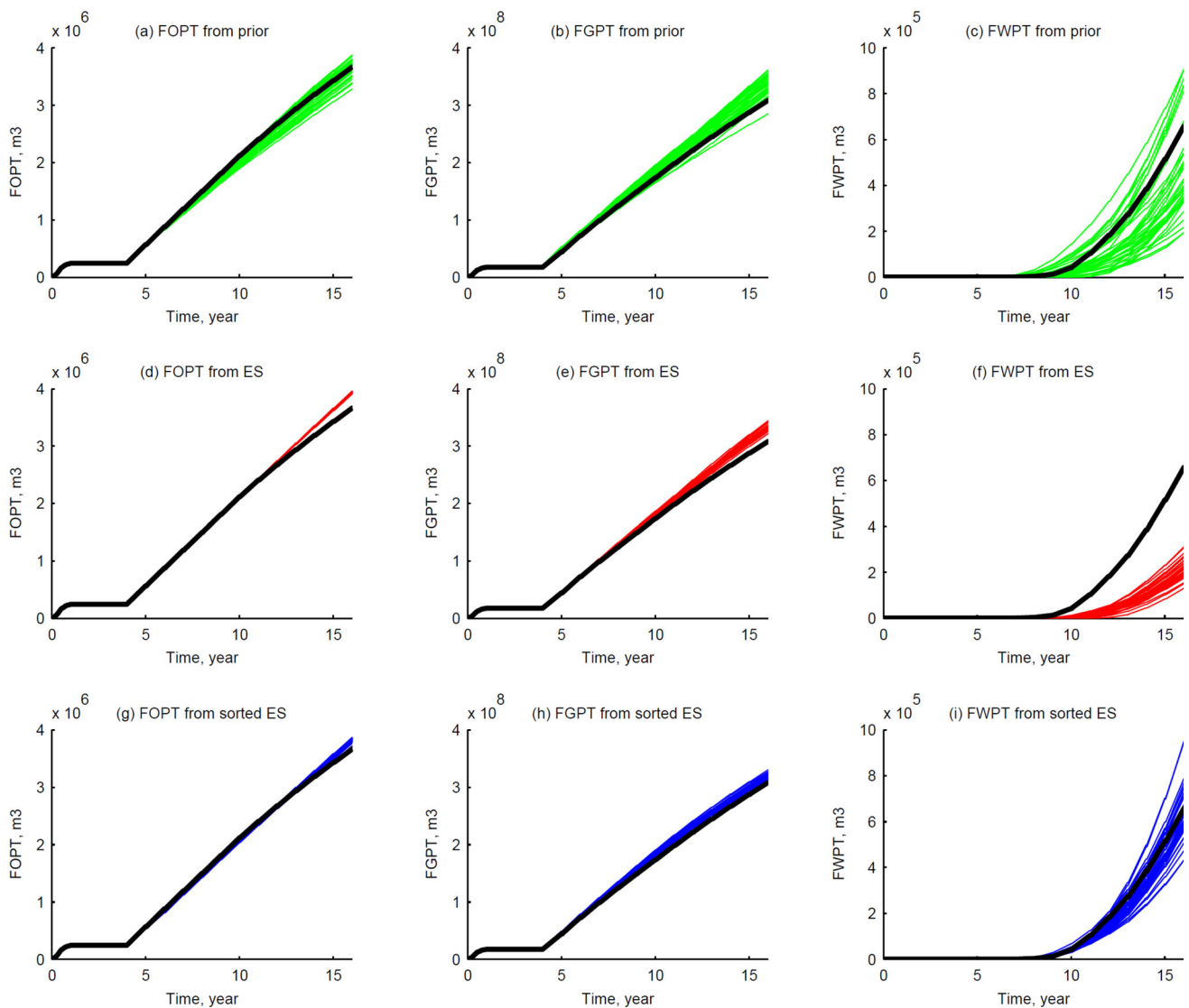
**Fig. 13** Percentiles of observations: **a** oil production rate at P6; **b** water production rate at P6; **c** bottom hole pressure at I4. The black line shows the exact result from the true reference. The green lines show the P10, P50 and P90 (over 1,000 ensembles) of the mean estimation in prior.

The red lines show the P10, P50 and P90 of the mean estimation in ES. The blue lines show the P10, P50 and P90 of the mean estimation in sorted ES



**Fig. 14** History matching and prediction of well PRO-11: the first row is from the prior; the second row is from the ES; the third row is from the sorted ES. The first column is the well bottom hole pressure; the

second column is the well oil production rate; the third column is the well gas–oil ratio; and the fourth column is the well water cut



**Fig. 15** History matching and prediction of whole field: the first row is from the prior; the second row is from the ES; the third row is from the sorted ES. The first column is the field oil production total; the second

column is the field gas production total; the third column is the field water production total

and assign the dimensions with large variances to the dominant components. By this means, the dominant components are more likely to be updated than other components, and thus, there is a higher chance that the main features are captured.

Four numerical tests were presented to evaluate the performance of the ES and the sorted ES, from the simple synthetic test to the three-dimensional black-oil benchmark test. It was seen that the sorted ES provided smaller root mean square errors than the standard ES. The error reduction mainly came from the reduction in the ensemble mean error, while keeping a similar spread value. In addition, we analyzed the effect of prior statistics and found that the improvement is not sensitive to the variance of the input parameter, but is more obvious

if the correlation length is larger. In conclusion, the sorting process improved the ES estimation, especially when the time-lapse data are observed. This idea can also be extended to other ensemble-based methods such as ensemble Kalman filter.

### References

- Zhong, H.; Li, Y.; Zhang, W.; Yin, H.; Lu, J.; Guo, D.: Microflow mechanism of oil displacement by viscoelastic hydrophobically associating water-soluble polymers in enhanced oil recovery. *Polymers* **10**, 628 (2018)
- Chen, S.Y.; Hsu, K.C.; Fan, C.M.: Improvement of generalized finite difference method for stochastic subsurface flow modeling. *J. Comput. Phys.* **429**, 110002 (2021)

3. Lei, G.; Liao, Q.; Lin, Q.; Zhang, L.; Xue, L.; Chen, W.: Stress dependent gas-water relative permeability in gas hydrates: a theoretical model. *Adv. Geo Energy Res.* **4**, 326–338 (2020)
4. Wang, S.; Qin, C.; Feng, Q.; Javadpour, F.; Rui, Z.: A framework for predicting the production performance of unconventional resources using deep learning. *Appl. Energy* **295**, 117016 (2021)
5. El-Amin, M.; Kou, J.; Sun, S.; Salama, A.: Adaptive time-splitting scheme for two-phase flow in heterogeneous porous media. *Adv. Geo Energy Res.* **1**, 182–189 (2017)
6. Rostami, A.; Daneshi, A.; Miri, R.: Proposing a rigorous empirical model for estimating the bubble point pressure in heterogeneous carbonate reservoirs. *Adv. Geo Energy Res.* **4**, 126–134 (2020)
7. Liao, Q.; Zeng, L.; Chang, H.; Zhang, D.: Efficient history matching using the Markov-chain Monte Carlo method by means of the transformed adaptive stochastic collocation method. *SPE J.* **24**, 1468–1489 (2019)
8. Yang, E.; Fang, Y.; Liu, Y.; Li, Z.; Wu, J.: Research and application of microfoam selective water plugging agent in shallow low-temperature reservoirs. *J. Pet. Sci. Eng.* **193**, 107354 (2020)
9. Wang, Y.; Li, X.J.: Physicochemical modeling of barium and sulfate transport in porous media and its application in seawater-breakthrough monitoring. *SPE J.* **26**, 1–22 (2021)
10. Tang, Y.; Hou, C.; He, Y.; Wang, Y.; Chen, Y.; Rui, Z.: Review on pore structure characterization and microscopic flow mechanism of CO<sub>2</sub> flooding in porous media. *Energy Technol.* **9**, 2000787 (2021)
11. Reichle, R.H.; McLaughlin, D.B.; Entekhabi, D.: Hydrologic data assimilation with the ensemble Kalman filter. *Mon. Weather Rev.* **130**, 103–114 (2002)
12. Houtekamer, P.; Zhang, F.: Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **144**, 4489–4532 (2016)
13. Chen, Y.; Zhang, D.: Data assimilation for transient flow in geologic formations via ensemble Kalman filter. *Adv. Water Res.* **29**, 1107–1122 (2006)
14. Xue, L.; Zhang, D.: A multimodel data assimilation framework via the ensemble Kalman filter. *Water Resour. Res.* **50**, 4197–4219 (2014)
15. Xue, L.; Zhang, D.; Guadagnini, A.; Neuman, S.P.: Multimodel Bayesian analysis of groundwater data worth. *Water Resour. Res.* **50**, 8481–8496 (2014)
16. Zhang, J.; Vrugt, J.A.; Shi, X.; Lin, G.; Wu, L.; Zeng, L.: Improving simulation efficiency of MCMC for inverse modeling of hydrologic systems with a Kalman-inspired proposal distribution. *Water Resour. Res.* **56**, e2019WR025474 (2020)
17. Zhu, P.; Shi, L.; Zhu, Y.; Zhang, Q.; Huang, K.; Williams, M.: Data assimilation of soil water flow via ensemble Kalman filter: infusing soil moisture data at different scales. *J. Hydrol.* **555**, 912–925 (2017)
18. Aanonsen, S.I.; Nævdal, G.; Oliver, D.S.; Reynolds, A.C.; Vallès, B.: The ensemble Kalman filter in reservoir engineering—a review. *SPE J.* **14**, 393–412 (2009)
19. Gu, Y.; Oliver, D.S.: An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. *SPE J.* **12**, 438–446 (2007)
20. Liao, Q.; Zhang, D.: Data assimilation for strongly nonlinear problems by transformed ensemble Kalman filter. *SPE J.* **20**, 202–221 (2015)
21. Oliver, D.S.; Chen, Y.: Recent progress on reservoir history matching: a review. *Comput. Geosci.* **15**, 185–221 (2011)
22. Zeng, L.; Zhang, D.: A stochastic collocation based Kalman filter for data assimilation. *Comput. Geosci.* **14**, 721–744 (2010)
23. Liao, Q.; Alsamadony, K.; Lei, G.; Awotunde, A.; Patil, S.: Reservoir history matching by ensemble smoother with principle component and sensitivity analysis for heterogeneous formations. *J. Pet. Sci. Eng.* **198**, 108140 (2021)
24. Gu, Y.; Oliver, D.S.: The ensemble Kalman filter for continuous updating of reservoir simulation models. *J. Energy Resour.* **128**, 79–87 (2006)
25. Li, G.; Reynolds, A.C.: Iterative ensemble Kalman filters for data assimilation. *SPE J.* **14**, 496–505 (2009)
26. Bailey, R.; Baù, D.: Ensemble smoother assimilation of hydraulic head and return flow data to estimate hydraulic conductivity distribution. *Water Resour. Res.* **46**, W12543 (2010)
27. Emerick, A.A.; Reynolds, A.C.: Ensemble smoother with multiple data assimilation. *Comput. Geosci.* **55**, 3–15 (2013)
28. Evensen, G.: *Data Assimilation: the Ensemble Kalman Filter*. Springer, Berlin (2007)
29. Skjervheim, J.A.; Evensen, G.: An ensemble smoother for assisted history matching. In: *SPE Reservoir Simulation Symposium*, The Woodlands, Texas (2011)
30. Evensen, G.; Eikrem, K.S.: Conditioning reservoir models on rate data using ensemble smoothers. *Comput. Geosci.* (2018). <https://doi.org/10.1007/s10596-018-9750-8>
31. Chang, H.; Zhang, D.: History matching of statistically anisotropic fields using the Karhunen-loeve expansion-based global parameterization technique. *Comput. Geosci.* **18**, 265–282 (2014)
32. He, J.; Sarma, P.; Durloufsky, L.J.: Reduced-order flow modeling and geological parameterization for ensemble-based data assimilation. *Comput. Geosci.* **55**, 54–69 (2013)
33. Reynolds, A.C.; He, N.; Chu, L.; Oliver, D.S.: Reparameterization techniques for generating reservoir descriptions conditioned to variograms and well-test pressure data. *SPE J.* **1**, 413–426 (1996)
34. Sarma, P.; Durloufsky, L.J.; Aziz, K.: Kernel principal component analysis for efficient, differentiable parameterization of multipoint geostatistics. *Math. Geosci.* **40**, 3–32 (2008)
35. Oliver, D.S.; Chen, Y.: Improved initial sampling for the ensemble Kalman filter. *Comput. Geosci.* **13**, 13 (2009)
36. Uyeda, J.C.; Caetano, D.S.; Pennell, M.W.: Comparative analysis of principal components can be misleading. *Syst. Biol.* **64**, 677–689 (2015)
37. Bear, J.: *Dynamics of Fluids in Porous Media*. Elsevier, New York (1972)
38. Dagan, G.: *Flow and Transport in Porous Formations*. Springer, New York (1989)
39. Ghanem, R.G.; Spanos, P.D.: *Stochastic Finite Elements: a Spectral Approach*. Springer, New York (1991)
40. Jafarpour, B.; McLaughlin, D.B.: History matching with an ensemble Kalman filter and discrete cosine parameterization. *Comput. Geosci.* **12**, 227–244 (2008)
41. Floris, F.J.T.; Bush, M.D.; Cuypers, M.; Roggero, F.; Syversveen, A.-R.: Methods for quantifying the uncertainty of production forecasts. *Pet. Geosci.* **7**, S87–S96 (2001)
42. Gu, Y.; Oliver, D.S.: History matching of the PUNQ-S3 reservoir model using the ensemble Kalman filter. *SPE J.* **10**, 51–65 (2005)
43. PUNQ-S3 model: [https://www.imperial.ac.uk/earth-science/research/research-groups/perm/standard-models/\(2021\)](https://www.imperial.ac.uk/earth-science/research/research-groups/perm/standard-models/(2021))

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.