Stochastic Environmental Research and Risk Assessment (2018) 32:1919–1937
https://doi.org/10.1007/s00477-018-1521-5

ORIGINAL PAPER

Joint inversion of physical and geochemical parameters in groundwater models by sequential ensemble-based optimal design

Tian Lan[1] · Xiaoqing Shi[1] · Beilei Jiang[2] · Yuanyuan Sun[1] · Jichun Wu[1]

Published online: 21 February 2018

Abstract

Joint inversion of physical and geochemical parameters in groundwater reactive transport models is still a great challenge due to the intrinsic heterogeneities of natural porous media and the scarcity of observation data. In this study, we make use of a sequential ensemble-based optimal design (SEOD) method to jointly estimate physical and geochemical parameters of groundwater models. The effectiveness and efficiency of the SEOD method are illustrated by the comparison between the sequential optimization strategy and the conventional strategy (using fixed sampling locations) for two synthetic cases. Since the SEOD method is an optimization method based on the ensemble Kalman filter (EnKF), it invokes the time-consuming genetic algorithm at every assimilation step of the EnKF to obtain the optimal sampling locations. To enhance its computational efficiency, we improve the SEOD method by replacing the EnKF with the ensemble smoother with multiple data assimilation. Furthermore, the influence factors of the original and improved SEOD method are also discussed. Our results show that the SEOD method provides an effective designed sampling strategy to accurately estimate heterogeneous distribution of physical and geochemical parameters. Moreover, the improved SEOD method is more advantageous than the original one in computational efficiency, making this SEOD framework more promising for future application.

Keywords Optimal sampling strategy · Physical and geochemical heterogeneity · Parameter estimation · Reactive transport model · Data assimilation

1 Introduction

Joint inversion of physical and geochemical parameters in groundwater reactive transport models is critical for reliable contaminant plume prediction, remediation and management, but it is still a great challenge due to the intrinsic heterogeneities of natural porous media and the scarcity of observation data. The subsurface environment is highly variable in its physical and chemical composition. Heterogeneity of physical parameters (e.g., hydraulic conductivity) has been shown to exert a key control on the mixing and spreading of conservative solutes (Dagan 1984; Rubin 1991; Sudicky 1986). For reactive solutes, their transport and reactions are simultaneously influenced by geochemical parameters (Atchley et al. 2014; Li et al. 2010; Scheibe et al. 2006). Similar to physical heterogeneity, the heterogeneity of geochemical parameters exists as well, which may be caused, for example, by spatial variability in the activity of bacteria related to biodegradation (Fennell et al. 2001; Sandrin et al. 2004). Therefore, it is important to jointly estimate the spatial distribution of physical and geochemical parameters in groundwater reactive transport models.

Inverse methods are often used by conditioning on observation data to characterize the spatial variation of parameters, which has been extensively investigated in the literature (e.g., Carrera et al. 2005; Dagan 1985; Doherty

✉ Xiaoqing Shi
shixq@nju.edu.cn

✉ Jichun Wu
jcwu@nju.edu.cn

[1] Key Laboratory of Surficial Geochemistry, Ministry of Education and School of Earth Sciences and Engineering, Nanjing University, Nanjing 210023, China

[2] National Key Laboratory of Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China

2004; Gómez-Hernández et al. 2003; Hendricks Franssen et al. 2009; Neuman 1973; Oliver et al. 1997; Zhou et al. 2014). The ensemble Kalman filter (EnKF, Evensen 2003, 2009) is one of the most popular inverse methods over the last decade (Aanonsen et al. 2009; Oliver and Chen 2011; Zhou et al. 2014), recently used in parameter estimation and state prediction (Chen and Zhang 2006; Huang et al. 2009; Tong et al. 2010). It is a variant of the Kalman Filter (KF, Kalman 1960) based on the Monte Carlo method. Unlike the KF, the EnKF was developed for nonlinear problems (Evensen 2003, 2009), its efficiency and effectiveness in nonlinear problems with high dimensionality have been illustrated (Chen and Zhang 2006; Hendricks Franssen and Kinzelbach 2008; Moradkhani et al. 2005; Sorensen et al. 2004). In addition to the EnKF, the Ensemble Smoother (ES, Van Leeuwen and Evensen 1996) and its iterative variants, like the Ensemble Smoother with multiple data assimilation (ES-MDA, Emerick and Reynolds 2013), are popular as well. Unlike the EnKF, the ES and the ES-MDA perform global update rather than sequential update during the data assimilation, avoiding restarting models again and again, so they are of more simplicity and computational efficiency than the EnKF.

Much research has focused on developing better methods based on the EnKF to broaden its implementation scale and improve its accuracy (Chen and Oliver 2010; Emerick and Reynolds 2011; Gu and Oliver 2007; Li and Reynolds 2009), with the sampling locations fixed during the data assimilation (called the conventional strategy in the following discussion). However, it is intuitive that the data worth of measurements is dramatically influenced by sampling locations, and the parameter estimation result can be improved if the measurements are more informative even though the number of sampling locations is the same. There has been much research revealed the effect of sampling strategies on the parameter uncertainty and predictive uncertainty in groundwater models (Carrera and Neuman 1986; Cleveland and Yeh 1990; Knopman and Voss 1987; Nowak et al. 2010; Sun and Yeh 2007; Ushijima and Yeh 2015; Zhang et al. 2015). In view of these two aspects, Man et al. (2016) integrated a sequential optimal design and the information theory into the EnKF framework seamlessly to provide the most informative measurements for more accurate parameter estimation, and proposed a sequential ensemble-based optimal design (SEOD) method. Man et al. (2016) demonstrated the effectiveness of this method by estimating only physical parameters in unsaturated flow models, assimilating only piezometric head data. However, the SEOD method developed by Man et al. (2016) invokes the optimization algorithm (the genetic algorithm) at each assimilation step, so its computational efficiency is not very satisfying. Furthermore, to the best of our knowledge, few studies have focused on joint inversion of physical and geochemical parameters by assimilating multiple kinds of data.

The objective of this study is to estimate both physical and geochemical parameters accurately in groundwater models by using the recent proposed SEOD method, and to enhance the computational efficiency of the SEOD method by replacing the EnKF with the ES-MDA. The rest of the paper is organized as follows. In Sect. 2, the groundwater reactive transport model and the SEOD method are described. In Sect. 3, synthetic one-dimensional and two-dimensional groundwater reactive transport model cases are constructed to jointly estimate the physical and geochemical parameters by using the SEOD method. In Sect. 4, the comparison between the sequential optimization strategy and the conventional strategy, and the effects of the ensemble size and the number of optimal sampling locations are discussed. Furthermore, we improve the SEOD method by replacing the EnKF with the ES-MDA to enhance its computational efficiency, and make a comparison of the original and the improved SEOD method in Sect. 4.4. Conclusions are summarized in Sect. 5.

2 Methodologies

2.1 Groundwater reactive transport model

In this work, transient flow is assumed, as the following governing equation (Bear 1972),

$$\nabla \cdot (\mathbf{K}\nabla H) + W = \mu_s \frac{\partial H}{\partial t} \tag{1}$$

where $\nabla\cdot$ is the divergence operator; ∇ is the gradient operator; $\boldsymbol{K}$ is the hydraulic conductivity (LT^{-1}); H is the hydraulic head (L); W is the volumetric injection (pumping) flow rate per unit volume of the aquifer (T^{-1}); μ_s is the specific storage of the aquifer (L^{-1}); t is the time (T).

The governing equation for the transport and reactions of aqueous species is defined as (Zheng 2006; Prommer and Post 2010):

$$\frac{\partial C_n}{\partial t} = \nabla \cdot (\mathbf{D} \cdot \nabla C_n) - \nabla \cdot (vC_n) + r_{\mathrm{reac},n} + \frac{q_s}{\theta} C_n^s \tag{2}$$

where C_n is the aqueous concentration of the nth component (ML^{-3}); t is the time (T); $\boldsymbol{D}$ is the diffusion coefficient ($\mathrm{L}^2\mathrm{T}^{-1}$); $v = (-K\nabla H)/\theta$ (LT^{-1}); $r_{reac,n}$ is the concentration change of the nth component caused by reactions (ML^{-3}); q_s is the volumetric flow rate per unit volume of the aquifer (T^{-1}); θ is the effective porosity; and C_n^s is the concentration of the source or sink flux of the nth component (ML^{-3}).

Equation (1) is solved by the numerical code MODFLOW-2000 (Harbaugh et al. 2000), and Eq. (2) is solved by the numerical code MT3DMS (Zheng 2006).

2.2 Sequential ensemble-based optimal design (SEOD) method

The sequential ensemble-based optimal design (SEOD) method is a new recently proposed optimal method based on the EnKF (Man et al. 2016). At each recursive step, the SEOD method provides an optimal sampling strategy, giving the maximum value of information metric. Then, the analysis equation of the EnKF is used to update estimated parameters by assimilating the most informative measurements, obtained based on the optimal sampling strategy.

In this work, relative entropy (*RE*), also known as the Kullback–Leibler divergence (Kullback 1997), is used to measure the information content of the posterior probability density function (pdf) relative to the prior pdf. If these two distributions are both *n*-dimensional Gaussian, *RE* between these two distributions is defined as:

$$RE = J_b + [\ln \det(\mathbf{BA}^{-1}) + \mathrm{Tr}(\mathbf{AB}^{-1}) - n]/2 \tag{3}$$

where $J_b = (\mathbf{a}-\mathbf{b})^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{a}-\mathbf{b})/2$ is the signal part of *RE*; $\det(\cdot)$ denotes the determinant; $\mathrm{Tr}(\cdot)$ denotes the trace; **a** and **A** denote the mean and covariance matrix of prior statistics respectively; **b** and **B** denote the mean and covariance matrix of posterior statistics respectively.

The loop of the SEOD method for parameter estimation is briefly recalled. More details can be found in Man et al. (2016). In the EnKF, all the parameters of interest **p** are augmented with state variables **h** into a joint state vector $\mathbf{x} = [\mathbf{p}\,\mathbf{h}]^{\mathrm{T}}$. Before the forecast step, an ensemble of N_e realizations of parameters is generated.

I. Forecast step

Rerun the forward model *G* from time 0 to time step *j* + 1 with parameters updated at time step *j* [Eq. (4)].

$$\mathbf{X}^{f}_{i,j+1} = G\left(\mathbf{x}^{a}_{i,j}\right), \quad i = 1, 2, \ldots, N_e \tag{4}$$

In the above equation, *i* is the ensemble member index, *j* is the time step index, superscripts *f* and *a* denote forecast and analysis, respectively.

II. Optimal design

Given a specific sampling strategy $\mathbf{H}'$, the possible realizations of measurements can then be expressed as $\mathbf{d}'_i = \mathbf{H}'\mathbf{x}^f_i + \xi_i$. With the realizations of measurements, the updated ensemble can be obtained from the Eq. (6). According to the prior and posterior statistics (mean and covariance), the information metrics *RE* of each candidate sampling strategies can be calculated. By comparing the *RE* values of different candidate sampling strategies, the optimal sampling design $\mathbf{H}_{\mathrm{opt}}$ can be determined by solving the following optimization problem [Eq. (5)] with the help of the genetic algorithm (GA, Whitley 1994).

$$\mathbf{H}_{\mathrm{opt}} = \arg\max RE(\mathbf{H}) \tag{5}$$

III. Analysis step

After obtaining the optimal sampling strategy, the actual measurements **d** can be obtained and used in the analysis step [Eq. (6)].

$$\mathbf{x}^{a}_{i,j+1} = \mathbf{x}^{f}_{i,j+1} + \mathbf{C}_{\mathrm{YD}}(\mathbf{C}_{\mathrm{DD}} + \mathbf{C}_{\mathrm{D}})^{-1}(\mathbf{d}_{obs_{i,j+1}} - \mathbf{d}_{i,j+1}), \quad i = 1, 2, \ldots, N_e \tag{6}$$

In the above equation, $\mathbf{C}_{\mathrm{YD}}$ is the cross-covariance matrix between the forecast state and the predicted data, $\mathbf{C}_{\mathrm{DD}}$ is the covariance matrix of the predicted data, $\mathbf{C}_{\mathrm{D}}$ is the covariance matrix of the measurements error, $\mathbf{d}_{obs}$ is the perturbed observations with noise of covariance $\mathbf{C}_{\mathrm{D}}$, and **d** is the predicted data.

After the analysis step, the updated ensemble X^{a} is obtained. Then, go back to step (I), the updated ensemble obtained this step is implemented for the next step.

To evaluate the performance of parameter estimation, two commonly used indicators, the *RMSE* and the *Ensemble Spread*, are defined as:

$$RMSE = \sqrt{\frac{1}{N_m}\sum_{i=1}^{N_e}(\overline{Y_i} - Y_i)^2} \tag{7}$$

$$Ensemble\ Spread = \sqrt{\frac{1}{N_m}\sum_{i=1}^{N_e} var(Y_i)} \tag{8}$$

where $\overline{Y}$ and Y are the estimated and the reference field respectively; $\mathrm{var}(Y)$ is the ensemble variance of the field; N_m is the total number of nodes in the study domain; N_e is the ensemble size; *i* is the node index. The *RMSE* measures the accuracy of the estimation, while the *Ensemble Spread* measures the uncertainty of the estimation.

3 Case studies

3.1 Case 1: One-dimensional synthetic case

In this case, a one-dimensional confined aquifer with a starting head of 100 m is constructed, in which saturated transient flow is assumed. As shown in Fig. 1a, we choose the horizontal aquifer to be 5 m × 150 m and the grid space to be 5 m both in horizontal *x* and *y* direction. Then, a Trichloroethylene (TCE) leaking area with an initial concentration of 1000 mg/L is introduced into the aquifer, and the degradation of TCE is assumed to follow first-order kinetic reaction. Furthermore, an injection well and a

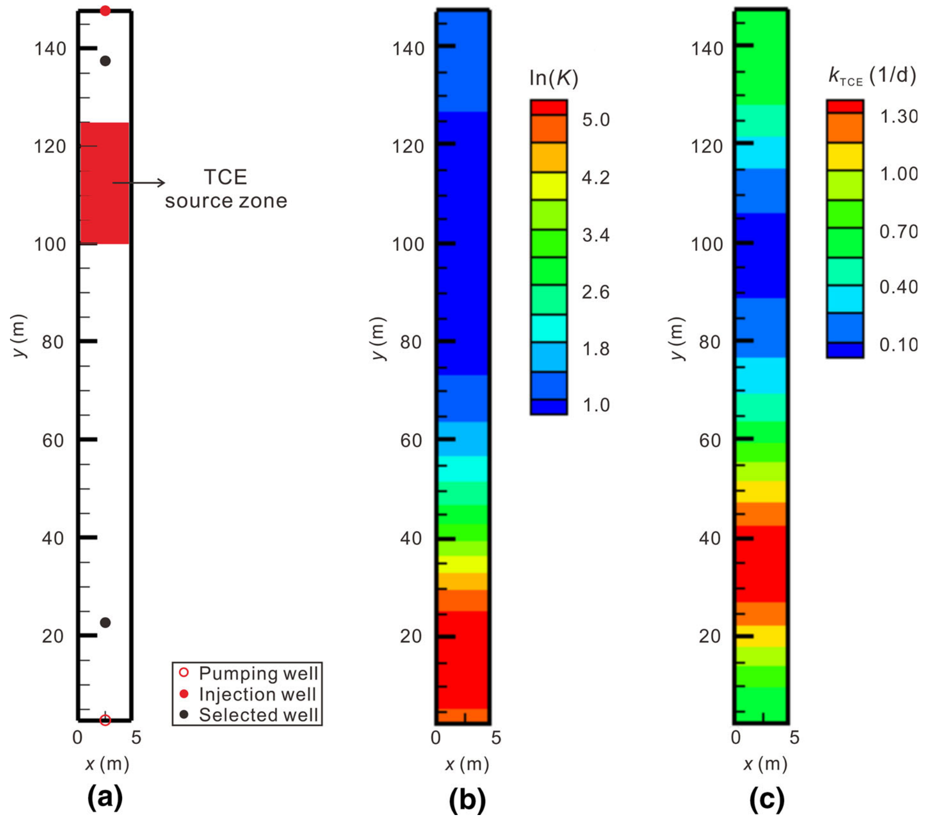


Fig. 1 The conceptual model (**a**), the reference fields of the hydraulic conductivity (**b**) and the first-order rate constant (**c**) for Case 1

pumping well are set upstream and downstream respectively, and all boundaries of the aquifer are assumed to be impermeable. In this case, the spatial distribution of the hydraulic conductivity (K) and the first-order rate constant (k_{TCE}) (Fig. 1b, c) are jointly estimated. At every assimilation step 2 optimal sampling locations are selected from 30 candidate locations to provide the most informative measurements. More details are given in Tables 1 and 2.

Table 1 Flow and transport parameters used in Case 1

Flow simulation	Transient state
Total simulation time (days)	10
Stress period	1
Time steps	100
Grid spacing (m)	5 × 5 × 5
Model length (m)	150
Model width (m)	5
Model height (m)	5
Starting head (m)	100
Porosity	0.3
Specific storage (m^{-1})	0.0001
Longitudinal dispersivity (m)	10
Injection rate per well (m^3/day)	50
Pumping rate per well (m^3/day)	45

The log saturated hydraulic conductivity $Y_1 = \ln(K)$ and the first-order rate constant $Y_2 = k_{\mathrm{TCE}}$ are assumed to be Gaussian distributed, with mean $\mu_{Y_1} = 1$ and $\mu_{Y_2} = 0.17$ and variance $\sigma_{Y_1} = 1$ and $\sigma_{Y_2} = 0.47$ respectively. Two arbitrary locations (x_1, y_1) and (x_2, y_2) in the random field are assumed to be correlated in the following form:

$$C_Y(x_1, x_2) = C_Y(x_1, y_1; x_2, y_2) = \sigma^2 \exp\left[-\frac{|x_1 - x_2|}{\lambda_x} - \frac{|y_1 - y_2|}{\lambda_y}\right] \tag{9}$$

where the horizontal correlation length $\lambda_x = 5$ m, and the vertical correlation length $\lambda_y = 30$ m. Here we use the Karhunen–Loeve (K–L) expansion (Zhang and Lu 2004) to parameterize the random field so as to achieve the reference fields and initial ensemble members. The measurement errors of the head and concentration data are assumed to follow the standard normal distribution with the standard deviation of 0.01 m and 10^{-6} mg/L respectively. Since the SEOD method is sequential, the uncertainty changes as real-time measurements are assimilated, which leads to the optimal sampling locations changing with time. The optimal sampling locations at 10 assimilation steps are shown in Fig. 2. It shows that the optimal sampling locations change with the flow and concentration fields so as to obtain the most informative measurements. It is interesting to note that the most optimal sampling locations are located at the front of

Table 2 Data assimilation related parameters used in different cases

Case name	Dimension	Number of ensemble (N_e)	Number of assimilation step (N_s)	Optimize or not	Number of optimal sampling locations
Case 1	1	300	10	Y	2
Case 2	2	100	10	Y	2
Case 3	2	50	10	Y	2
Case 4	2	300	10	Y	2
Case 5	2	500	10	Y	2
Case 6	2	1000	10	Y	2
Case 7	2	100	10	Y	1
Case 8	2	100	10	Y	5
Case 9	2	100	10	Y	10
Case 10	2	100	10	Y	20
Case 11	2	100	10	N	(2 fixed)
Case 12	2	100	10	N	(10 fixed)
Case 13	2	100	10	N	(20 fixed)
Case 14	2	100	8 (50, 50)[a]	Y	2
Case 15	2	100	8 (50, 50)	Y	5
Case 16	2	100	12 (20, 30, 50)	Y	5
Case 17	2	100	12 (50, 30, 20)	Y	5

[a] The numbers in the parentheses are the group division of observation time, and the number in front of the parentheses is the number of assimilation steps. For example, 8(50, 50) represents that there are 8 steps in the assimilation and the observation time is divided into two groups with each group having an observation time of 50 days

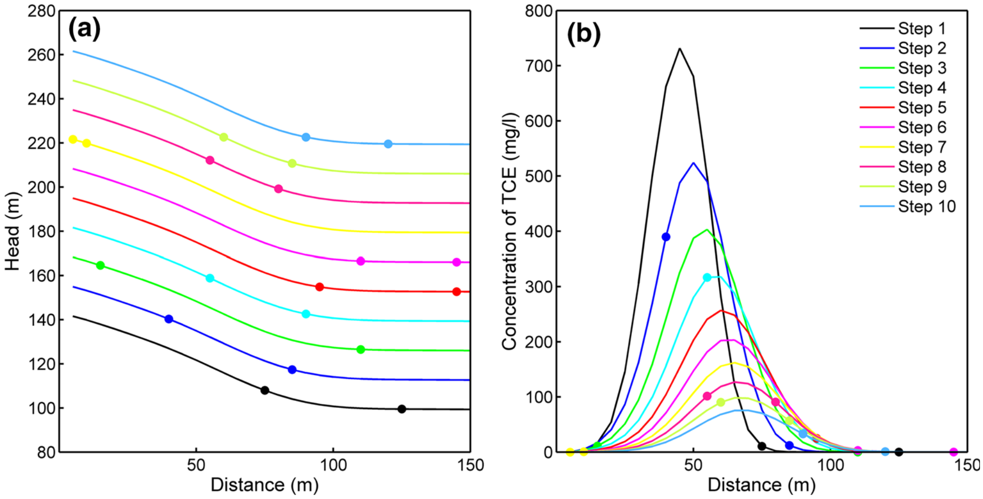


Fig. 2 The calculated flow field (**a**) and concentration field (**b**) in Case 1. The circles denote the optimal sampling locations proposed by the SEOD method

the contaminant plume, which means that these locations can provide the most informative measurements.

In Fig. 3, we plot the curves of the ensemble mean and the standard deviation at different assimilation steps. It shows that, for both Y_1 and Y_2, the ensemble mean at the final assimilation step is very close to the reference field. Furthermore, the ensemble standard deviation is high at the early steps; however, it reduces dramatically after assimilating the most informative measurements from the optimal sampling locations. As shown in Fig. 4, the *RMSE* and the *Ensemble Spread* of Y_1 and Y_2 decrease as the assimilation step increases, which also suggests that estimated fields are close to their reference fields and of low uncertainty.

To further illustrate the accuracy and uncertainty of the estimation, we also evaluate the performance of data match and model prediction. Considering the limitation of space, two wells (marked with black circles in Fig. 1a) are selected randomly to show the following evaluation results.

The initial and final ensembles of Y_1 and Y_2 are taken into the synthetic model respectively to calculate the head and

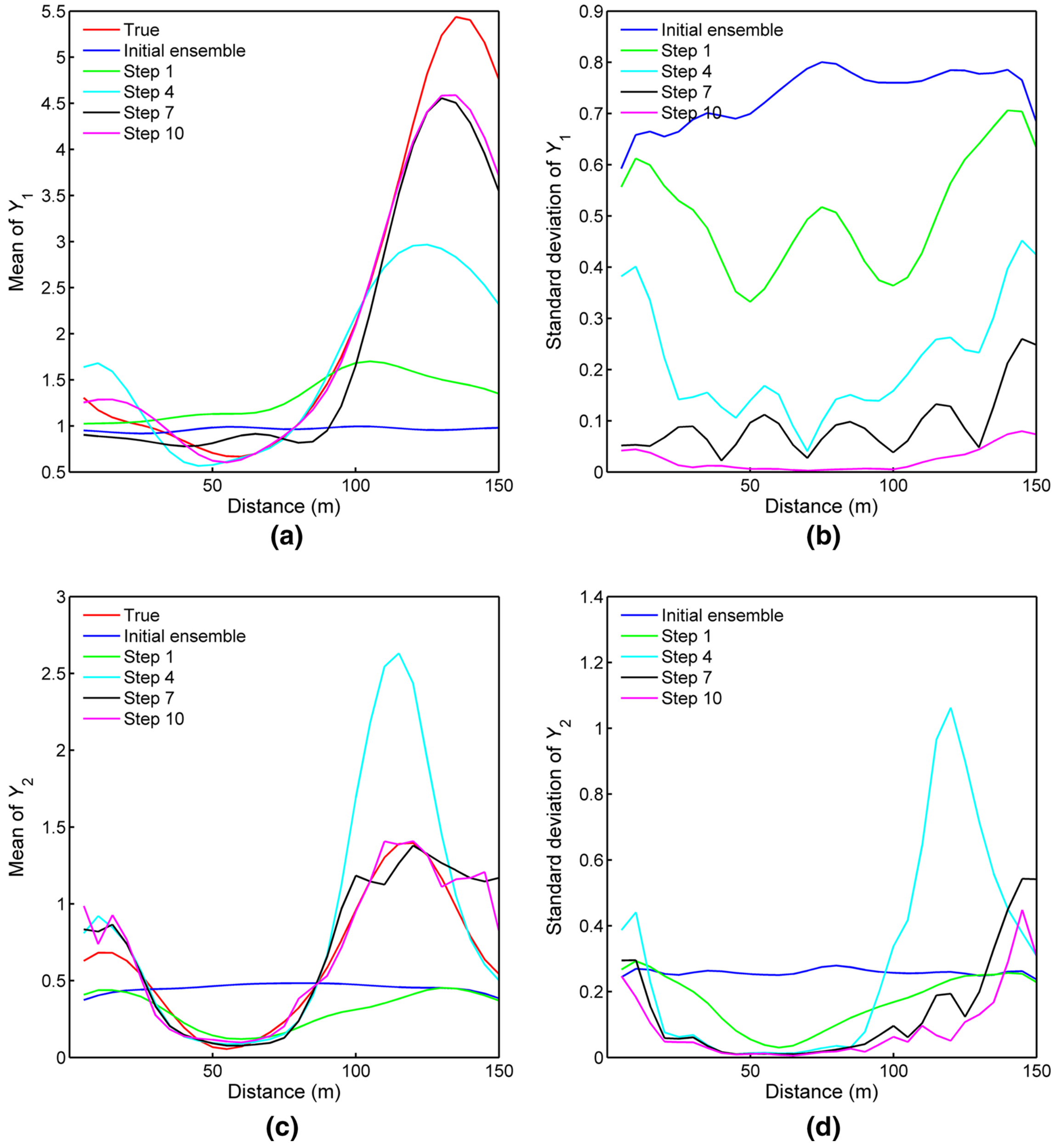


Fig. 3 The ensemble mean and the standard deviation of field Y_1 and Y_2 in Case 1. (**a**) and (**b**) are the ensemble mean and the standard deviation of field Y_1 respectively, while (**c**) and (**d**) are the ensemble mean and the standard deviation of field Y_2 respectively

concentration data, which are then compared with the real observations. After data assimilation, the data calculated by the final ensemble become closer to the real observations, as shown in Figs. 5 and 6. Overall, model uncertainty is significantly reduced after assimilating the most informative measurements from the optimal sampling locations.

3.2 Case 2: Two-dimensional synthetic case

In this case, saturated transient flow is assumed in a two-dimensional confined aquifer with a starting head of 50 m. As shown in Fig. 7, we choose the horizontal aquifer to be 105 m × 65 m and the grid space to be 5 m both in horizontal x and y direction. Three TCE leaking sources with the constant injection flow of 80 m^3/day and the constant concentration of 50 mg/L per well are set upstream in the aquifer. Furthermore, the liner sorption reaction of the TCE is considered in this case. Besides three injection wells, two pumping wells with the constant pumping flow of 120 m^3/day per well are set downstream (Fig. 7). In addition, all boundaries of the aquifer are assumed to be impermeable. In this case, the hydraulic conductivity

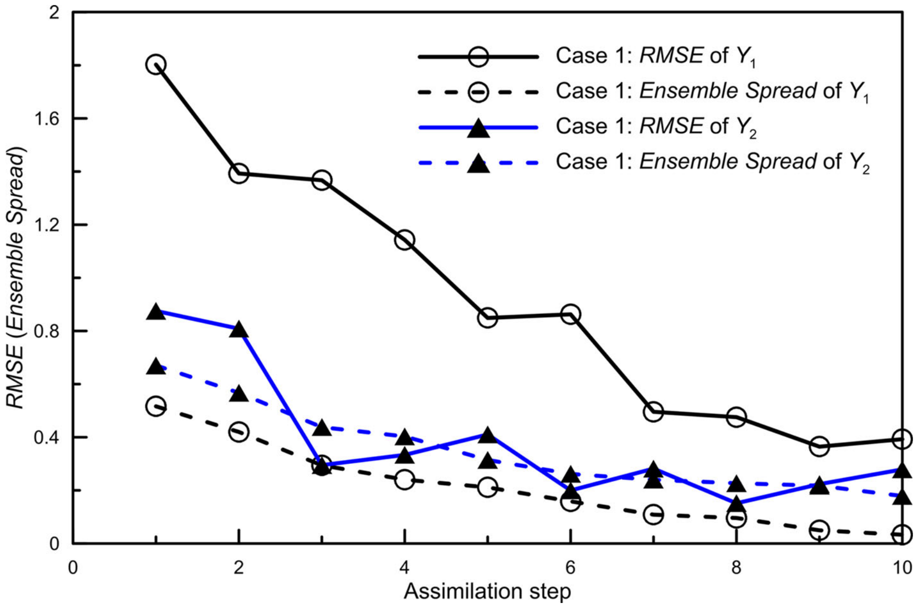


Fig. 4 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for field Y_1 and Y_2 in Case 1

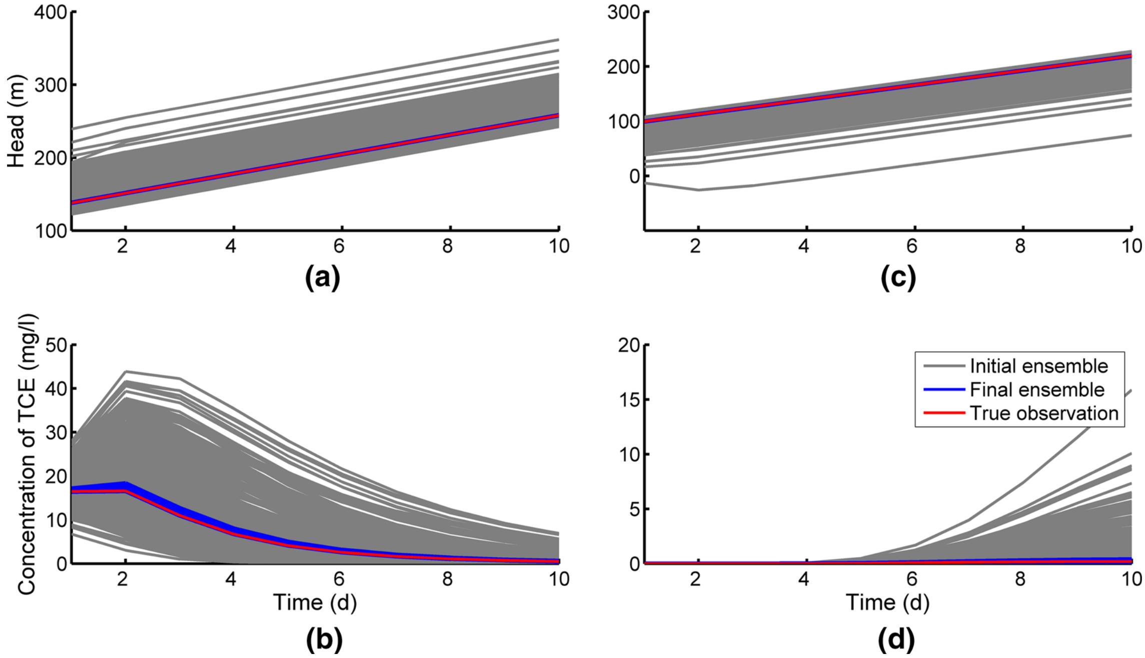


Fig. 5 The performance of data match. (**a**), (**c**) show the data match of the head of two selected wells respectively, while (**b**), (**d**) show the data match of the TCE concentration data of two selected wells respectively

(K) and the liner sorption constant (k_d) are assumed to be spatially heterogeneous (Fig. 9). At every assimilation step, 2 optimal sampling locations are selected from 77 candidate locations (Fig. 7) to provide the most informative measurements. More details are given in Tables 2 and 3.

The log saturated hydraulic conductivity $Y_1 = \ln(K)$ is assumed to be Gaussian distributed with mean $\mu_{Y_1} = 1$ and variance $\sigma_{Y_1} = 1$. The horizontal and vertical correlation lengths of Y_1 are 40 and 20 m respectively. With these statistics, the reference field and initial ensemble members of Y_1 can be generated by the K–L decomposition based on Eq. (9). For field Y_2, it is assumed that there is a positive correlation between $Y_2 = \ln(k_d)$ and Y_1, i.e. $Y_2 = 0.5 \times Y_1 - 15.95$, on which the generation of reference field Y_2 and its initial ensemble members are based. In addition, the measurement errors of the head and concentration data are assumed to follow the standard normal distribution with the standard deviation of 0.01 m and 10^{-6} mg/L respectively.

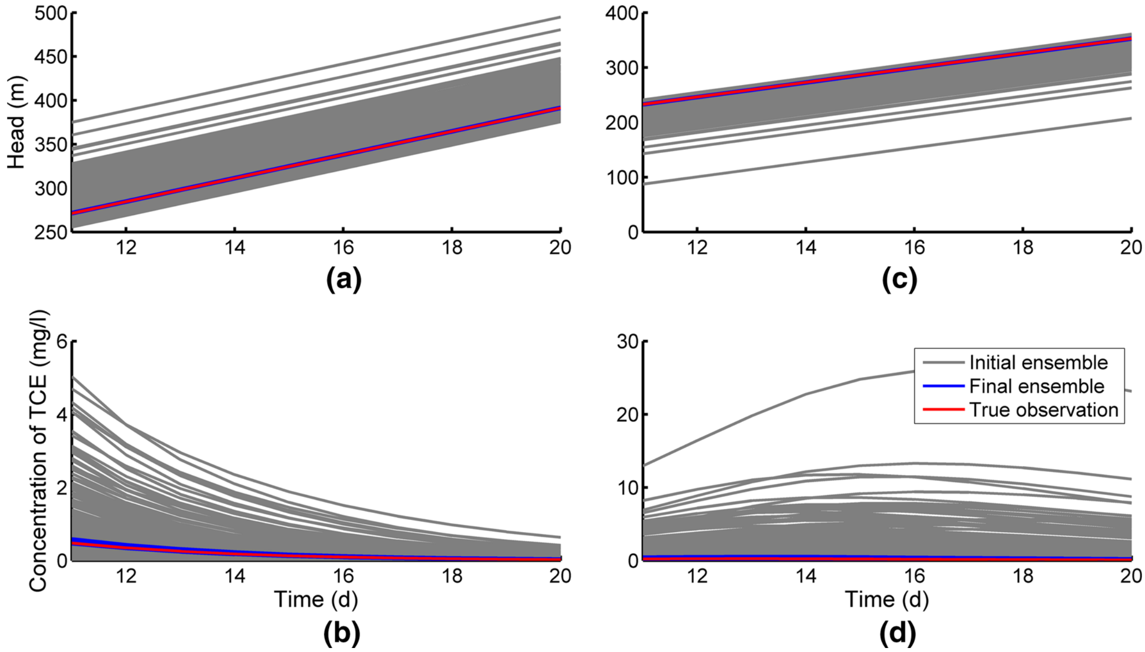
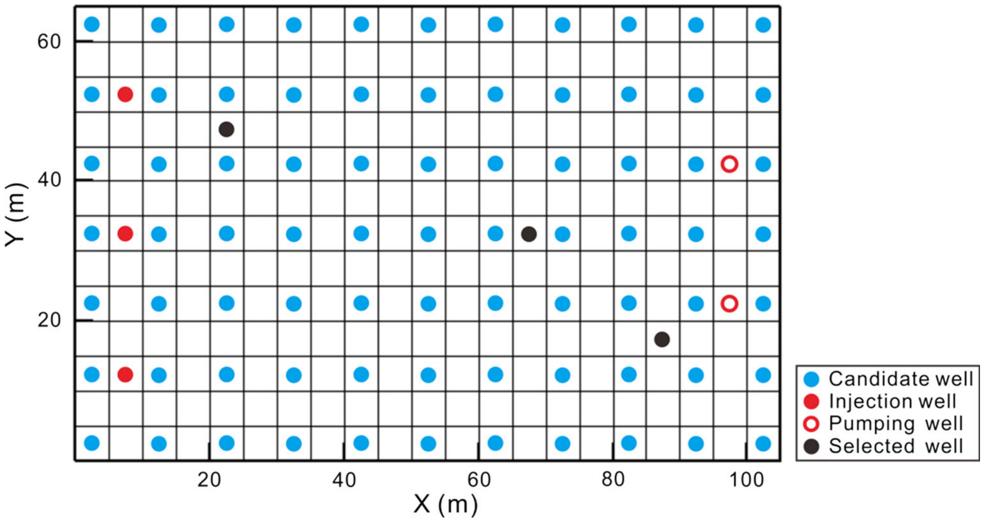


Fig. 6 The performance of model prediction. (**a**), (**c**) show the prediction of the head of two selected wells respectively, while (**b**), (**d**) show the prediction of the TCE concentration data of two selected wells respectively

Fig. 7 The schematic of the two-dimensional conceptual model



In Fig. 8, we plot the optimal sampling locations at each assimilation step. It shows the tendency that locations of large gradient are more likely to be selected as the optimal sampling locations. Overall, it shows that the choice of the optimal sampling locations at each assimilation step changes with the flow and concentration fields to obtain the most informative measurements.

The contour maps of the ensemble mean and the standard deviation at different assimilation steps are plotted in Fig. 9. It shows that, for both Y_1 and Y_2, the contour maps of the ensemble mean exhibit a pattern very similar to the reference fields. Even just after 4 assimilation steps, the contour maps of the ensemble mean recover the major features of the reference fields of Y_1 and Y_2. Furthermore, the ensemble standard deviations reduce dramatically after assimilating the most informative measurements from the optimal sampling locations, indicating that the optimal sampling strategy does play a crucial role in the model inversion though only 2 sampling locations are selected at every assimilation step.

The *RMSE* and the *Ensemble Spread* of Y_1 and Y_2 are plotted in Fig. 10. It shows that, these two indicators gradually decrease as assimilation step increases and finally reach a low value, suggesting that the estimations of

Table 3 Flow and transport parameters used in Case 2

Flow simulation	Transient state
Total simulation time (days)	100
Stress period	1
Time steps	200
Grid spacing (m)	5 × 5 × 5
Model length (m)	105
Model width (m)	65
Model height (m)	5
Starting head (m)	50
Porosity	0.3
Specific storage (m^{-1})	0.0001
Longitudinal dispersivity (m)	10
horizontal transverse dispersivity (m)	1
Injection rate per well (m^3/day)	80
Pumping rate per well (m^3/day)	120
TCE injection concentration per well (mg/L)	50

Y_1 and Y_2 in this case are accurate and effective. Meanwhile, the difference between the *RMSE* and the *Ensemble Spread* is small, indicating that the SEOD method estimates the uncertainty properly.

To evaluate the performance of data match and model prediction, the initial and final ensembles of Y_1 and Y_2 are taken into the synthetic model respectively to calculate the head and concentration data, which are then compared with the real observations. It should be noted that only three wells (black circles in Fig. 7) are selected randomly from the study domain to show the evaluation results due to the space limitation. As shown in Figs. 11 and 12, the data calculated by the final ensemble are very close to the real observations, performing much better than those calculated by the initial ensemble. It indicates that the estimated fields of Y_1 and Y_2 are both of low uncertainty after assimilating the most informative measurements from the optimal sampling locations.

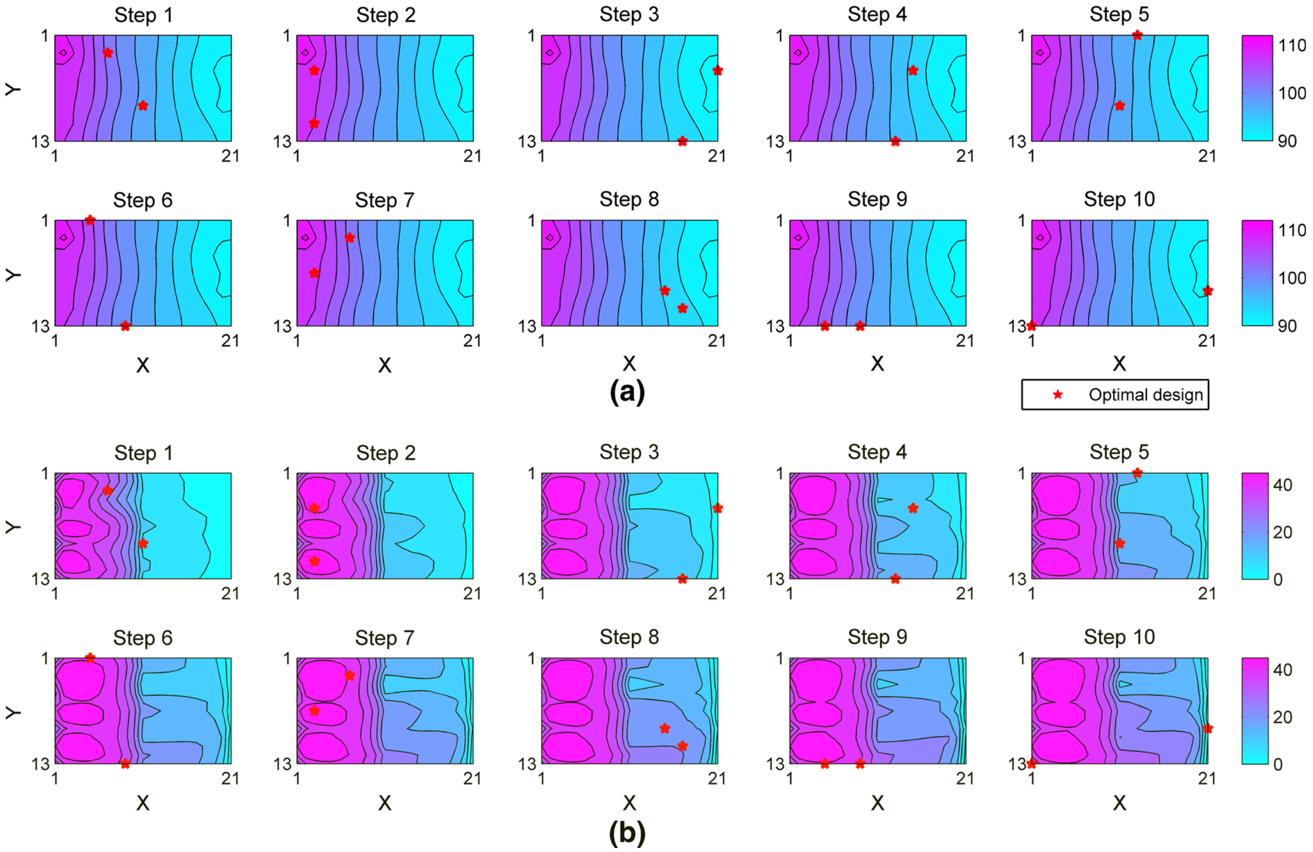


Fig. 8 The optimal sampling locations (red stars) at every assimilation step of Case 2. (**a**) and (**b**) are the contour maps of the flow field and the concentration field, respectively

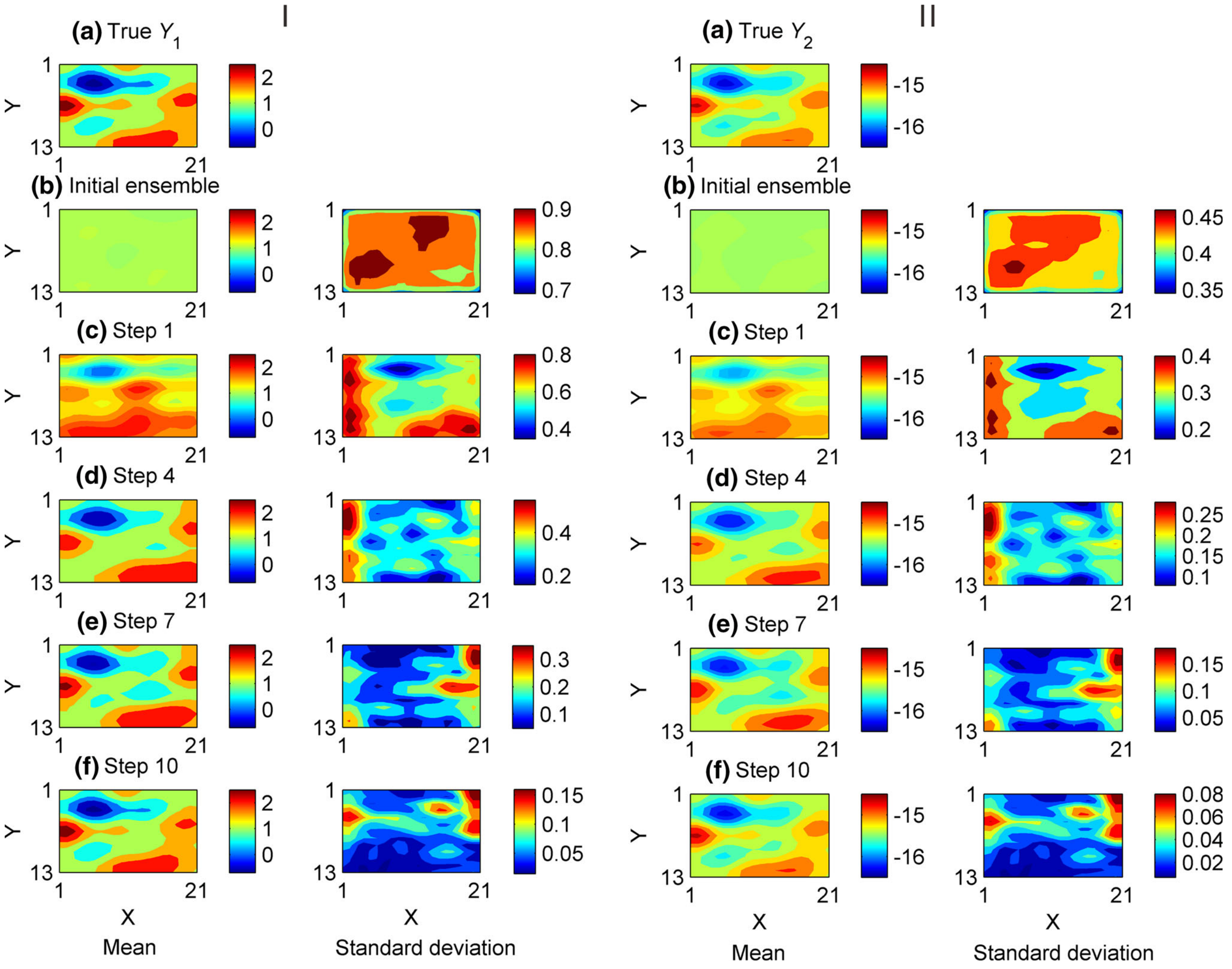
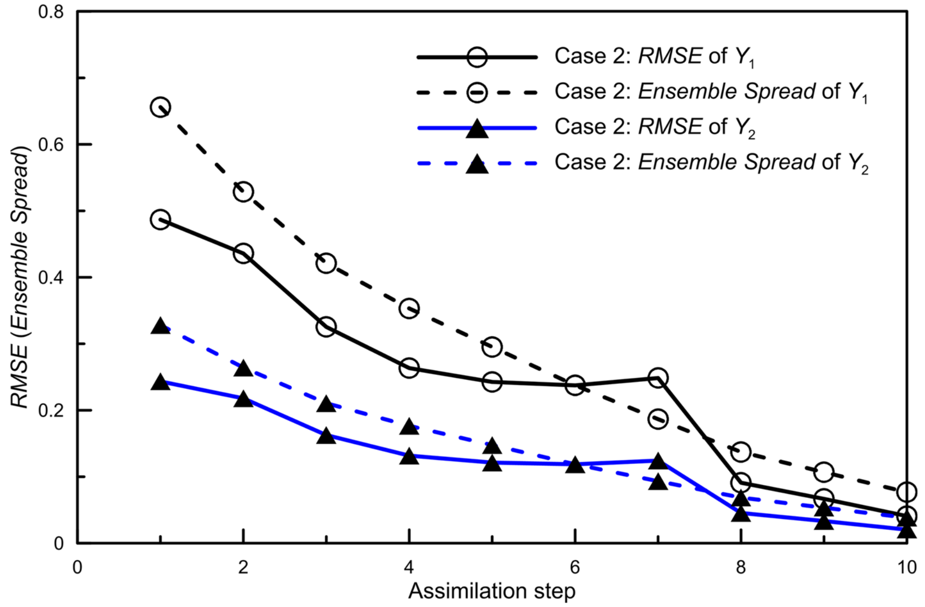


Fig. 9 The ensemble mean and the standard deviation of field Y_1 and Y_2 in Case 2. **I** for Field Y_1, **II** for Field Y_2

Fig. 10 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for field Y_1 and Y_2 in Case 2



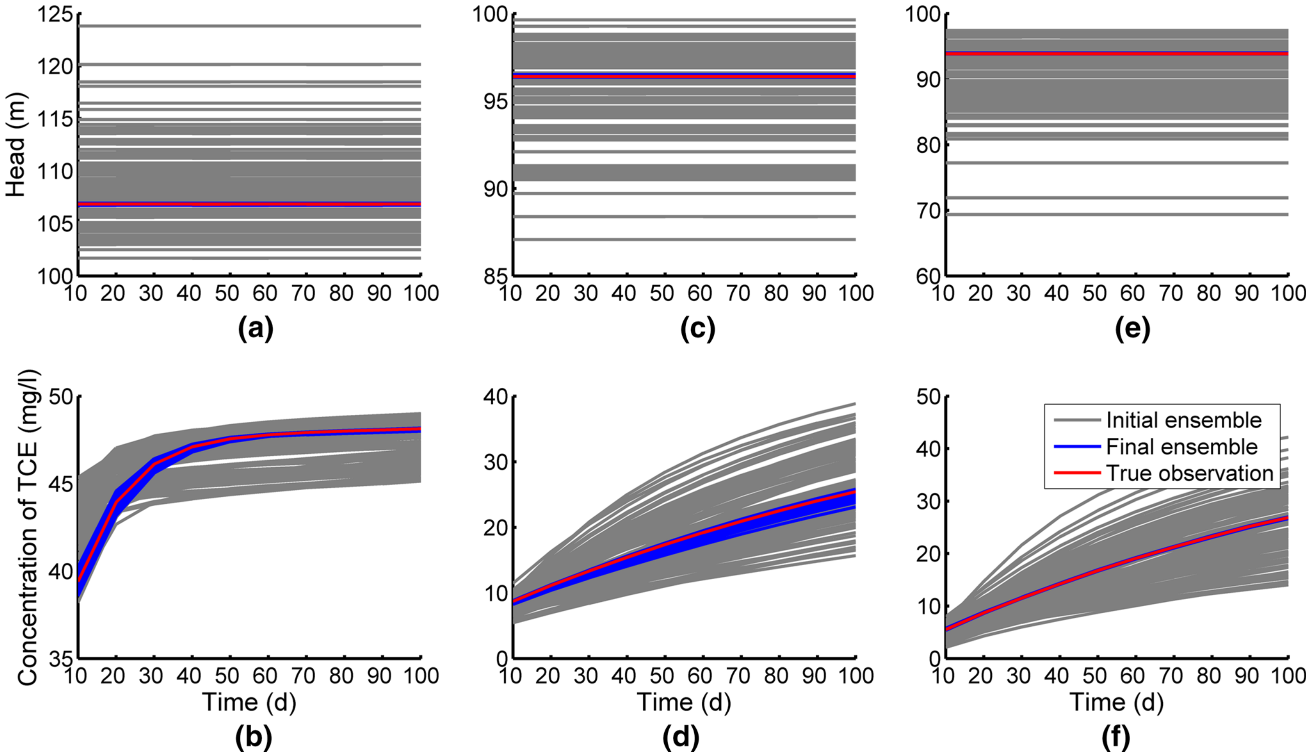


Fig. 11 The performance of data match. (**a**), (**c**), (**e**) show the data match of the head of three selected wells respectively, while (**b**), (**d**), (**f**) show the data match of the TCE concentration data of three selected wells respectively

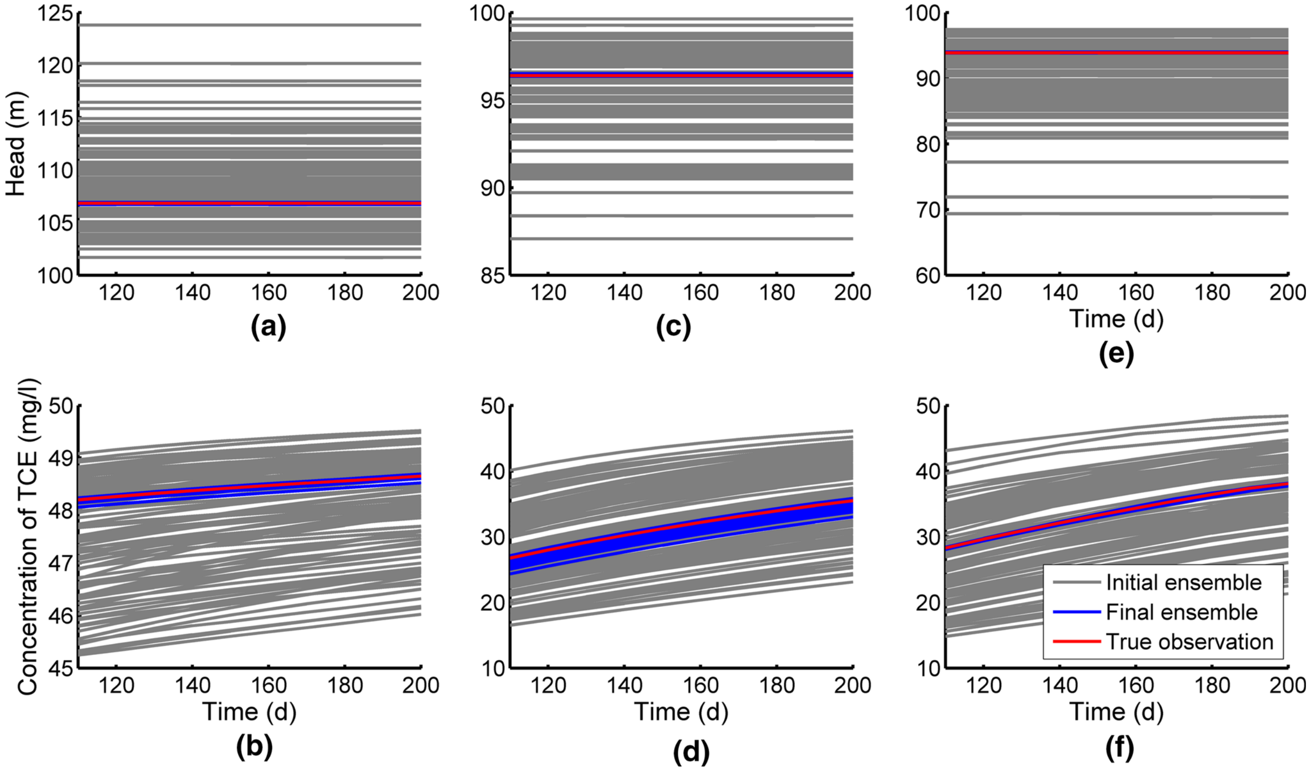


Fig. 12 The performance of model prediction. (**a**), (**b**), (**c**), (**e**) show the prediction of the head of three selected wells respectively, while (**b**), (**d**), (**f**) show the prediction of the TCE concentration data of three selected wells respectively

4 Discussion

4.1 Comparison of sequential optimization strategy and conventional strategy

In order to illustrate and demonstrate the effectiveness and efficiency of the SEOD method for jointly estimating physical and geochemical parameters, the sequential optimization strategy is compared with the conventional strategy in this subsection. For the convenience of comparison, several synthetic cases (Case 11, 12 13) are constructed based on Case 2 by replacing the sequential optimization strategy with the conventional strategy (different fixed sampling locations numbers for different cases). Except this, the other model parameters of cases constructed here are the same as those of Case 2. More details are given in Table 2.

Figure 13 shows the *RMSE* and the *Ensemble Spread* for different cases. It illustrates that the sequential optimization strategy obtains better performance of the parameter estimation when the number of sampling locations is the same. Even more, the sequential optimization strategy with 2 optimal sampling locations (Case 2) performs better than the conventional strategy with 10 fixed sampling locations (Case 12). Besides, the conventional strategy with a large number of fixed sampling locations could result in the *Ensemble Spread* becoming very small at the first few assimilation steps, which could prevent assimilating further measurements.

4.2 Effect of ensemble size

All results shown so far of Case 2 are based on an ensemble of 100 realizations. To evaluate the impact of the ensemble size on the parameter estimation, an analysis with an ensemble of 50, 300, 500, 1000 realizations (Table 2) is performed here.

The *RMSE* and the *Ensemble Spread* of different cases are shown in Fig. 14 below. It shows that an appropriate ensemble size is important for the parameter estimation. If the ensemble size is too small (Case 3), ensemble collapse, a phenomenon in which the *Ensemble Spread* is artificially small relative to its *RMSE*, could happen. If the ensemble size is too large (Case 5, 6), it could lead to more computational burden and introduce more observation errors into the model as the SEOD method is based on the Monte Carlo method. It shows that the *RMSE* and the *Ensemble Spread* of Y_1 and Y_2 are small and close to each other when the ensemble size is 100, suggesting that the estimations of Y_1 and Y_2 are accurate and the model uncertainty is estimated properly. Accordingly, the ensemble size is set to 100 in the cases discussed below.

4.3 Effect of the number of optimal sampling locations

Optimizing too many sampling locations could bring a heavy computational burden. Here, to explore the impact of the number of optimal sampling locations on the parameter estimation, several synthetic cases with different numbers of optimal sampling locations are constructed. More details are given in Table 2.

As shown in Fig. 15 below, the *RMSE* is no longer sensitive to the number of optimal sampling locations when the number of optimal sampling locations is large enough, suggesting that there could be a threshold value of the number of optimal sampling locations in this synthetic model. On the one hand, if the number of optimal sampling locations is too large, the *Ensemble Spread* becomes extremely small at the first few assimilation steps, which could prevent assimilating further measurements into the model. On the other hand, too many optimal sampling locations could lead to high economic cost and heavy computational burden. Therefore, 2–5 optimal sampling locations are enough and appropriate in this model.

4.4 Improvement of the SEOD method

In the original SEOD method, since the EnKF is a sequential history matching method, the optimization algorithm part (GA) needs to be invoked N_s times to obtain the optimal sampling design at each assimilation step, which is time-consuming. To enhance its computational efficiency, we improve the original SEOD method by replacing the EnKF with the ES-MDA. The loop of the improved SEOD method is shown in Fig. 16.

Figure 16 shows that the loop of the improved SEOD method is divided into outer loop and inner loop. The outer loop is similar to the original SEOD method, which consists of a forecast step, an optimal design step and an analysis step. The inner loop of the improved SEOD method is part of the ES-MDA. Unlike the EnKF, the ES-MDA performs N_a times global update so as to assimilate the same data (all available data) multiple times without restarting the forward model, which helps enhance the computational efficiency. In the improved SEOD method, we divide all N_s assimilation steps in the original SEOD method into N_g groups, the following loop is performed for each group in chronological order (from 1 to N_g). Note that N_a of all cases in this subsection (Case 14, 15, 16, 17) is set to 4 (Emerick and Reynolds 2013). More details are given in Table 2.

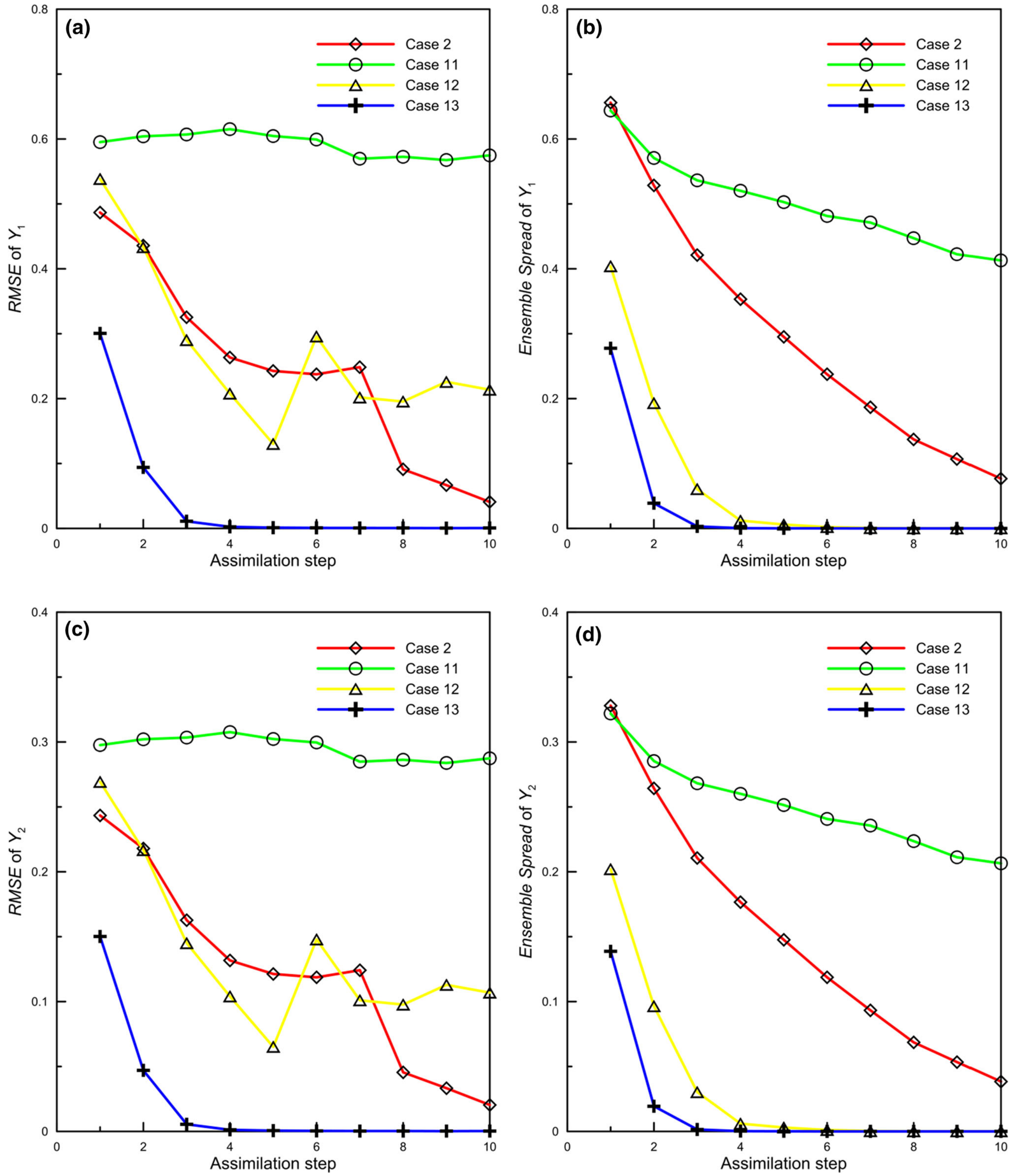


Fig. 13 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different sampling strategies

1. Forecast step

Run the forward model G from beginning time step of the group $j + 1$ to the end time step of the group $j + 1$ with updated parameters from the group j [Eq. (10)].

$$\mathrm{x}^f_{i,j+1} = G\left(\mathrm{x}^a_{i,j}\right), \quad i = 1, 2, \ldots, N_e \tag{10}$$

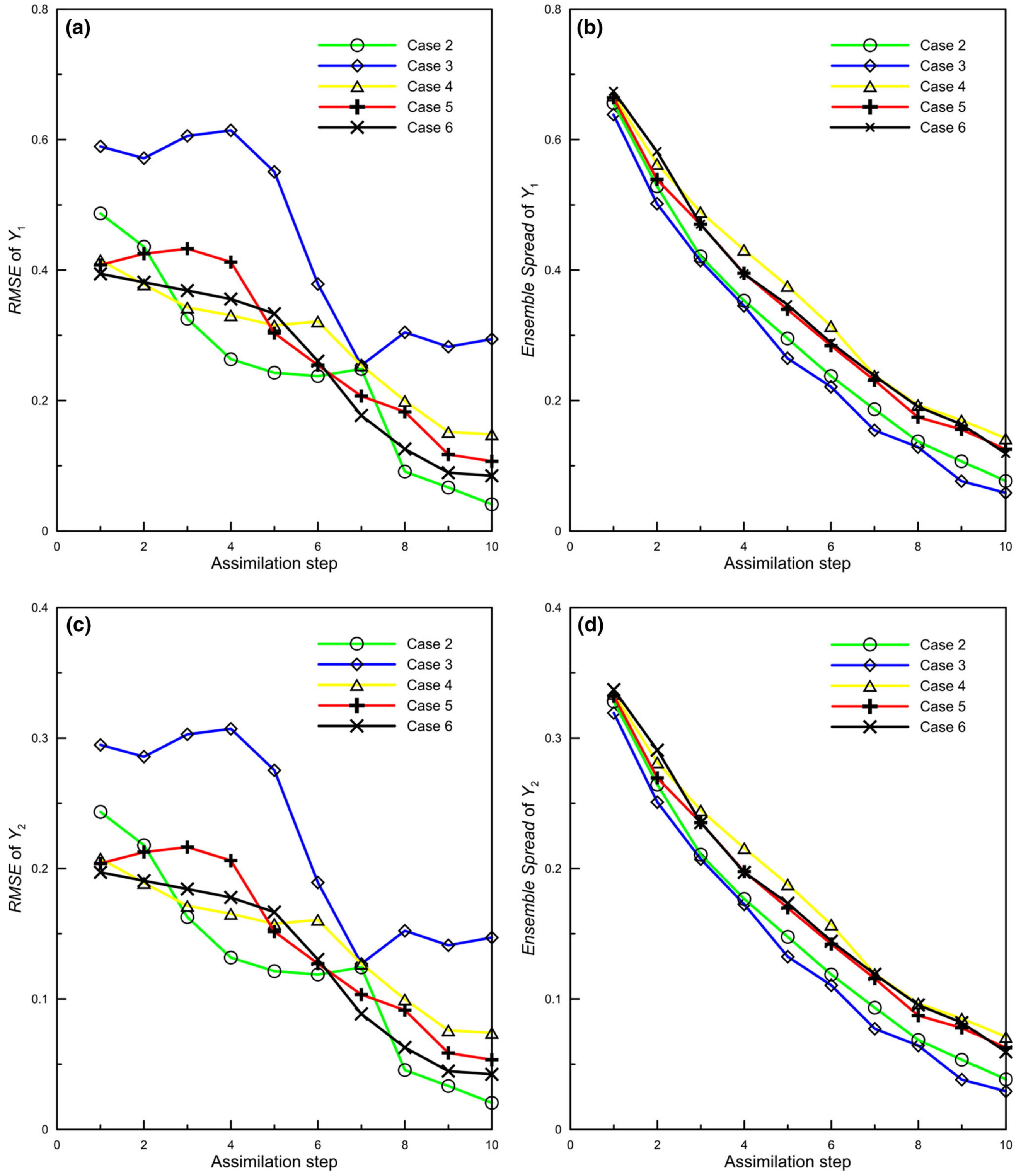


Fig. 14 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different ensemble sizes

In the above equation, i is the ensemble member index, j is the group index (from 1 to N_g), superscripts f and a denote forecast and analysis, respectively.

II. Optimal design

This step is similar with the original SEOD method, using the GA to solve an optimization problem to obtain the most informative measurements from the optimal sampling design.

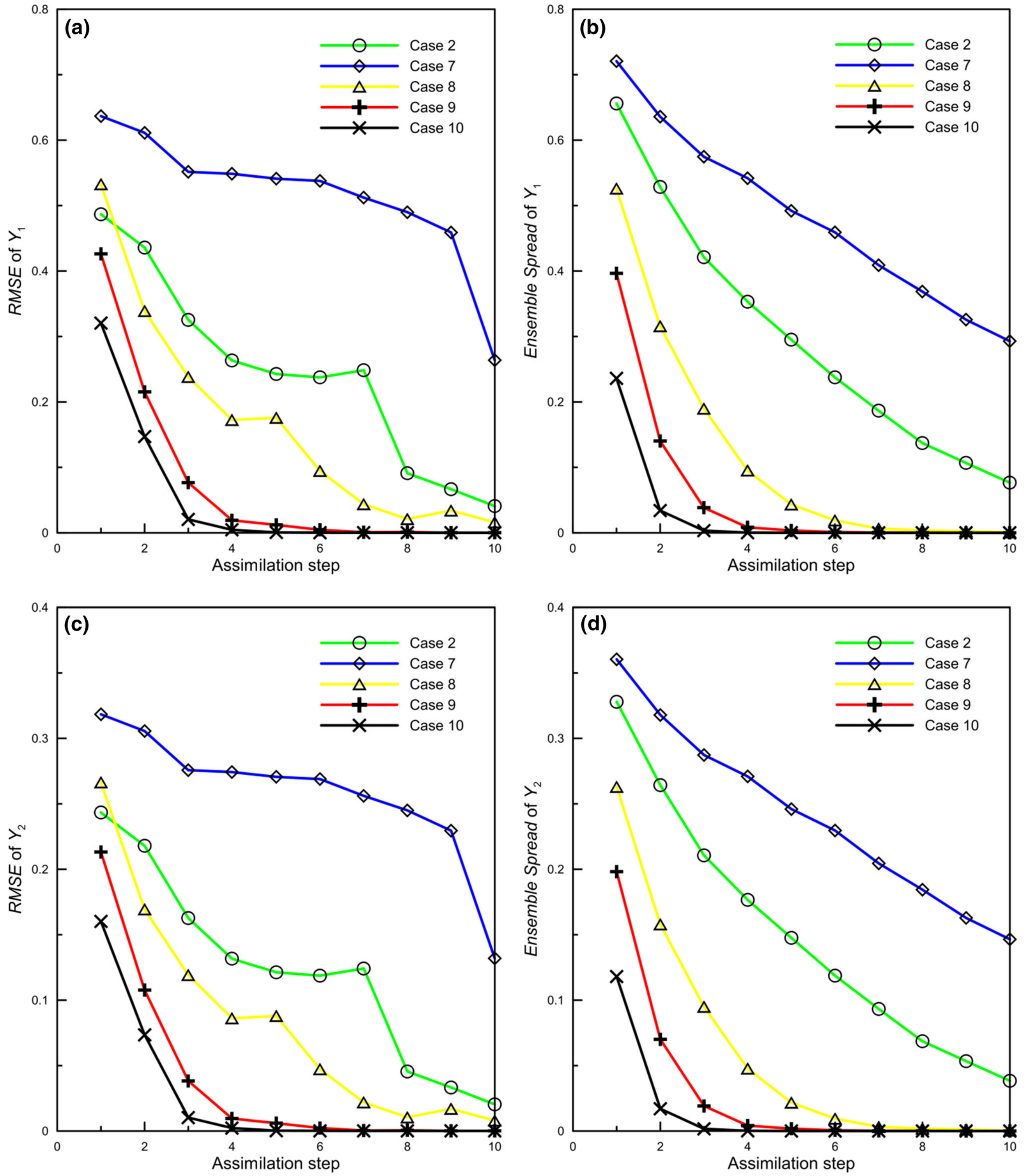


Fig. 15 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different numbers of optimal sampling locations

III. Analysis step

The following update equation [Eq. (11)] of the ES-MDA is different from that of the EnKF.

$$\mathbf{x}^{a}_{i,j+1} = \mathbf{x}^{f}_{i,j+1} + \mathbf{C}_{\mathrm{YD}}(\mathbf{C}_{\mathrm{DD}} + \alpha_l \mathbf{C}_{\mathrm{D}})^{-1}(\mathbf{d}_{obs_{i,j+1}} - \mathbf{d}_{i,j+1}),\quad i = 1, 2, \ldots, N_e \tag{11}$$

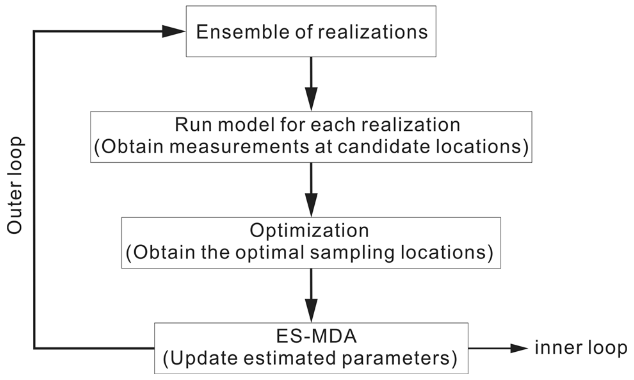

Ensemble of realizations
Run model for each realization
(Obtain measurements at candidate locations)
Optimization
(Obtain the optimal sampling locations)
ES-MDA
(Update estimated parameters)
Outer loop
inner loop

Fig. 16 The loop diagram of the improved SEOD method

In the above equation, l is the times index of the ES-MDA, $l = 1, 2, \ldots, N_a$; $\mathbf{d}_{obs}$ is the perturbed observations with noise of covariance $\alpha_l \mathbf{C}_D$ ($\alpha_1 = 9.333$, $\alpha_2 = 7.0$, $\alpha_3 = 4.0$ and $\alpha_4 = 2.0$, Emerick and Reynolds 2013). Other letters in this equation have the same meaning as those in Eq. (6).

After N_a times global update, the updated ensemble X^a of the group $j + 1$ is obtained here. Then, go back to step (I), the updated ensemble is implemented for the next group. Through this improvement, the times of invoking the GA decrease from N_s to N_g, which helps enhance the computational efficiency.

To compare the improved SEOD method with the original one and discuss the influence factors of the improved one, several cases are constructed with all model parameters the same as those in Case 2. More details can be found in Table 2. The results of these cases are shown in Fig. 17.

Figure 17 shows that the number of optimal sampling locations dominantly affects the results of the improved SEOD method. In Case 2, two optimal sampling locations are chosen at each assimilation step. During the whole data assimilation, total 20 sampling locations are used to obtain measurements at most (2 × 10, if the optimal sampling locations are different at each step). Therefore, if the number of optimal sampling locations in the improved SEOD method is too small, the improved SEOD method can’t estimate the parameters of the entire study domain well just through a few sampling locations (e.g., Case 14 with total 4 optimal sampling locations at most). When the number of sampling locations is enough, the result of the improved SEOD method is acceptable (e.g., Case 15 with total 10 optimal sampling locations at most). The result of Case 15 is comparable with the result of Case 2, but the computer cost of Case 15 is much less than that of Case 2 because that Case 15 just invokes the GA only twice while Case 2 invokes the GA 10 times. Therefore, if the number of optimal sampling locations is not set to a very small value, the computational efficiency will be enhanced by using the improved SEOD method (Table 4).

Furthermore, the way of dividing the observation time (assimilation steps in the original SEOD method) into several groups (called the group division strategy in the following context) affects the results of the improved method as well. From the comparison of Cases 15, 16 and 17, it is obvious that Case 16, whose observation time in each divided groups is progressively increasing, has a better data assimilation result than the other two cases, whose observation time in each divided groups is equivalent or progressively decreasing. It is an interesting phenomenon, which is worth further research. For now, we think it is probably because more and more precise measurements in the early stage of the data assimilation would lead to excessive update of estimated parameters (Burgers et al. 1998; Evensen 2009). Therefore, future studies should focus on optimizing the group division strategy to obtain a more accurate estimation of model parameters.

5 Conclusions

In this study, we make use of a sequential ensemble-based optimal design (SEOD) method to jointly estimate physical and geochemical parameters of groundwater models.

Both physical and geochemical parameters are estimated accurately in the one-dimensional and two-dimensional synthetic cases by using the SEOD method. Uncertainties of both physical and geochemical parameters decrease after assimilating the most informative measurements at the optimal sampling locations, and the accuracy of model prediction increases meanwhile. Furthermore, several comparison cases are tested and analyzed, results illustrate and demonstrate the effectiveness and efficiency of the SEOD method on jointly estimating high-dimensional physical and geochemical parameters in groundwater models.

The ensemble size and the number of optimal sampling locations have impacts on the parameter estimation based on the SEOD method. A too small ensemble size would lead to the ensemble collapse. Furthermore, when the number of optimal sampling locations is too large, heavier computational burden and more observation errors would be caused, and the *RMSE* is no longer sensitive to the number of optimal sampling locations. How to determine the optimal ensemble size and sampling locations number for different scenarios is worth further investigation.

The original SEOD method has a heavy computational burden because it invokes the GA too many times. To enhance its computational efficiency, we proposed an improved SEOD method in this study by replacing the EnKF with the ES-MDA. The results of comparison cases

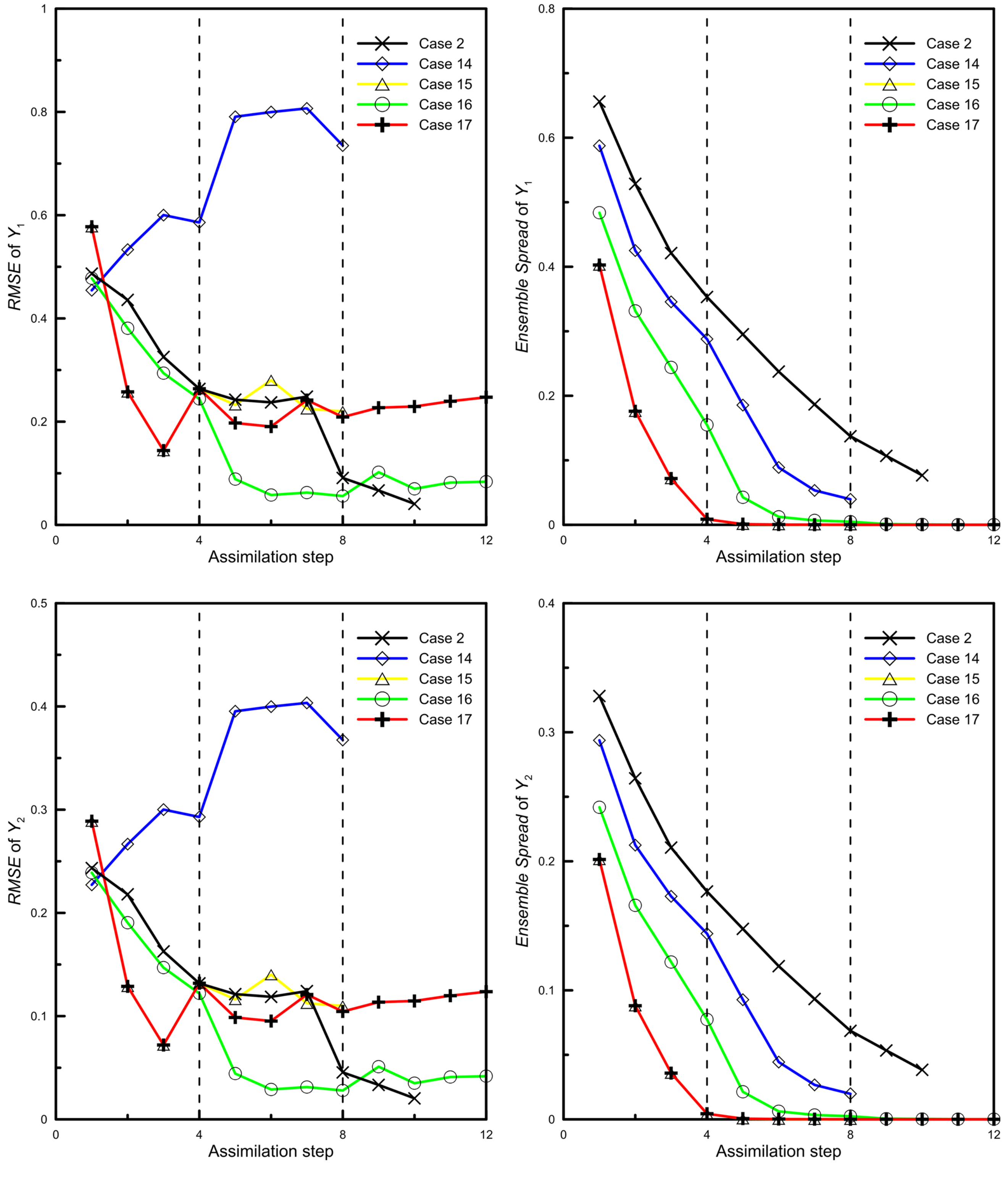


Fig. 17 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different cases using the improved SEOD method

show that the improved SEOD method is advantageous than the original one, which makes the SEOD framework more promising for the parameter estimation and the optimal sampling strategy design. The number of optimal sampling locations and the strategy of dividing groups would affect the results of the improved SEOD method.

Table 4 Computational costs

Case name	Times of model invoking	Times of GA invoking	computing time (using the same computer)
Case 2	10	10	2 h
Case 14	8	2	26 min
Case 15	8	2	26 min
Case 16	12	3	25 min
Case 17	12	3	25 min

It is noted that only two kinds of measurements (head and concentration) are assimilated in this work. More kinds of measurements (e.g., hydraulic conductivity, porosity, temperatures and hydrogeophysical data) can be assimilated simultaneously so as to make use of more hard and soft data to improve the accuracy of parameter estimation in further study.

Acknowledgements The authors would like to thank the anonymous referees for their insightful comments and suggestions that have helped improve the paper. This work was financially supported by the National Nature Science Foundation of China grants (Nos. U1503282, 41672229 and 41172206). We would like to thank Mr. Jun Man from Zhejiang University for providing the SEOD code.

References

Aanonsen SI, Nævdal G, Oliver DS, Reynolds AC, Vallès B (2009) The ensemble Kalman filter in reservoir engineering—a review. SPE J 14:393–412

Atchley AL, Navarre-Sitchler AK, Maxwell RM (2014) The effects of physical and geochemical heterogeneities on hydro-geochemical transport and effective reaction rates. J Contam Hydrol 165:53–64. https://doi.org/10.1016/j.jconhyd.2014.07.008

Bear J (1972) Dynamics of fluids in porous materials. Dover, New York

Burgers G, Leeuwen P, Evensen G (1998) Analysis scheme in the ensemble Kalman filter. Month Weather Rev 126:1719–1724

Carrera J, Neuman SP (1986) Estimation of aquifer parameters under transient and steady state conditions: 3. Application to synthetic and field data. Water Resour Res 22:228–242

Carrera J, Alcolea A, Medina A, Hidalgo J, Slooten LJ (2005) Inverse problem in hydrogeology. Hydrogeol J 13(1):206–222

Chen Y, Oliver DS (2010) Cross-covariances and localization for EnKF in multiphase flow data assimilation. Comput Geosci 14:579–601

Chen Y, Zhang D (2006) Data assimilation for transient flow in geologic formations via ensemble Kalman filter. Adv Water Resour 29:1107–1122. https://doi.org/10.1016/j.advwatres.2005.09.007

Cleveland TG, Yeh WWG (1990) Sampling network design for transport parameter identification. J Water Resour Plan Manag 116:764–783

Dagan G (1984) Solute transport in heterogeneous porous formations. J Fluid Mech 145:151. https://doi.org/10.1017/s0022112084002858

Dagan G (1985) Stochasti modeling of groundwater flow by unconditional and conditional probabilities: the inverse problem. Water Resour Res 21(1):65–72

Doherty J (2004) PEST: model-independent parameter estimation, user's manual, 5th edn. Watermark Numerical Computing, Oxley

Emerick AA, Reynolds AC (2011) Combining sensitivities and prior information for covariance localization in the ensemble Kalman filter for petroleum reservoir applications. Comput Geosci 15:251–269

Emerick AA, Reynolds AC (2013) Ensemble smoother with multiple data assimilation. Comput Geosci Uk 55:3–15. https://doi.org/10.1016/j.cageo.2012.03.011

Evensen G (2003) The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn 53:343–367

Evensen G (2009) Data assimilation: the ensemble Kalman filter. Springer, Berlin

Fennell DE, Carroll AB, Gossett JM, Zinder SH (2001) Assessment of indigenous reductive dechlorinating potential at a TCE-contaminated site using microcosms, polymerase chain reaction analysis, and site data. Environ Sci Technol 35:1830–1839

Gómez-Hernández JJ, Hendricks Franssen HJ, Sahuquillo A (2003) Stochastic conditional inverse modeling of subsurface mass transport: a brief review and the self-calibrating method. Stoch Environ Res Risk Assess 17(5):319–328

Gu Y, Oliver DS (2007) An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. SPE J 12:1990–1995

Harbaugh AW, Banta ER, Hill MC, McDonald MG (2000) MODFLOW-2000, The US geological survey modular ground-water model—User guide to modularization concepts and the groundwater flow process. US Geological Survey Open-File Report 00–92, 121 p

Hendricks Franssen HJ, Kinzelbach W (2008) Real-time groundwater flow modeling with the ensemble Kalman filter: joint estimation of states and parameters and the filter inbreeding problem. Water Resour Res 44:354–358

Hendricks Franssen HJ, Alcolea A, Riva M, Bakr M, van der Wiel N, Stauffer F, Guadagnini A (2009) A comparison of seven methods for the inverse modelling of groundwater flow. Application to the characterisation of well catchments. Adv Water Resour 32(6):851–872

Huang C, Hu BX, Li X, Ye M (2009) Using data assimilation method to calibrate a heterogeneous conductivity field and improve solute transport prediction with an unknown contamination source. Stoch Env Res Risk Assess 23(8):1155

Kalman RE (1960) A new approach to linear filtering and prediction problems. Trans ASME J Basic Eng 82(D):35–45

Knopman DS, Voss CI (1987) Behavior of sensitivities in the one-dimensional advection-dispersion equation: implications for parameter estimation and sampling design. Water Resour Res 23:253–272

Kullback S (1997) Information theory and statistics. Courier Corporation, North Chelmsford

Li G, Reynolds AC (2009) Iterative ensemble Kalman filters for data assimilation. SPE J 14:496–505

Li L, Steefel CI, Kowalsky MB, Englert A, Hubbard SS (2010) Effects of physical and geochemical heterogeneities on mineral transformation and biomass accumulation during biostimulation experiments at Rifle, Colorado. J Contam Hydrol 112:45–63. https://doi.org/10.1016/j.jconhyd.2009.10.006

Man J, Zhang J, Li W, Zeng L, Wu L (2016) Sequential ensemble-based optimal design for parameter estimation. Water Resour Res 52:7577–7592. https://doi.org/10.1002/2016wr018736

Moradkhani H, Sorooshian S, Gupta HV, Houser PR (2005) Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Adv Water Resour 28:135–147

Neuman SP (1973) Calibration of distributed parameter groundwater flow models viewed as a multiple objective decision process under uncertainty. Water Resour Res 9(4):1006–1021

Nowak W, De Barros FPJ, Rubin Y (2010) Bayesian geostatistical design: task-driven optimal site investigation when the geostatistical model is uncertain. Water Resour Res 46(3):374–381. https://doi.org/10.1029/2009WR008312

Oliver DS, Chen Y (2011) Recent progress on reservoir history matching: a review. Comput Geosci 15(1):185–221

Oliver DS, Cunha LB, Reynolds AC (1997) Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. Math Geol 29(1):61–91

Prommer CH, Post V (2010) A reactive multicomponent transport model for saturated porous media. Groundwater 48(5):627–632

Rubin Y (1991) Transport in heterogeneous porous media: prediction and uncertainty. Water Resour Res 27:1723–1738

Sandrin SK, Brusseau ML, Piatt JJ, Bodour AA, Blanford WJ, Nelson NT (2004) Spatial variability of in situ microbial activity: biotracer tests. Groundwater 42:374–383

Scheibe TD, Fang Y, Murray CJ, Roden EE, Chen J, Chien YJ, Brooks SC, Hubbard SS (2006) Transport and biogeochemical reaction of metals in a physically and chemically heterogeneous aquifer. Geosphere 2(4):220–235. https://doi.org/10.1130/Ges00029.1

Sorensen JVT, Madsen H, Madsen H (2004) Data assimilation in hydrodynamic modelling: on the treatment of non-linearity and bias. Stoch Environ Res Risk Assess 18(7):228–244

Sudicky EA (1986) A natural gradient experiment on solute transport in a sand aquifer: spatial variability of hydraulic conductivity and its role in the dispersion process. Water Resour Res 22:2069–2082. https://doi.org/10.1029/WR022i013p02069

Sun NZ, Yeh WWG (2007) Development of objective-oriented groundwater models: 2. Robust experimental design. Water Resour Res. https://doi.org/10.1029/2006wr004888

Tong J, Hu BX, Yang J (2010) Using data assimilation method to calibrate a heterogeneous conductivity field conditioning on transient flow test data. Stoch Environ Res Risk Assess 24(8):1211–1223

Ushijima TT, Yeh WWG (2015) Experimental design for estimating unknown hydraulic conductivity in an aquifer using a genetic algorithm and reduced order model. Adv Water Resour 86:193–208

Van Leeuwen PJ, Evensen G (1996) Data assimilation and inverse methods in terms of a probabilistic formulation. Mon Weather Rev 124:2898–2913

Whitley D (1994) A genetic algorithm tutorial. Stat Comput 4(2):65–85

Zhang D, Lu Z (2004) An efficient, high-order perturbation approach for flow in random porous media via Karhunen–Loève and polynomial expansions. J Comput Phys 194:773–794

Zhang J, Zeng L, Chen C, Chen D, Wu L (2015) Efficient Bayesian experimental design for contaminant source identification. Water Resour Res 51(1):576–598

Zheng C (2006) MT3DMS v5.2 supplemental user's guide: technical report to the US Army Engineer Research and Development Center, Department of Geological Sciences, University of Alabama, p 24

Zhou HY, Gómez-Hernández JJ, Li LP (2014) Inverse methods in hydrogeology: evolution and recent trends. Adv Water Resour 63:22–37