Stochastic Environmental Research and Risk Assessment

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

--Manuscript Draft--

Manuscript Number:	SERR-D-17-00190R1
Full Title:	Joint inversion of physical and geochemical parameters in groundwater models by sequential ensemble-based optimal design
Article Type:	Original research
Keywords:	Optimal sampling strategy; Physical and geochemical heterogeneity; Parameter estimation; Reactive transport model; Data assimilation.
Corresponding Author:	Xiaoqing Shi Nanjing University Nanjing, Jiangsu CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Nanjing University
Corresponding Author's Secondary Institution:	
First Author:	Tian Lan
First Author Secondary Information:	
Order of Authors:	Tian Lan
	Xiaoqing Shi
	Beilei Jiang, Ph.D.
	Yuanyuan Sun, Ph.D.
	Jichun Wu, Ph.D.
Order of Authors Secondary Information:	
Funding Information:	National Nature Science Foundation of China grant (U1503282) — Professor Jichun Wu
	National Nature Science Foundation of China grant (41672229) — Dr. Xiaoqing Shi
	National Nature Science Foundation of China grant (41172206) — Dr. Xiaoqing Shi
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 (GA) 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 (ES-MDA). 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.
Response to Reviewers:	More details please see the attached Response Letter.

Dear Editor,

We greatly appreciate your concern in our research paper. Your suggestions and both reviewers' comments contributed to its improvement. The reviewers' comments are all considered and the answers are given in the response letter. All the corrections are shown in the annotated copy of our manuscript.

Thank you for your consideration of this manuscript.

Sincerely,

Xiaoqing Shi

Reply to review comments on "Joint inversion of physical and geochemical parameters in groundwater models by sequential ensemble-based optimal design" (SERR-D-17-00190)

(Original review comments in black; Reply from the authors in red)

Comments from Chief Editor George Christakos:

Based on the advice received, I feel that your manuscript could be reconsidered for publication should you be prepared to incorporate major revisions. When preparing your revised manuscript, you are asked to carefully consider the reviewer comments which can be found below, and submit a list of responses to the comments. You are kindly requested to also check the website for reviewer attachment.

Response: We thank the editor for giving us the opportunity to address the AE's and reviewers' comments. We have carefully made the revisions corresponding to all the comments and provided a detailed point-by-point response as given below.

Comments from Associate Editor:

1- I received comments from two reviewers, reviewer 1 suggested revision and reviewer 2 suggested major revision. In general, reviewer 1 was positive to the manuscript, but asked the authors to provide more clear description of the SEOD and methodology used in this study. Reviewer 2, however, concerned about the novelty of this manuscript. The reviewer pointed out that the method used in the manuscript, SEOD method, was a known model, the authors need to highlight the innovation in the paper.

After receiving the comments, I carefully read the manuscript again and concur with the reviewers in general. I suggest major revision. The authors need to clearly highlight the manuscript's novelty and contribution. They need to reply the comments and question in detail and make revision to their manuscript accordingly.

Response: We would like to thank the AE for the constructive comments and for keeping the door for addressing reviewers' comments. Although the sequential ensemble-based optimal design (SEOD) method is a known method (developed by Man et al., 2016), the novelties and contributions of this study were highlighted in three aspects:

(1) To our knowledge, the SEOD method is firstly applied for groundwater reactive transport modeling in this study. Major challenge for subsurface reactive transport models is the correct description of the subsurface processes using a limited amount of measurements. For this reason, to avoid complexity not justified by the available data, the spatial distributions of physical and geochemical parameters are often simplified. Man et al. (2016) developed the SEOD method, and demonstrated the effectiveness of this method by estimating only physical parameters in unsaturated flow models. In this study, we applied the SEOD method into jointly estimating physical and geochemical parameters of reactive transport model.

(2) To distinguish this study from a mere case study, the worth of multiple observation data (piezometric head and concentration) was analyzed in the context of parameter estimation and prediction uncertainty reduction. The beneficial impact of assimilating dynamic multiple observation data in a transient flow and transport model is patent for the EnKF framework. Few studies have attempted to assimilate head and concentration to accurately estimate heterogeneous distribution of both physical and geochemical parameters. It is noted that the SEOD method was assimilated only head data in groundwater flow model as demonstrated in Man et al. (2016).

(3) Additionally, we improved the SEOD method using the Ensemble Smoother with multiple data assimilation (ES-MDA) to replace the EnKF to enhance its computational efficiency. The comparison was illustrated in the section 4 of the revised manuscript.

According to the reviewer's constructive comments, we rewrote the Introduction, Abstract and Conclusion to highlight our contributions. To illustrate how we improved the SEOD framework using the ES-MDA, we also added details about the procedure of the SEOD method in section 2 of Methodologies.

2- In addition, the authors must link directly the paper's analysis to papers that have appeared in SERRA. This is because there is a question of relevancy, i.e., whether SERRA readers will benefit maximally by a paper that does not show strong evidence of relevancy to the journal's contents.

Response: To benefit maximally for SERRA readers, we added more literature reviews on widespread use of EnKF in parameter estimation and state prediction published in SERRA in section of "Introduction", for example:
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).

Comments from Reviewer #1:

This paper proposes to implement SEOD method into the groundwater transport modeling to improve the sampling strategy and parameter estimation process. The authors tested the SEOD method in two synthetic groundwater modeling cases and showed that it performs better than the traditional data assimilation method.

General comments:

1- I believe the manuscript is well written. I do like the objectives, and simple logic flow of this paper. The techniques can be helpful for other groundwater modelers.

Response: Thank you for the positive comment.

2- However, some points of this paper need to be clarified and more discussions are needed, as listed below. I would be in favor of publication after the authors addressed the comments given below.

Response: We have revised the manuscript carefully according to your comments and provided a detailed point-by-point response as given below.

Specific comments:

1. Since SEOD method is essential for this manuscript, I believe more details of this algorithm should be provided in the methodology section instead of just mentioning it (Page 8).

Response: Accepted and clarified. The SEOD method used in this study was adopted from previous work published in a recently WRR paper (Man et al., 2016). The details could be found in the corresponding cited literature and hence were only briefly discussed in the original manuscript. According to your comments, especially to help readers to understand how we improve the SEOD framework using the ES-MDA, the details about the SEOD method were added in section 2 Methodologies in the revised manuscript.

2. The methodology used in this research was not described clearly enough, please provide more details about the SEOD, EnKF, and relative entropy (RE).

Response: Accepted. More details about the SEOD method, the EnKF and the definition of RE were added in the revised manuscript (see section 2.2).

3. The novelty of this work is not well presented, since the methodology of this research is not new, please add more descriptions about what the new knowledge we can learn from this work (especially in the discussion and conclusion sections).

Response: Accepted and clarified. The Introduction, Discussion and Conclusion were modified to highlight the primary objective and key findings. This point has been addressed in our response to the AE's first comment (above). Although the SEOD method was a known method (developed by Man et al., 2016), the novelties and contributions of this study were highlighted in three aspects: (1) To our knowledge, the SEOD method is first applied for groundwater reactive transport modeling. In this study, we applied the SEOD method into jointly estimating physical and geochemical parameters of reactive transport model. (2) To distinguish this study from a mere case study, the worth of multiple observation data (piezometric head and concentration) was analyzed in the context of parameter estimation and prediction uncertainty reduction. The beneficial impact of assimilating dynamic multiple observation data in a transient flow and transport model is patent for the EnKF framework. Few studies have attempted to assimilate head and concentration to accurately estimate heterogeneous distribution of both physical and geochemical parameters. (3) Additionally, we improved the SEOD method using the ES-MDA to replace the EnKF to enhance its computational efficiency. The comparison was illustrated in the section 4 of the revised manuscript.

4. Please explain why use synthetic cases instead of realistic ones.

Response: Clarified. As we mentioned before, this study firstly explores an effective designed sampling strategy to accurately estimate heterogeneous distribution of physical and geochemical parameters in groundwater reactive transport models. Synthetic cases are good choices to illustrate the performance of a new proposed method due to the complexity and large computational burden of realistic ones. It is easily to extend our work to real practical applications in further study.

5. Page 6, line 8-9: please provide some new references about uncertainty study.

Response: Accepted. We have added several recently references, which relate to designing optimal sampling strategy in groundwater models for improving parameter identification and minimizing prediction uncertainty. Please see the section "Introduction" in the revised manuscript.

6. Please avoid using the terms like "firstly" in the manuscript, just "first".

Response: Accepted. We have checked and corrected in the revised manuscript.

7. The figure formats have some issues, please keep them consistent and put all the legends in the same place and add color bars for each line of subplots (i.e., Fig.8).

Response: Accepted. To our best, we have checked and corrected the figure formats, i.e., added color bars for each line, kept the consistent legends and put them in the same place. The legends were usually put in the upper right corner. However, for special cases that if there are important messages in the upper right corner of plots, we would change the location of legend correspondingly to make the plots more clearly for readers.

Comments from Reviewer #2:

1- The authors applied the ensemble Kalman filter (EnKF) method in joint inversion of physical and geochemical parameters of reactive transport model. And they used the sequential ensemble-based optimal design (SEOD) method to obtain the informative measurements at every assimilation step of the EnKF method. English is very easy to read. Authors have done much interesting work,

Response: Thank you for the positive comment.

2- however, SEOD method is a known model, authors do not describe its research situation to highlight their innovation in the paper.

Response: Accepted and clarified. This point has been addressed in our response to the AE and the first reviewer's comment 3 (above). The Introduction, Discussion and Conclusion sections were modified to highlight the innovation of this study in the revised manuscript.

3- Also they just present figures and analyze them simply, and the discussion section should be rewritten with more analysis. So I recommend major revision of this paper to the SERRA journal with its present format.

Response: Accepted. The content of discussion section was indeed not abundant enough in the original manuscript. In the revised manuscript, we focused on rewriting this section.
The original SEOD method is time-consuming because it invokes the Genetic Algorithm many times to estimate the optimal sampling design. To improve its computational efficiency, we replaced the EnKF with the ES-MDA. In discussion section, we compared the performance of the original and improved SEOD method. The effect of optimal sampling number and group division strategy on the result of this improved method were also discussed. The improvement increases the value of this method and the implications of this study.

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,*]

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

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

*Corresponding author:

Xiaoqing Shi, Email: shixq@nju.edu.cn; (86) 25-89680839

Jichun Wu, Email: jcwu@nju.edu.cn; (86) 25-89680705

Highlights:

1. Both the heterogeneous physical and geochemical parameters of groundwater models are accurately estimated by using the sequential ensemble-based optimal design (SEOD) method.
2. The effectiveness and efficiency of the SEOD method are illustrated and demonstrated by the comparison between the sequential optimization strategy and the conventional strategy.

3. To enhance its computational efficiency, the SEOD method is improved by replacing the EnKF with the Ensemble Smoother with multiple data assimilation (ES-MDA).

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 (GA) 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 (ES-MDA). 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 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 Section 2, the groundwater reactive transport model and the SEOD method are described. In Section 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 Section 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 Section 4.4. Conclusions are summarized in Section 5.

2 Methodologies

2.1 Groundwater reactive transport model

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

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

where $\nabla \cdot$ is the divergence operator; ∇ is the gradient operator; 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 [LT^{-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):

$$\frac{\partial C_n}{\partial t} = \nabla \cdot \left(D \cdot \nabla C_n \right) - \nabla \cdot \left(v C_n \right) + r_{\text{reac},n} + \frac{q_s}{\theta} C_n^s \tag{2}$$

where C_n is the aqueous concentration of the nth component [ML^{-3}]; t is the time [T]; D is the diffusion coefficient [L^2T^{-1}]; $v = (-K \nabla H)/\theta$ [L^2T^{-1}]; $r_{reac,n}$ is the concentration change of the nth component caused by reactions; 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}].

Eq. (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 1977), 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:

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

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)).

$$\mathbf{x}_{i,j+1}^{f} = G(\mathbf{x}_{i,j}^{a}),\ \ i = 1,2,...,N_e \tag{4}$$

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 $\mathbf{d}_i^{'} = \mathbf{H}^{'}\mathbf{x}_i^{f} + \boldsymbol{\xi}_i$. 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 $\mathbf{H}_{\text{opt}}$ can be determined by solving the following optimization problem (Eq. (5)) with the help of the Genetic Algorithm (GA, Whitley 1994).

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

(III) Analysis step

After obtaining the optimal sampling strategy, the actual measurements $\mathbf{d}$ can be obtained and used in the analysis step (Eq. (6)).

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

In the above equation, $\mathbf{C}_{\text{YD}}$ is the cross-covariance matrix between the forecast

state and the predicted data, $\mathbf{C}_{DD}$ is the covariance matrix of the predicted data, $\mathbf{C}_D$ is the covariance matrix of the measurements error, $\mathbf{d}_{obs}$ is the perturbed observations with noise of covariance $\mathbf{C}_D$, and $\mathbf{d}$ 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:

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

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

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. 1(a), 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 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. 1(b), (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 Table 1 and 2.

[Figure 1]

[Table 1]

[Table 2]

The log saturated hydraulic conductivity $Y_1 = \ln(K)$ and the first-order rate constant $Y_2 = k_{TCE}$ are assumed to be Gaussian distributed, with mean μ_{Y_1}=1 and μ_{Y_2}=0.17 and variance σ_{Y_1}=1 and σ_{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:

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

where the horizontal correlation length λ_x = 5 m, and the vertical correlation length λ_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 field so as to obtain the most informative measurements. It is interesting to note that most optimal sampling locations are located at the front of the contaminant plume, which means that these locations can provide the most informative measurements.

[Figure 2]

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.

[Figure 3]

[Figure 4]

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. 1(a)) 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 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 Fig. 5 and 6. Overall, model uncertainty is significantly reduced after assimilating the most informative measurements from the optimal sampling locations.

[Figure 5]

[Figure 6]

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/d 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/d 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 (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 Table 2 and 3.

[Figure 7]

[Table 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 m 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 *Y2*, 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.

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 field to obtain the most informative measurements.

[Figure 8]

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.

[Figure 9]

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 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.

[Figure 10]

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 Fig. 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.

[Figure 11]

[Figure 12]

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.

Fig.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.

[Figure 13]

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.

[Figure 14]

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 to 5 optimal sampling locations are enough and appropriate in this model.

[Figure 15]

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]

Fig. 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.

(I) 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)).

$$\mathrm{x}_{i,j+1}^{f} = G(\mathrm{x}_{i,j}^{a}), \quad i = 1, 2, ..., N_e \tag{10}$$

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.

(III) Analysis step

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

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

In the above equation, l is the times index of the ES-MDA, $l = 1, 2, ..., N_a$; $\mathbf{d}_{obs}$ is the perturbed observations with noise of covariance $\alpha_l \mathbf{C}_{\mathrm{D}}$ (α_1=9.333, α_2=7.0,

α_3=4.0 and α_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]

Fig. 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.

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 Case 15, Case 16 and Case 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 increase 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 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.

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 (No. 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 Journal 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 doi: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, Alcolea A, Medina A, Hidalgo J, Slooten LJ (2005) Inverse problem in hydrogeology. Hydrogeology Journal 13(1):206–222

Carrera J, Neuman SP (1986) Estimation of Aquifer Parameters Under Transient and Steady State Conditions: 3. Application to Synthetic and Field Data. Water Resources Research 22:228-242

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

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

Cleveland TG, Yeh WWG (1990) Sampling network design for transport parameter identification. Journal of Water Resources Planning & Management 116:764-783

Dagan G (1984) Solute transport in heterogeneous porous formations. Journal of Fluid Mechanics 145:151 doi:10.1017/s0022112084002858

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

Doherty J (2004) PEST: Model-Independent Parameter Estimation,User's Manual (5th edition). Watermark Numerical Computing, Australia

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

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

Evensen G (2003) The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dynamics 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. Environmental Science & Technology 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. Stochastic Environmental Research and Risk Assessment 17(5):319–328

Gu Y, Oliver DS (2007) An Iterative Ensemble Kalman Filter for Multiphase Fluid Flow Data Assimilation. Spe Journal 12:1990 - 1995

Harbaugh AW, Banta ER, Hill MC, McDonald MG (2000) MODFLOW-2000, The U. S. Geological Survey Modular Ground-Water Model—User Guide to Modularization Concepts and the Ground-Water Flow Process. U.S. Geological Survey Open-File Report 00–92, 121 p

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. Advances in Water Resources 32(6):851–872

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

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. Stochastic Environmental Research and Risk Assessment 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 Resources Research 23:253-272

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

Li G, Reynolds AC (2009) Iterative Ensemble Kalman Filters for Data Assimilation. Spe Journal 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 doi: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 doi:10.1002/2016wr018736

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

Neuman SP (1973) Calibration of distributed parameter groundwater flow models viewed as a multiple objective decision process under uncertainty. Water Resources Research 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 Resources Research 46(3): 374-381 doi:10.1029/2009WR008312

Oliver DS, Chen Y (2011) Recent progress on reservoir history matching: a review. Computational Geosciences 15(1):185-221

Oliver DS, Cunha LB, Reynolds AC (1997) Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. Mathematical Geology 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 doi: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 doi:10.1029/WR022i013p02069

Sun NZ, Yeh WWG (2007) Development of objective-oriented groundwater models: 2. Robust experimental design. Water resources research 43(2) doi: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. Stochastic environmental research and risk assessment 24(8):1211-23

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

Van Leeuwen PJ, Evensen G (1996) Data Assimilation and Inverse Methods in Terms of a Probabilistic Formulation. Monthly Weather Review 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. Journal of Computational Physics 194:773-794

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

Zheng C (2006) MT3DMS v5.2 supplemental user's guide: Technical report to the U.S. 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. Advances in Water Resources 63:22-37

Tables

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/d)	50
Pumping rate per well (m^3/d)	45

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)*	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

* 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.

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/d)	80
Pumping rate per well (m^3/d)	120
TCE injection concentration per well (mg/L)	50

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

Figures

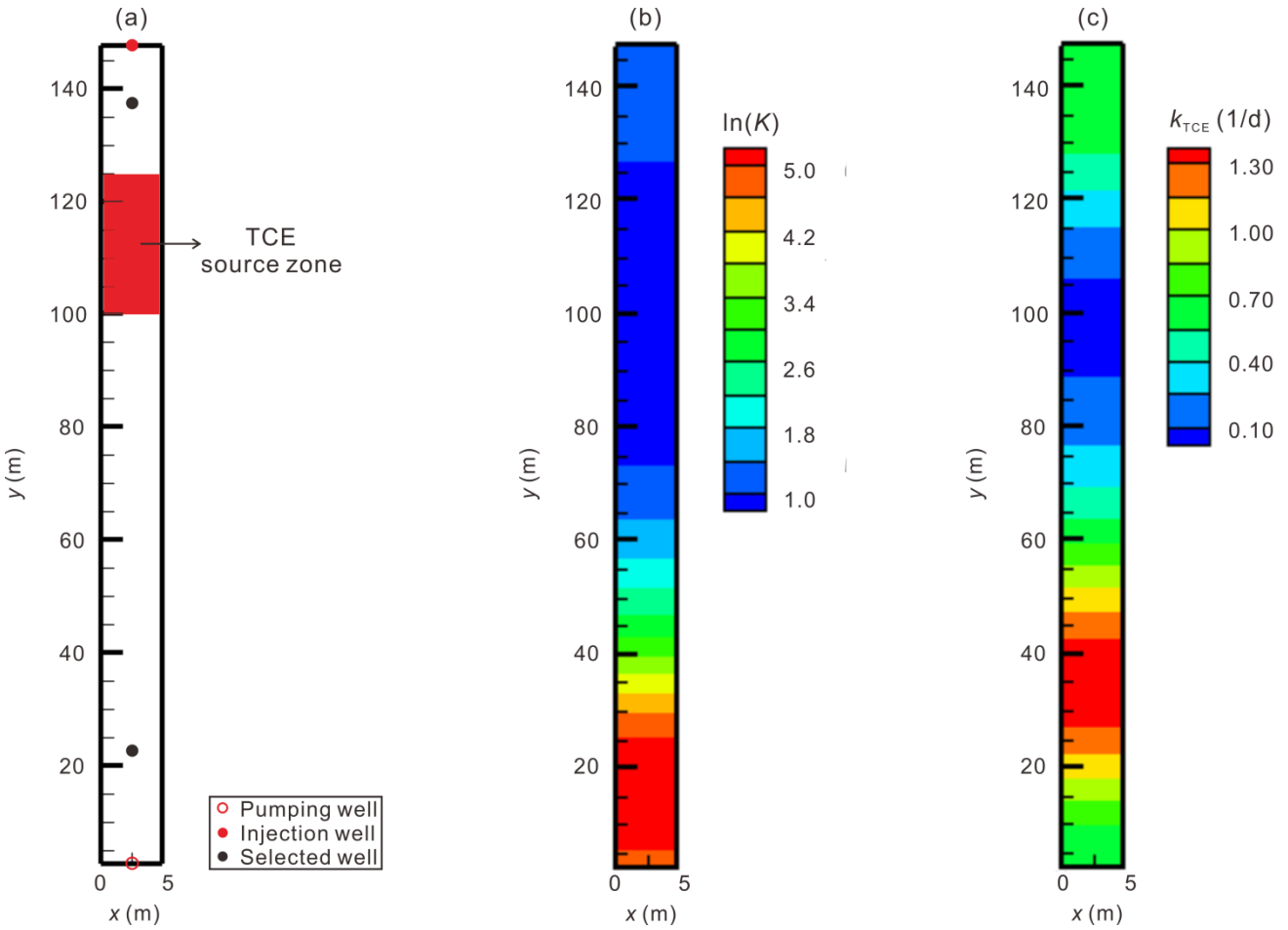


Fig. 1 The conceptual model (a), the reference fields of the hydraulic conductivity (b) and the first-order rate constant (c) for Case 1.

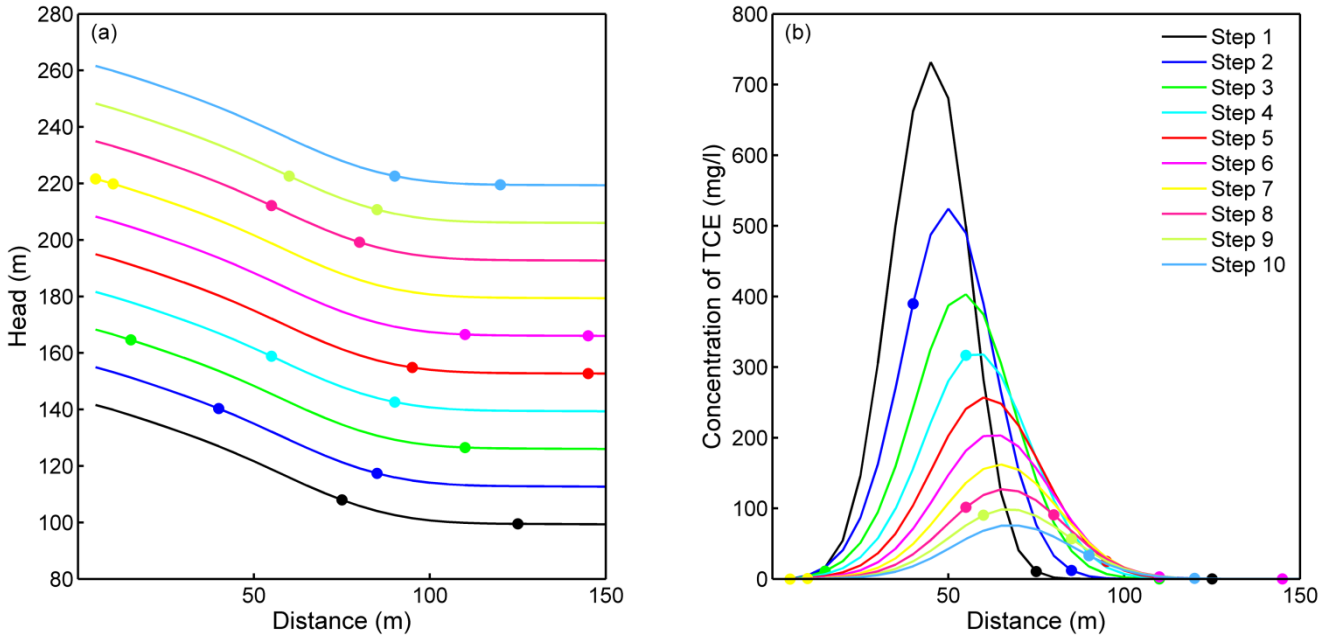


Fig. 2 The calculated flow field (a) and concentration field (b) in Case 1. The circles denote the optimal sampling locations proposed by the SEOD method.

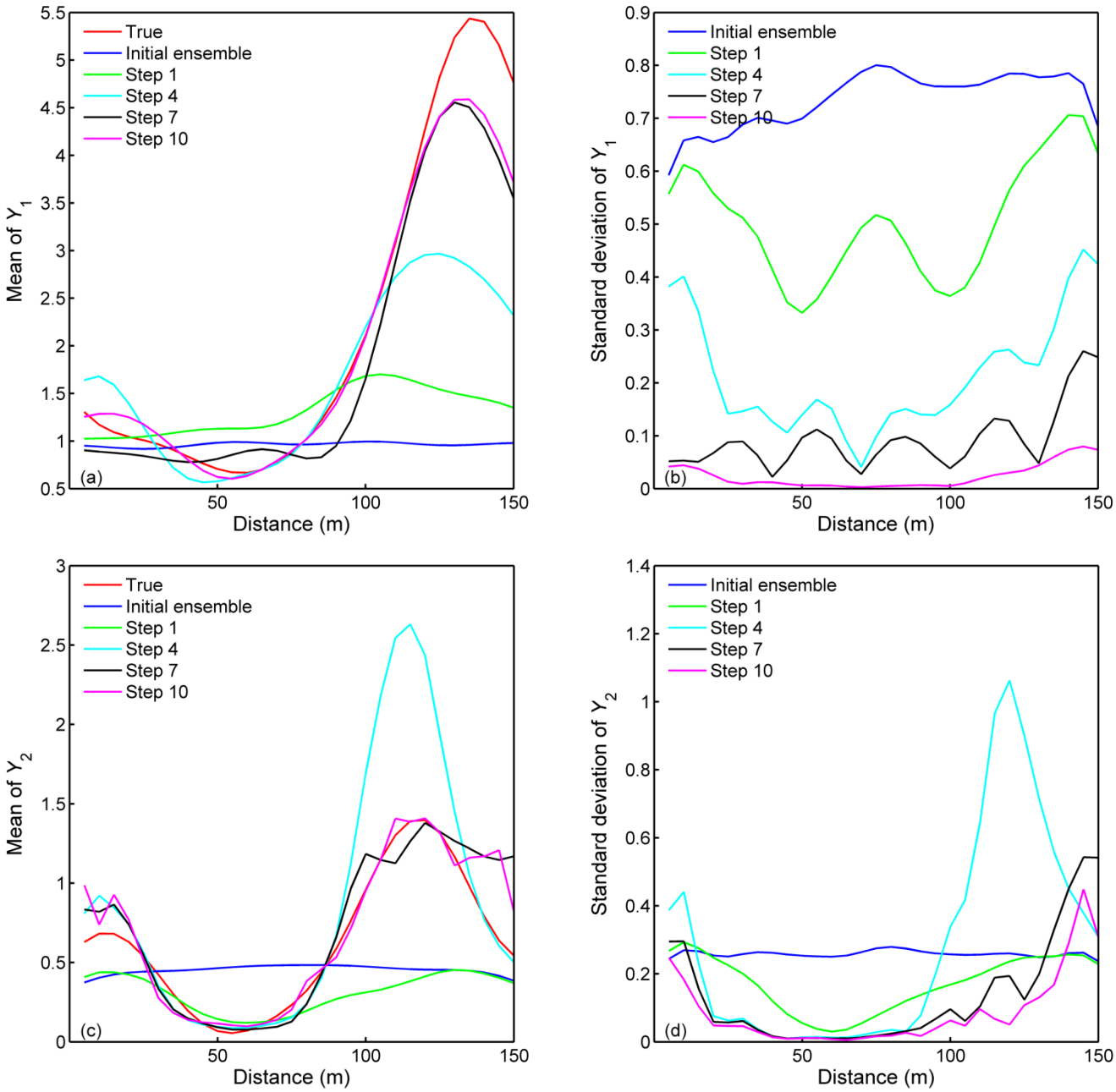


Fig. 3 The ensemble mean and the standard deviation of field Y_1 and Y_2 in Case 1. (a) and (b) are the ensemble mean and the standard deviation of field Y_1 respectively, while (c) and (d) are the ensemble mean and the standard deviation of field Y_2 respectively.

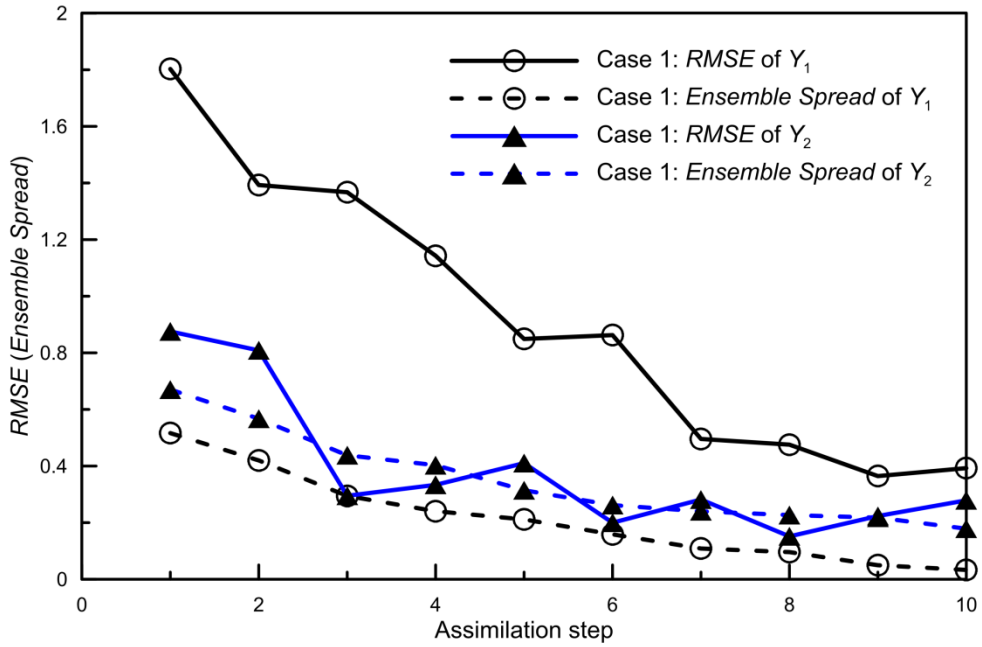


Fig. 4 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for field Y_1 and Y_2 in Case 1.

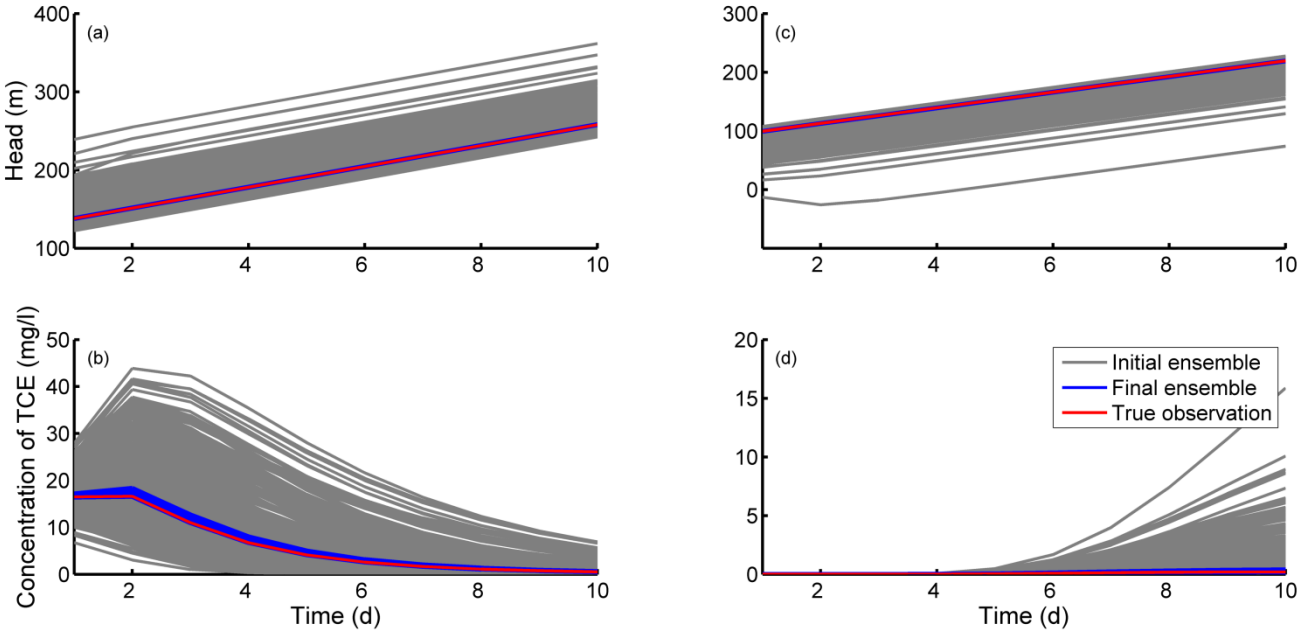


Fig. 5 The performance of data match. (a) and (c) show the data match of the head of two selected wells respectively, while (b) and (d) show the data match of the TCE concentration data of two selected wells respectively.

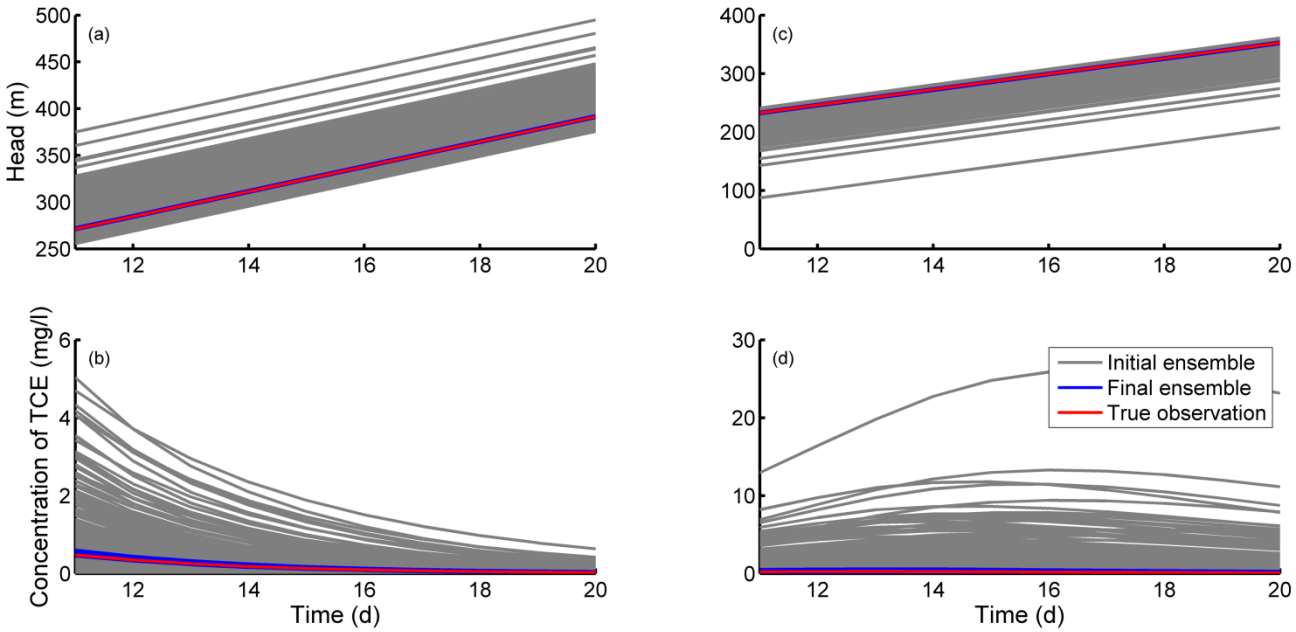


Fig. 6 The performance of model prediction. (a) and (c) show the prediction of the head of two selected wells respectively, while (b) and (d) show the prediction of the TCE concentration data of two selected wells respectively.

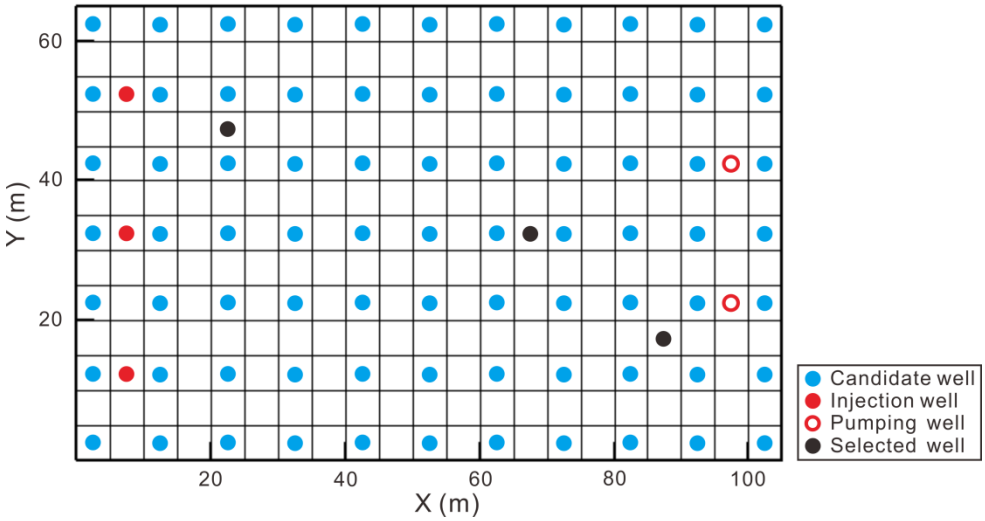


Fig. 7 The schematic of the two-dimensional conceptual model.

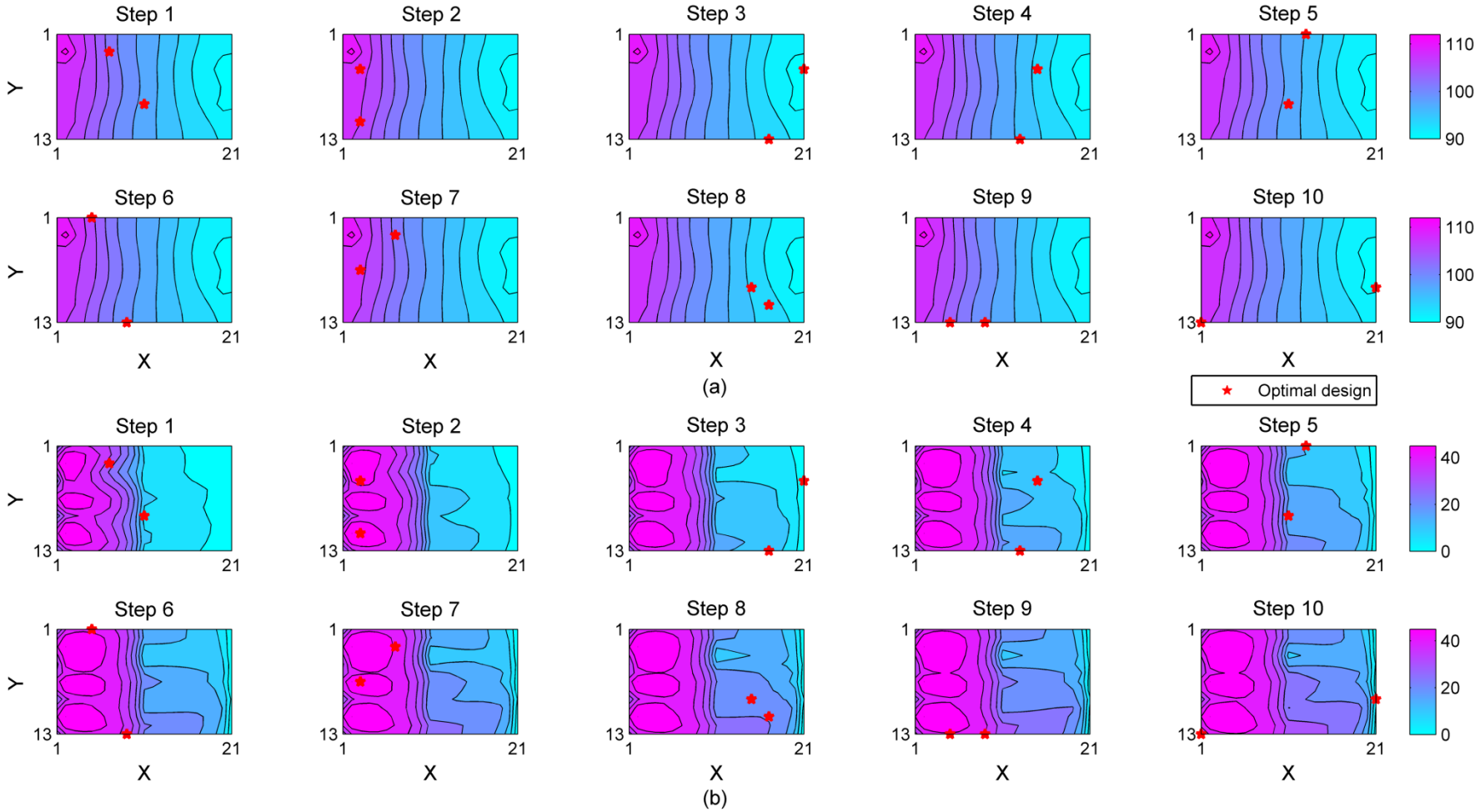


Fig. 8 The optimal sampling locations (red stars) at every assimilation step of Case 2. (a) and (b) are the contour maps of the flow filed and the concentration field, respectively.

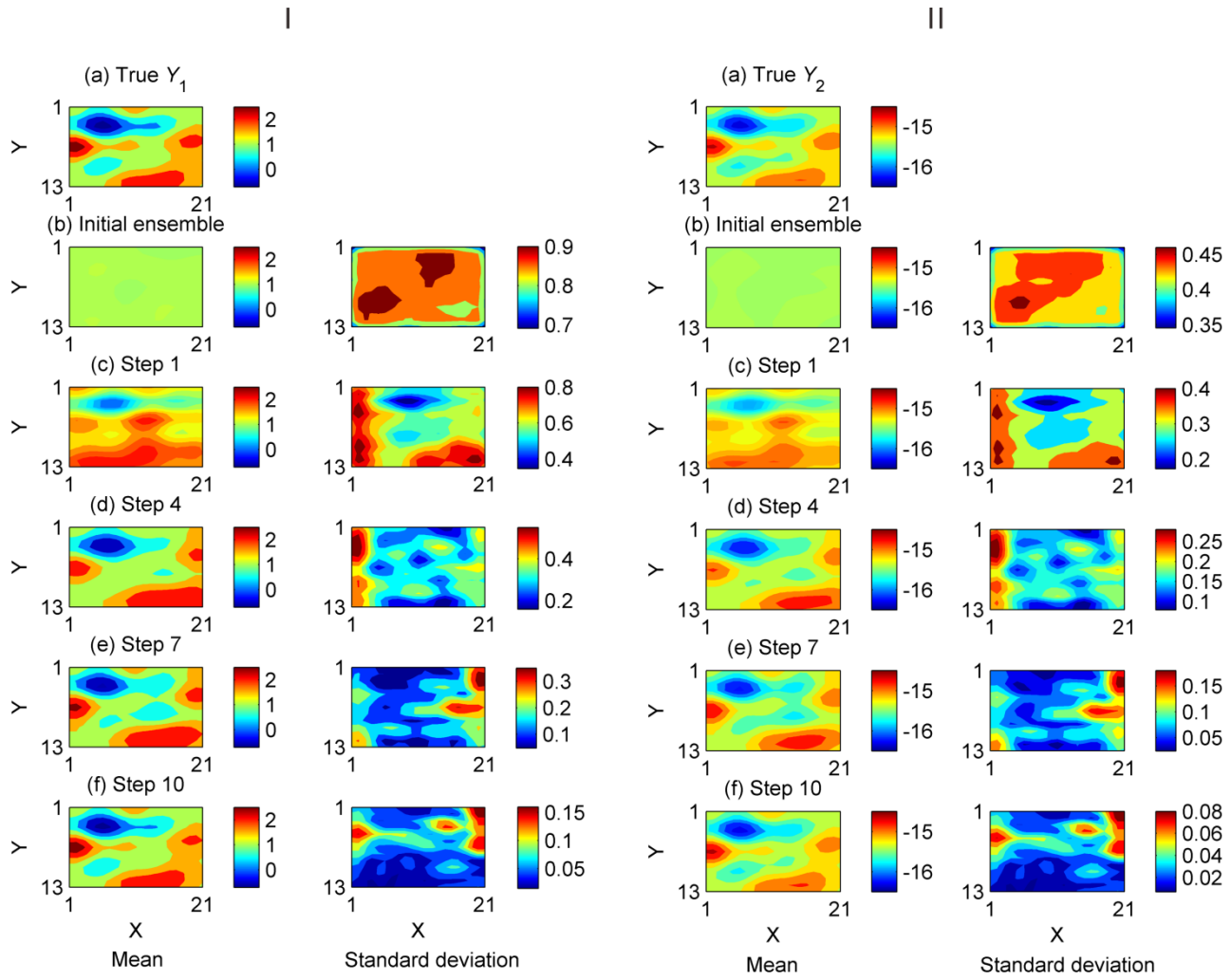


Fig. 9 The ensemble mean and the standard deviation of field Y_1 and Y_2 in Case 2. I for field Y_1, II for field Y_2.

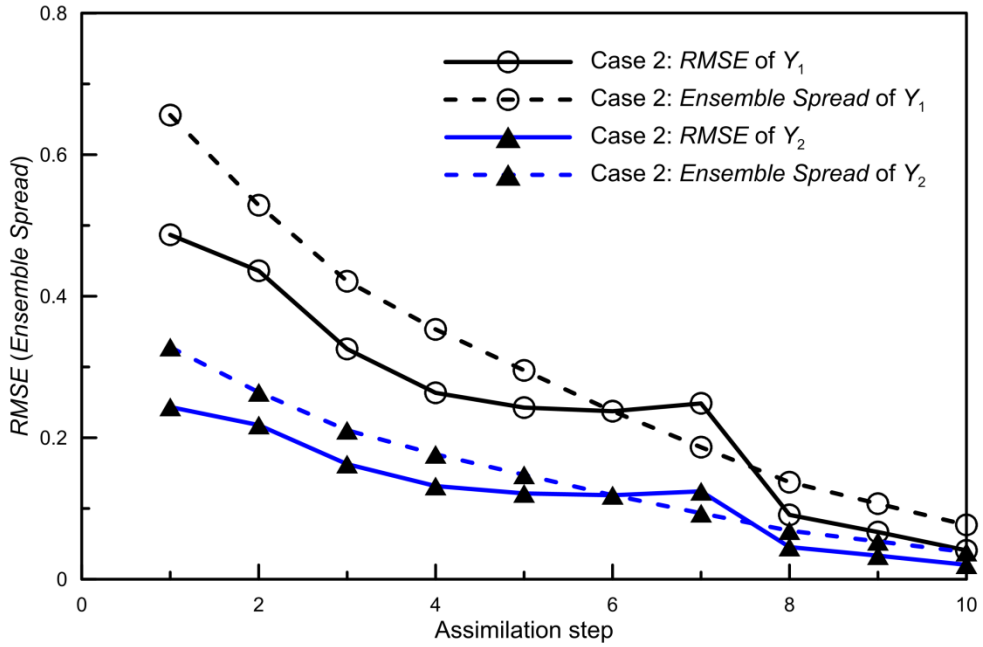


Fig. 10 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for field Y_1 and Y_2 in Case 2.

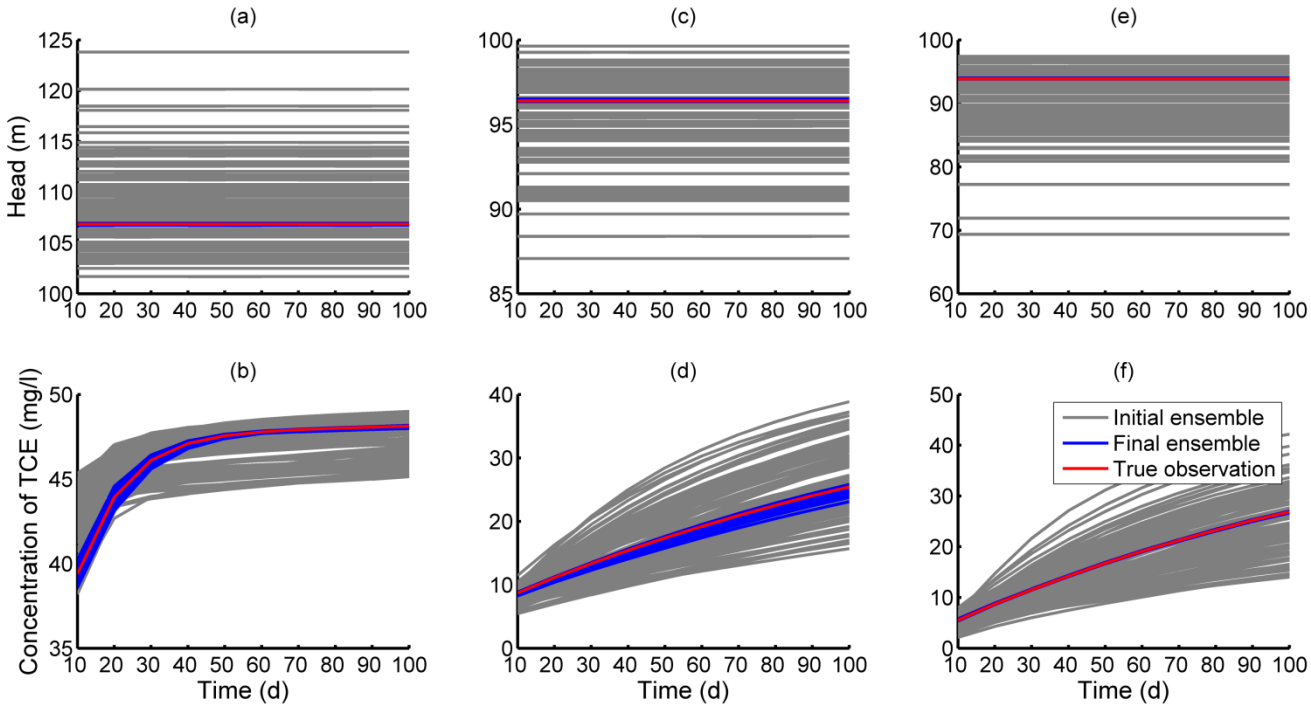


Fig. 11 The performance of data match. (a), (c) and (e) show the data match of the head of three selected wells respectively, while (b), (d) and (f) show the data match of the TCE concentration data of three selected wells respectively.

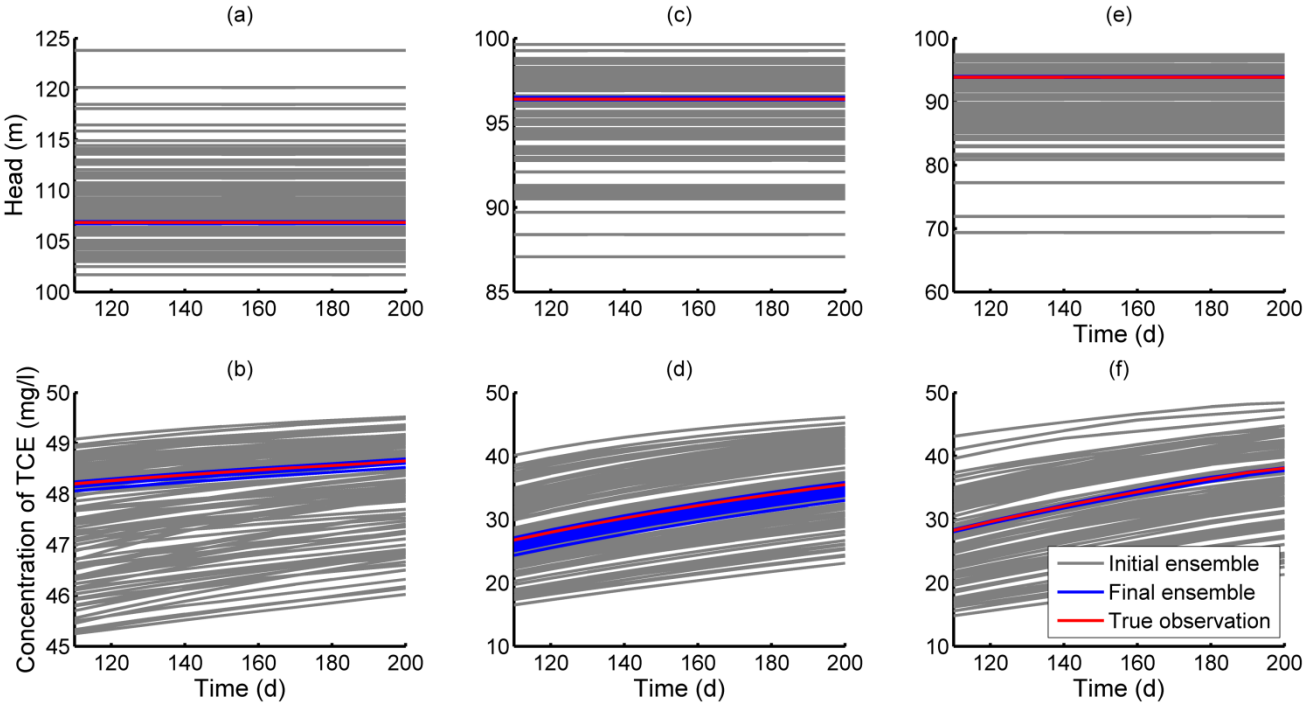


Fig. 12 The performance of model prediction. (a), (c) and (e) show the prediction of the head of three selected wells respectively, while (b), (d) and (f) show the prediction of the TCE concentration data of three selected wells respectively.

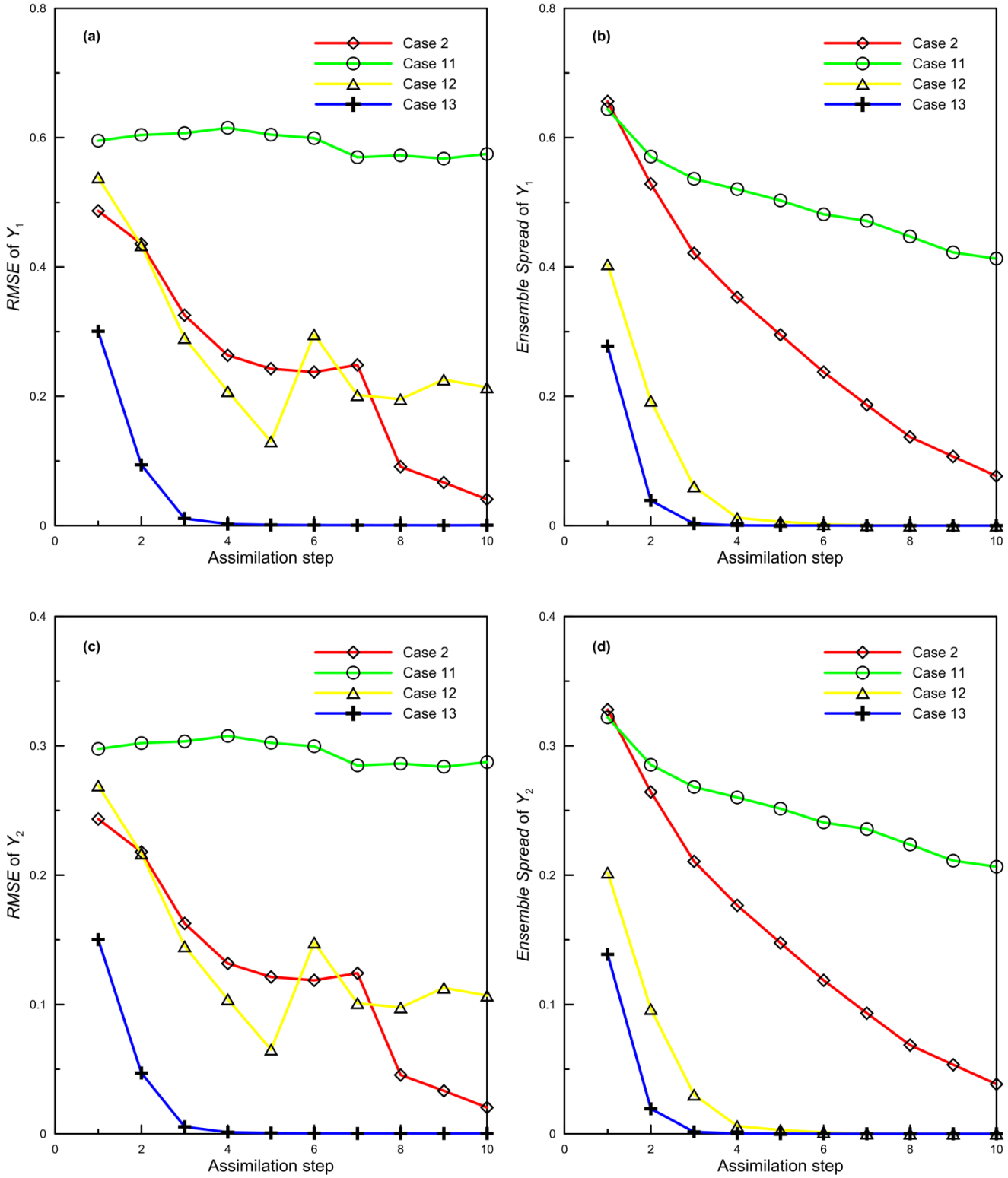


Fig. 13 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different sampling strategies.

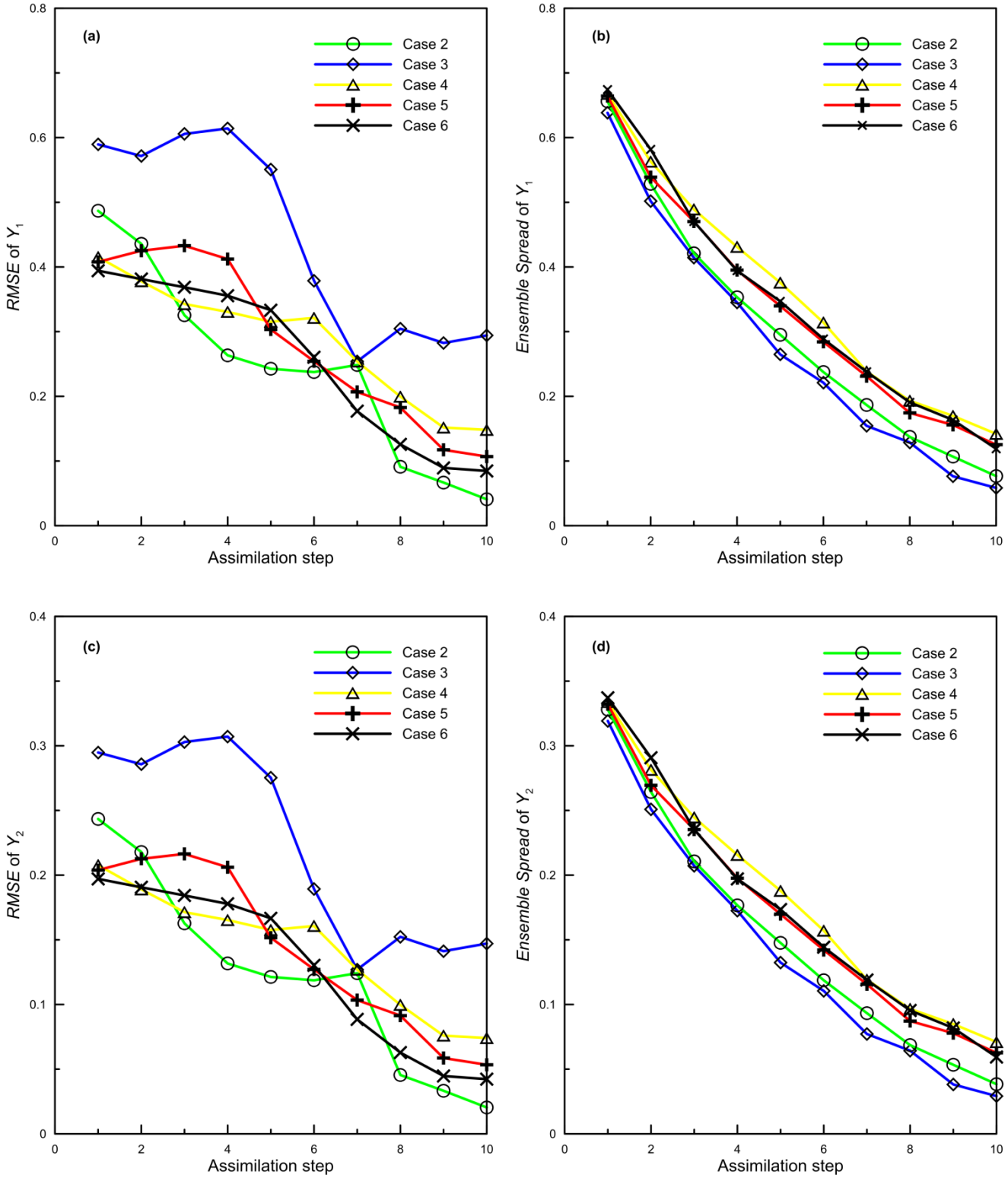


Fig. 14 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different ensemble sizes.

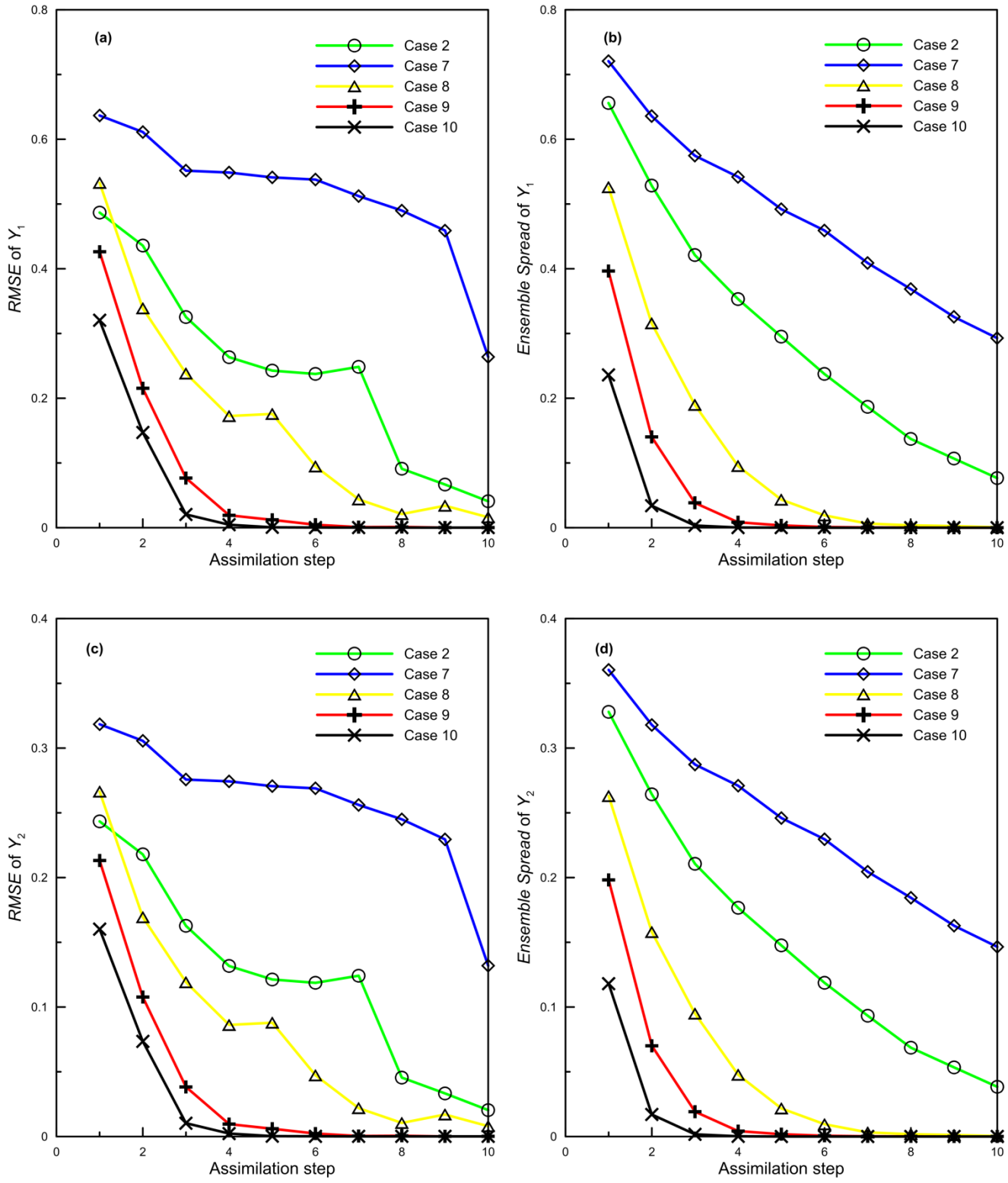


Fig. 15 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different numbers of optimal sampling locations.

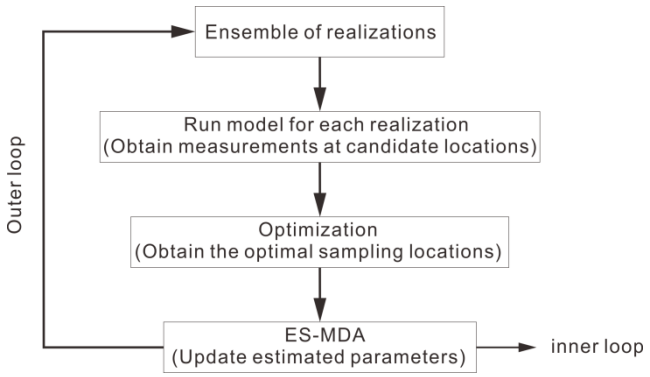


Fig. 16 The loop diagram of the improved SEOD method.

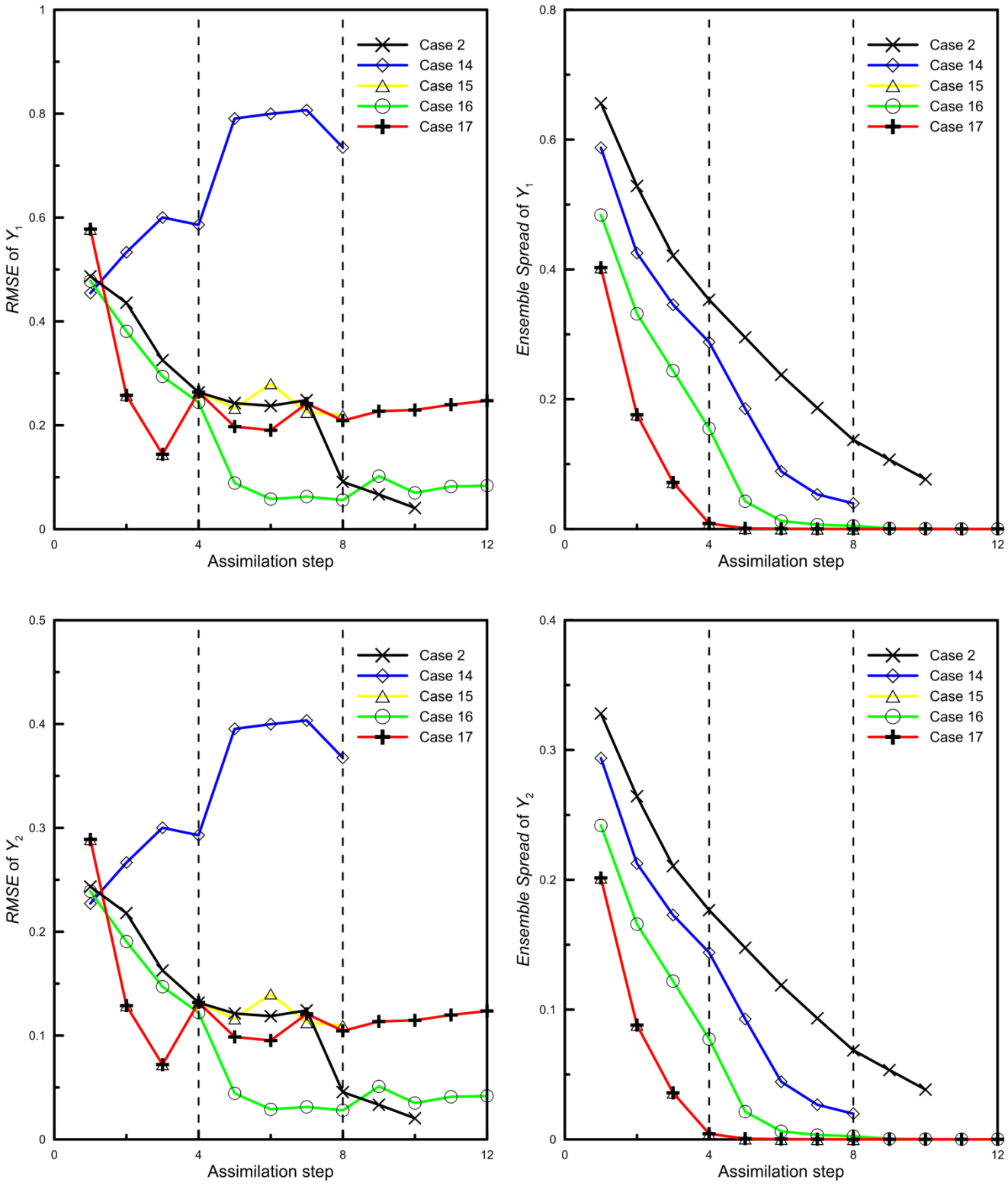


Fig. 17 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different cases using the improved SEOD method.

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,*]

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

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

*Corresponding author:

Xiaoqing Shi, Email: shixq@nju.edu.cn; (86) 25-89680839

Jichun Wu, Email: jcwu@nju.edu.cn; (86) 25-89680705

Highlights:

~~1. The heterogeneous physical and geochemical parameters of groundwater models are jointly estimated.~~

~~2. The SEOD method provides an efficient sampling strategy of sampling locations to accurately estimate both physical and geochemical parameters.~~

~~3. The effectiveness and efficiency of the SEOD method for jointly estimating physical and geochemical parameters are illustrated and demonstrated by the comparison with EnKF using fixed sampling locations during estimation.~~

1. Both the heterogeneous physical and geochemical parameters of groundwater models are accurately estimated by using the sequential ensemble-based optimal design (SEOD) method.
2. The effectiveness and efficiency of the SEOD method are illustrated and demonstrated by the comparison between the sequential optimization strategy and the conventional strategy.
3. To enhance its computational efficiency, the SEOD method is improved by replacing the EnKF with the Ensemble Smoother with multiple data assimilation (ES-MDA).

Abstract

The Ensemble Kalman Filter (EnKF) method has been widely applied into the parameter estimation of groundwater models. Its application in joint inversion of physical and geochemical parameters of reactive transport models remains challenging, partly due to the intrinsic heterogeneities of natural porous media and the scarcity of observation data. Since it is generally cost intensive and time-consuming to acquire measurements from the subsurface system, it is beneficial to select the optimal sampling locations that are most informative for model inversion. In this study, the sequential ensemble-based optimal design (SEOD) method is used to obtain the most informative measurements at every assimilation step of EnKF. The performances of SEOD are tested by jointly estimating heterogeneous physical and geochemical parameters in two synthetic cases. The results show that the SEOD method provides an effective designed sampling strategy of sampling locations to accurately estimate heterogeneous distribution of physical and geochemical parameters. We also discuss the impacts of ensemble size and number of optimal sampling locations on parameter estimation. Meanwhile, the comparison between sequential optimization strategy and conventional strategy (using fixed sampling locations during estimation) is also investigated. Results show that larger ensemble size and more optimal sampling locations improve the parameter estimation and the convergence of optimal sampling strategy, however, the computational cost increases accordingly. Besides, the effectiveness and efficiency of the SEOD method for jointly estimating physical and geochemical parameters are also illustrated and demonstrated by the comparison with

~~EnKF using fixed sampling locations during estimation.~~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 (GA) 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 (ES-MDA). 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

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 reaction are simultaneously influenced by geochemical parameters (Atchley et al. 2014; Li et al. 2010; Scheibe et al. 2006). Like physical heterogeneity, the heterogeneity of geochemical parameters exists as well, which could be caused, for example, by spatial variability in the activity of bacteria related to biodegradation (Fennell et al. 2001; Sandrin et al. 2004). Groundwater reactive transport model, combining solute transport model and geochemical model, is an important tool to characterize and explain the coupling processes. Identification of both physical and geochemical parameters of the subsurface environment is critical for reliable contaminant plume prediction, remediation and management. However, there is still great challenge of jointly estimating the spatial distribution of physical and geochemical parameters due to the intrinsic heterogeneities of natural porous media and the scarcity of observation data.

Many ensemble-based history matching methods have been developed and widely applied in parameter estimation of groundwater models, such as Ensemble Kalman Filter (EnKF) (Evensen 2009), ensemble smoother (Van Leeuwen and Evensen 1996), and their iterative variants (Emerick and Reynolds 2013; Li and Reynolds 2009). Among these ensemble-based history matching methods, EnKF is

one of the most popular methods for its simplicity and flexibility in implementation. EnKF is a variant of Kalman filter based on Monte Carlo method. In the EnKF algorithm, it is not necessary to calculate the sensitivities explicitly which are required in optimization-based approaches (Aanonsen et al. 2009, Zhou et al. 2014). Instead, the cross-correlations between model parameters and states calculated from realizations are used (Evensen 2003, 2009). Besides, though EnKF is based on the linear assumption, 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). To broaden its implementation scale and improve its accuracy, many variants have been developed. For example, Gu and Oliver (2007) proposed using the ensemble randomized maximum likelihood (EnRML) method to obtain better results in strong non-linear models; Zhou et al. (2011) proposed NS-EnKF method to solve non-Gaussian problems; Chen and Oliver (2010), Emerick and Reynolds (2011) introduced a distance-based localization scheme into EnKF to mitigate ensemble collapse.

Numerous variants based on EnKF are developed, however, few take sampling design into consideration. In most of EnKF variants mentioned above, the sampling locations were kept fixed during estimation, which is called conventional strategy in the following discussion. It is intuitive that sampling locations predominantly influence the data-worth of measurements, and the parameter estimation can be improved if the measurements are more informative even that the number of measurements is the same. Knopman and Voss (1987, 1988) showed that

concentration data collected at locations that exhibited high sensitivities to model parameters provided the most information. Catania et al. (2004) built and tested a decision model which could obtain the "optimal" parameter estimation with minimum parameters uncertainty and experimental cost. Cleveland and Yeh (1990, 1991) presented two optimization algorithms in an aquifer tracer test for designing a monitoring network, which provided minimum parameter uncertainty. Though many studies have focused on designing optimal sampling strategy for minimizing parameter uncertainty in groundwater models (Carrera and Neuman 1986; Knopman and Voss 1989; Knopman et al. 1991; Nishikawa and Yeh 1989), few related researches have been studied within the EnKF framework. Man et al. (2016) developed a sequential ensemble-based optimal design (SEOD) method, which firstly integrated sequential optimal design and information theory into the EnKF framework seamlessly to provide the most informative measurements for more accurate parameter estimation. He demonstrated the effectiveness of this method by estimating only physical parameters in unsaturated flow models, whereas we apply the SEOD method into jointly estimating physical and geochemical parameters of reactive transport models in this study.

In this work, the spatial distributions of physical and geochemical parameters are jointly estimated by assimilating both head and concentration measurements based on the SEOD method. The rest of the paper is organized as follows. In Section 2, the groundwater reactive transport model and the SEOD method are described. In Section 3, we construct synthetic one-dimensional and two-dimensional groundwater reactive

~~transport model cases. In Section 4, the effects of ensemble size and number of optimal sampling locations are discussed. We also compare the sequential optimization strategy and conventional strategy in section 4. Finally, conclusions are summarized in Section 5.~~ 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 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 Section 2, the groundwater reactive transport model and the SEOD method are described. In Section 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 Section 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 Section 4.4. Conclusions are summarized in Section 5.

2 Methodologies

2.1 Groundwater reactive transport model

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

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

where $\nabla \cdot$ is the divergence operator; ∇ is the gradient operator; 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 [LT^{-1}]; μ_s is the specific storage ~~coefficient~~ 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):

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

where C_n is the aqueous concentration of the nth component [ML^{-3}]; t is the time [T]; D is the diffusion coefficient [L^2T^{-1}]; $v = (-K \nabla H)/\theta$ [L^2T^{-1}]; $r_{reac,n}$ is the concentration change of the nth component caused by reactions; 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}].

Eq. (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~~, proposed recently~~ (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 1977), 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 ~~follows,~~:

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

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 in details can be found~~

~~in *Man et al.* (2016).~~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)).

$$\mathbf{x}_{i,j+1}^{f} = G(\mathbf{x}_{i,j}^{a}),\ \ i=1,2,...,N_e \tag{4}$$

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 $\mathbf{d}_i^{'} = \mathbf{H}^{'}\mathbf{x}_i^{f} + \boldsymbol{\xi}_i$. 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 $\mathbf{d}$ can be obtained and used in the analysis step (Eq. (6)).

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

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 $\mathbf{d}$ 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 ~~follows,~~:

$$RMSE = \sqrt{\frac{1}{N_m}\sum_{i=1}^{N_e}(\overline{Y}_i - Y_i)^2} \tag{74}$$

$$Ensemble\quad Spread\ (ES) = \sqrt{\frac{1}{N_m}\sum_{i=1}^{N_e}\mathrm{var}(Y_i)} \tag{85}$$

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 ~~mean~~ the estimation, while the *Ensemble Spread* measures the uncertainty ~~in~~ of the estimation.

3 Case studies

3.1 Case 1: One-dimensional synthetic case

In this case, a one-dimensional confined aquifer with ~~the~~ a starting head of 100 m

is constructed ~~created~~, in which saturated transient flow is assumed. As shown in Fig. 1(a), we choose the horizontal aquifer ~~is~~ to be 5 m×150 m and the grid space ~~is~~ to be 5 m both in horizontal *x* and *y* direction. Then, a ~~A~~ Trichloroethylene (TCE) leaking area with ~~the~~ an initial concentration of 1000 mg/L is ~~considered in~~ introduced into the aquifer, and the degradation of TCE is ~~regarded as~~ assumed to follow ~~the~~ first-order kinetic reaction. ~~An~~ Furthermore, an injection well and a 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. 1(b), (c)) are jointly estimated. At every assimilation step, 2 optimal sampling locations~~,~~ are selected from 30 candidate locations~~, are obtained~~ to provide the most informative measurements. More details are given in Table 1 and 2.

[Figure 1]

[Table 1]

[Table 2]

The log saturated hydraulic conductivity $Y_1 = \ln(K)$ and the first-order rate constant $Y_2 = k_{TCE}$ are assumed to be Gaussian distributed, with mean μ_{Y_1}=1 and μ_{Y_2}=0.17 and variance σ_{Y_1}=1 and σ_{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:

$$C_Y(x_1, x_2) = C_Y(x_1, y_1; x_2, y_2) = \sigma^2 \exp[-\frac{|x_1 - x_2|}{\lambda_x} - \frac{|y_1 - y_2|}{\lambda_y}] \quad (9\text{~~11~~})$$

where the horizontal correlation length λ_x = 5 m, and the vertical correlation length λ_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 field so as to obtain the most informative measurements. It is ~~of~~ interesting to note that most optimal sampling locations are ~~around~~ located at the front of the contaminant plume, which means that ~~sampling locations around the contaminant plume front~~ these locations can provide the most informative measurements.

[Figure 2]

In Fig. 3, we plot the curves of the ensemble mean and the standard deviation at different assimilation ~~stages~~ 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. ~~In addition~~ 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.

[Figure 3]

[Figure 4]

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

[Figure 5]

[Figure 6]

3.2 Case 2: Two-dimensional synthetic case

In this case, saturated transient flow is assumed in ~~the~~ a two-dimensional confined aquifer with ~~the~~ a starting head of 50 m. As shown in Fig. 7, we choose the horizontal aquifer ~~is~~ to be 105 m×65 m and the grid space ~~is~~ to be 5 m both in horizontal x and y direction. Three TCE leaking sources with the constant injection flow of 80 m^3/d and the constant concentration of 50 mg/L per well are set upstream in the aquifer.~~, and~~ 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/d per well are set downstream (Fig.7). ~~All~~ In addition, all boundaries of the aquifer are assumed to be impermeable. In this case, the hydraulic conductivity (K) and the liner sorption constant (k_d)~~, parameters to be estimated,~~ 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 Table 2 and 3.

[Figure 7]

[Table 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 m and 20 m respectively. With these statistics, the reference field and initial ensemble members of Y_1 can be ~~are~~ generated by the K-L ~~(Karhunen-Loeve) D~~decomposition based on Eq. (9~~6~~). For field *Y2*, 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 ~~on~~. 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.

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

field ~~so as~~ to obtain the most informative measurements.

[Figure 8]

The contour maps of the ensemble mean and the standard deviation at different assimilation ~~stages~~ steps are plotted in Fig. 9. It ~~is shown~~ 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 ~~Besides~~, 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 ~~it's~~ a crucial role in the model inversion though only 2 sampling locations are selected at every assimilation step.

[Figure 9]

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

[Figure 10]

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 ~~then~~. 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 ~~of data match and model prediction~~ due to the space limitation. As shown in Fig. 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.~~, which~~ 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.

[Figure 11]

[Figure 12]

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.

Fig.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.

[Figure 13]

4.~~1~~ 2 Effect of ensemble size

All results shown so far ~~until now~~ 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~~in this subsection~~.

The *RMSE* and the *Ensemble Spread* of different cases ~~ensemble sizes~~ are shown in Fig. ~~13~~ 14 below. It ~~is shown~~ 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 ~~that~~ 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 ~~is shown~~ 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 ~~following~~ cases discussed below.

[Figure ~~13~~14]

4.~~2~~ 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 ~~is explored.~~, several synthetic cases with different numbers of optimal sampling locations are constructed. More details are given in Table 2.

As shown in Fig. ~~14~~15 below, the *RMSE* is ~~not~~ no longer sensitive to the number of optimal sampling locations ~~any more~~ 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 ~~Besides~~, 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 ~~the method from~~ assimilating further measurements into the model. On the other hand ~~In addition~~, too many optimal sampling locations could lead to high economic cost and heavy computational burden. Therefore, 2 to 5 optimal sampling locations are enough and appropriate in this model.

[Figure ~~14~~15]

~~**4.3 Comparison between sequential optimization strategy and conventional strategy**~~

~~In order to demonstrate the effectiveness and efficiency of the SEOD method for jointly estimating physical and geochemical parameters, sequential optimization strategy is compared with conventional strategy in this subsection. More details are~~

~~given in Table 2.~~

~~The RMSE and Ensemble Spread of different cases are plotted in Fig. 15 below. It is intuitively seen that sequential optimization strategy obtains better performance of parameter estimation when the number of sampling locations is the same. Even more, sequential optimization strategy with 2 optimal sampling locations (Case 2) performs better than conventional strategy with 10 fixed sampling locations (Case 12). Besides, conventional strategy with large number of fixed sampling locations could lead that the Ensemble Spread becomes very small at first few assimilation steps, which could prevent assimilating further measurements.~~

~~[Figure 15]~~

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]

Fig. 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.

(I) 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)).

$$\mathrm{x}_{i,j+1}^{f} = G(\mathrm{x}_{i,j}^{a}), \quad i = 1, 2, ..., N_e \tag{10}$$

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.

(III) Analysis step

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

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

In the above equation, l is the times index of the ES-MDA, $l = 1, 2, ..., N_a$; $\mathbf{d}_{obs}$

is the perturbed observations with noise of covariance $\alpha_l \mathbf{C}_D$ (α_1=9.333, α_2=7.0, α_3=4.0 and α_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]

Fig. 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.

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 Case 15, Case 16 and Case 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, the effectiveness and efficiency of the sequential ensemble based optimal design (SEOD) method for jointly estimating the spatial distribution of physical and geochemical parameters are illustrated and demonstrated by two synthetic cases. The results indicate that uncertainties of both physical and~~

~~geochemical parameters decrease after assimilating the most informative measurements from optimal sampling locations.~~

~~Ensemble size and number of optimal sampling locations have impact on the parameter estimation. For ensemble size, too small ensemble size could lead to ensemble collapse, and heavier computational burden and more observation errors could be caused conversely. For the number of optimal sampling locations, when it is large enough, the *RMSE* is not sensitive to the number of optimal sampling locations any more. Overall, appropriate ensemble size and number of optimal sampling locations can estimate parameters well with low computational burden and economic cost.~~

~~Compared with conventional strategy, sequential optimization strategy has better performance of parameter estimation when the number of sampling locations is the same. Even more, sequential optimization strategy with 2 optimal sampling locations (Case 2) performs better than conventional strategy with 10 fixed sampling locations (Case 12).~~

~~In addition, only two kinds of measurements, head and concentration, are assimilated in this work. For further study, more kinds of measurements, such as temperature, geophysical data and so on, can be taken into data assimilation so as to make the most of both solid and soft measurements to improve the accuracy of parameter estimation.~~ 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 increase 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 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.

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 (No. 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 Journal 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 doi: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, Alcolea A, Medina A, Hidalgo J, Slooten LJ (2005) Inverse problem in hydrogeology. Hydrogeology Journal 13(1):206–222

Carrera J, Neuman SP (1986) Estimation of Aquifer Parameters Under Transient and Steady State Conditions: 3. Application to Synthetic and Field Data. Water Resources Research 22:228-242

~~Catania F, Massabó M, Paladino O (2004) Optimal Sampling Strategy for Parameters Estimation.~~

~~In: AGU Fall Meeting, 2004~~
Chen Y, Oliver DS (2010) Cross-covariances and localization for EnKF in multiphase flow data assimilation. Computational Geosciences 14:579-601
Chen Y, Zhang D (2006) Data assimilation for transient flow in geologic formations via ensemble Kalman filter. Advances in Water Resources 29:1107-1122 doi:10.1016/j.advwatres.2005.09.007
~~Cleveland TG, Yeh WG (1991) Optimal Configuration and Scheduling of Ground-Water Tracer Test. Journal of Water Resources Planning & Management 117:37-51~~
Cleveland TG, Yeh WWG (1990) Sampling network design for transport parameter identification. Journal of Water Resources Planning & Management 116:764-783
Dagan G (1984) Solute transport in heterogeneous porous formations. Journal of Fluid Mechanics 145:151 doi:10.1017/s0022112084002858
Dagan G (1985) Stochasti modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem. Water Resources Research 21(1):65–72
Doherty J (2004) PEST: Model-Independent Parameter Estimation,User's Manual (5th edition). Watermark Numerical Computing, Australia
Emerick AA, Reynolds AC (2011) Combining sensitivities and prior information for covariance localization in the ensemble Kalman filter for petroleum reservoir applications. Computational Geosciences 15:251-269
Emerick AA, Reynolds AC (2013) Ensemble smoother with multiple data assimilation. Comput Geosci-Uk 55:3-15 doi:10.1016/j.cageo.2012.03.011
Evensen G (2003) The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dynamics 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. Environmental Science & Technology 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. Stochastic Environmental Research and Risk Assessment 17(5):319–328
Gu Y, Oliver DS (2007) An Iterative Ensemble Kalman Filter for Multiphase Fluid Flow Data Assimilation. Spe Journal 12:1990 - 1995
Harbaugh AW, Banta ER, Hill MC, McDonald MG (2000) MODFLOW-2000, The U. S. Geological Survey Modular Ground-Water Model—User Guide to Modularization Concepts and the Ground-Water Flow Process. U.S. Geological Survey Open-File Report 00–92, 121 p
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. Advances in Water Resources 32(6):851–872
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
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. Stochastic Environmental Research and Risk Assessment 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 Resources Research 23:253-272

~~Knopman DS, Voss CI (1988) Further comments on sensitivities, parameter estimation, and sampling design in one-dimensional analysis of solute transport in porous media. Water Resources Research 24:225–238~~

~~Knopman DS, Voss CI (1989) Multiobjective sampling design for parameter estimation and model discrimination in groundwater solute transport. Water Resources Research 25:2245–2258~~

~~Knopman DS, Voss CI, Garabedian SP (1991) Sampling design for groundwater solute transport: Tests of methods and analysis of Cape Cod tracer test data. International Journal of Rock Mechanics & Mining Sciences & Geomechanics Abstracts 28:925–949~~

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

Li G, Reynolds AC (2009) Iterative Ensemble Kalman Filters for Data Assimilation. Spe Journal 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 doi: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 doi:10.1002/2016wr018736

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

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

~~Nishikawa T, Yeh WG (1989) Optimal pumping test design for the parameter identification of groundwater systems. Water Resources Research 25:1737–1747~~

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

Oliver DS, Chen Y (2011) Recent progress on reservoir history matching: a review. Computational Geosciences 15(1):185-221

Oliver DS, Cunha LB, Reynolds AC (1997) Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. Mathematical Geology 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 et al. (2006) Transport and biogeochemical reaction of metals in a physically and chemically heterogeneous aquifer. Geosphere 2:220-235 doi:10.1130/Ges00029.1~~

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 doi: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 doi:10.1029/WR022i013p02069

Sun NZ, Yeh WWG (2007) Development of objective-oriented groundwater models: 2. Robust experimental design. Water resources research 43(2) doi: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. Stochastic environmental research and risk assessment 24(8):1211-23

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

Van Leeuwen PJ, Evensen G (1996) Data Assimilation and Inverse Methods in Terms of a Probabilistic Formulation. Monthly Weather Review 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. Journal of Computational Physics 194:773-794

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

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

~~Zhou HY, Gómez-Hernández JJ, Franssen HJH, Li LP (2011) An approach to handling non-Gaussianity of parameters and state variables in ensemble Kalman filtering. Adv Water Resour 34:844-864 doi:10.1016/j.advwatres.2011.04.014~~

Zhou HY, Gómez-Hernández JJ, Li LP (2014) Inverse methods in hydrogeology: Evolution and recent trends. Advances in Water Resources 63:22-37

Tables

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/d)	50
Pumping rate per well (m^3/d)	45

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)*	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

* 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.

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/d)	80
Pumping rate per well (m^3/d)	120
TCE injection concentration per well (mg/L)	50

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

Figures

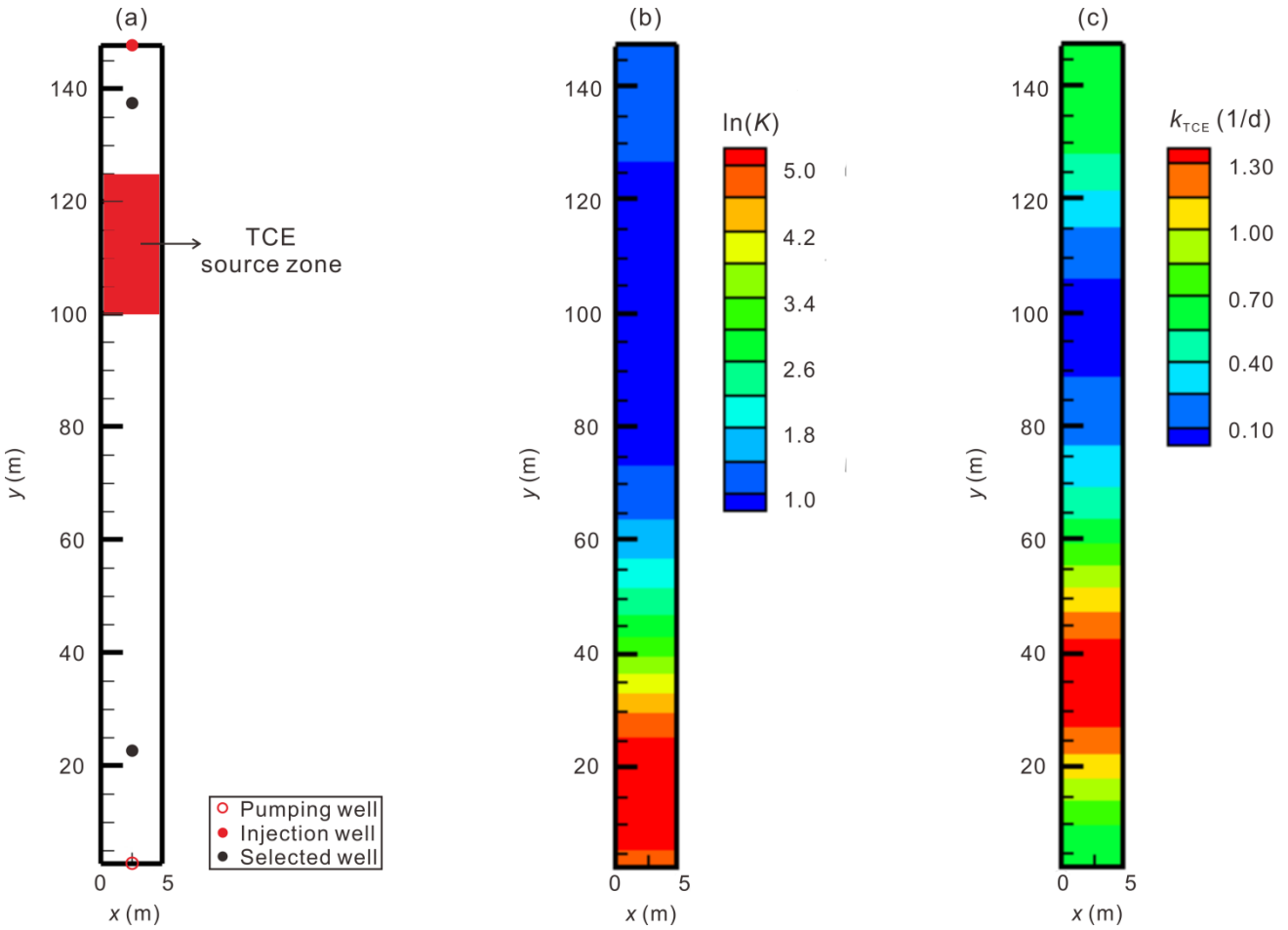


Fig. 1 The conceptual model (a), the reference fields of the hydraulic conductivity (b) and the first-order rate constant (c) for Case 1.

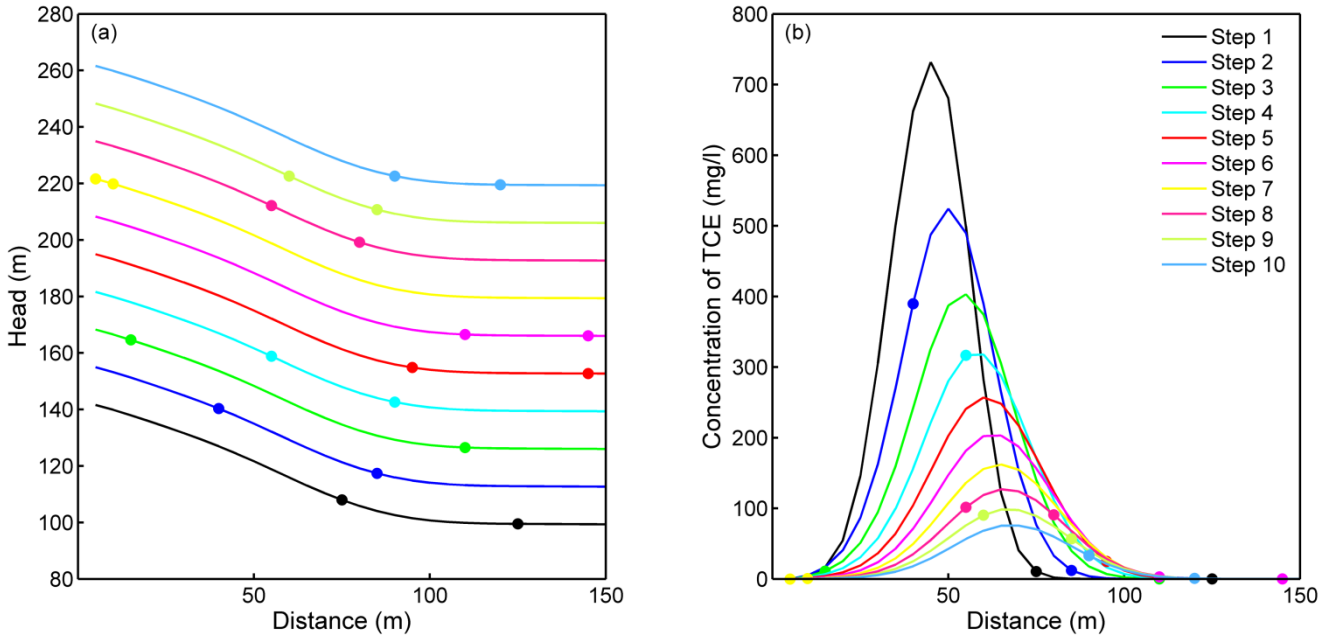


Fig. 2 The calculated flow field (a) and concentration field (b) in Case 1. The circles denote the optimal sampling locations proposed by the SEOD method.

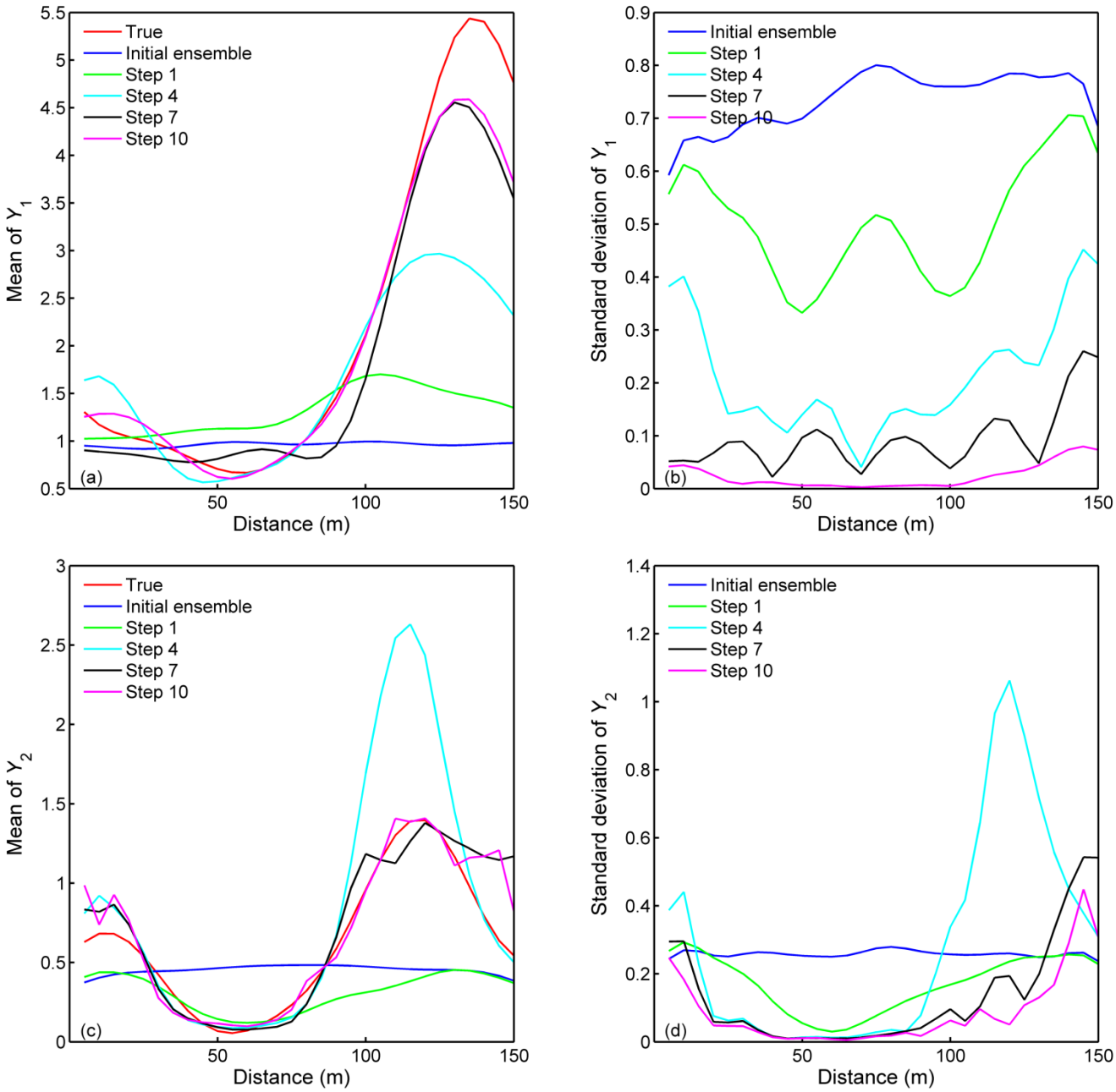


Fig. 3 The ensemble mean and the standard deviation of field Y_1 and Y_2 in Case 1. (a) and (b) are the ensemble mean and the standard deviation of field Y_1 respectively, while (c) and (d) are the ensemble mean and the standard deviation of field Y_2 respectively.

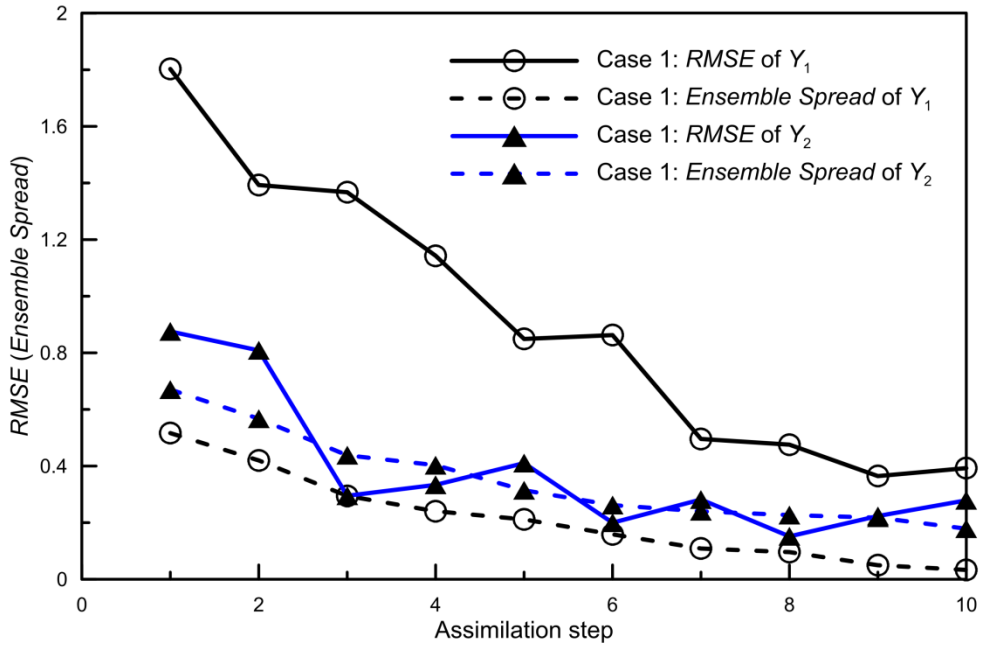


Fig. 4 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for field Y_1 and Y_2 in Case 1.

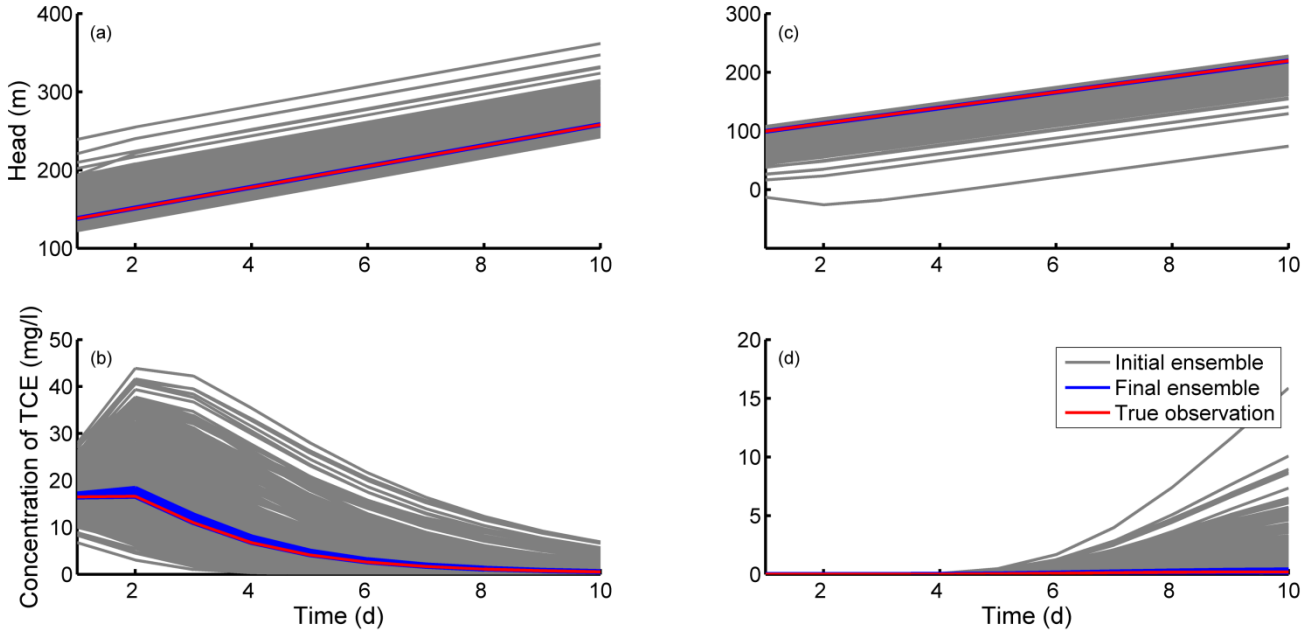


Fig. 5 The performance of data match. (a) and (c) show the data match of the head of two selected wells respectively, while (b) and (d) show the data match of the TCE concentration data of two selected wells respectively. ~~(a) and (b) is the plot of head and TCE concentration data match of well 1 respectively, while (c) and (d) is the plot of head and TCE concentration data match of well 2 respectively~~.

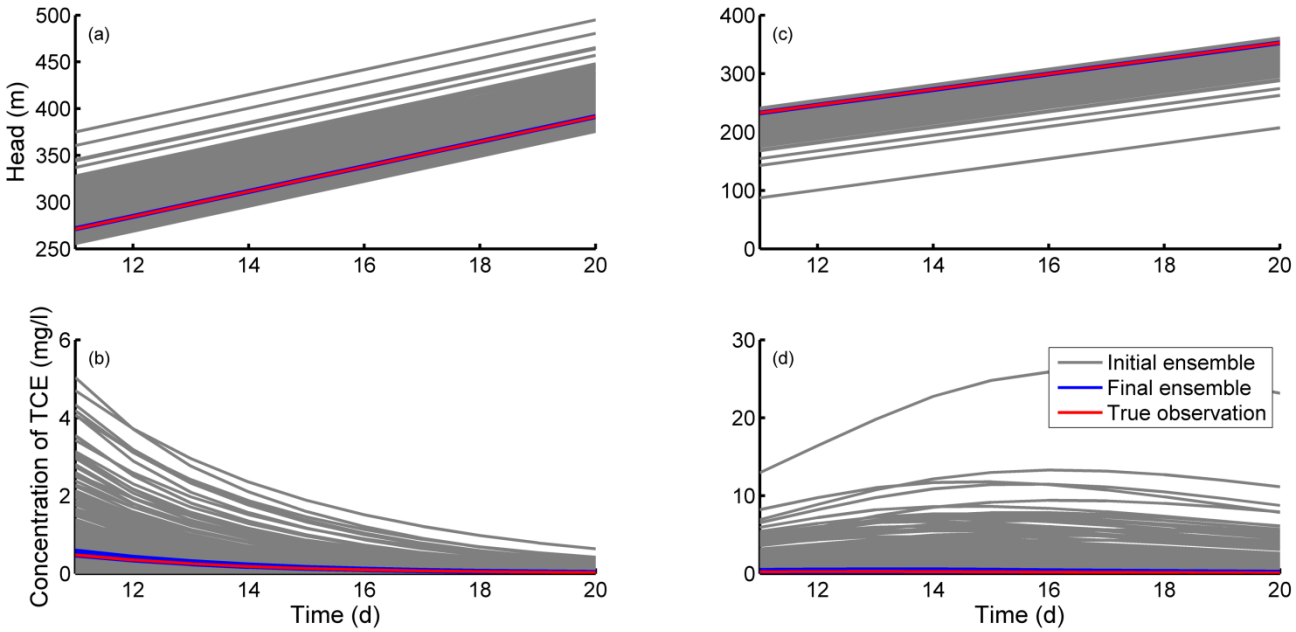


Fig. 6 The performance of model prediction. (a) and (c) show the prediction of the head of two selected wells respectively, while (b) and (d) show the prediction of the TCE concentration data of two selected wells respectively. ~~(a) and (b) is the prediction of head and TCE concentration of well 1 respectively, while (c) and (d) is the prediction of head and TCE concentration data match of well 2 respectively.~~

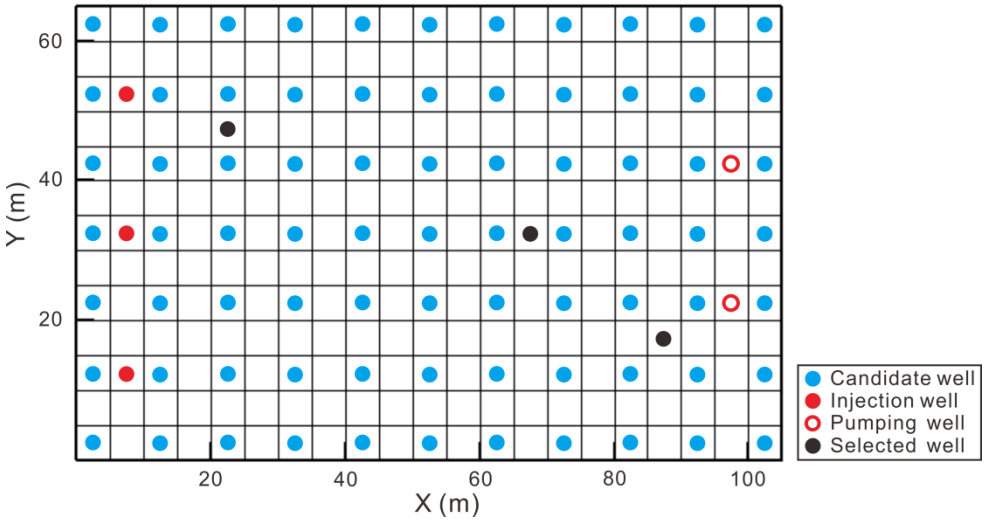


Fig. 7 The schematic of the two-dimensional conceptual model.

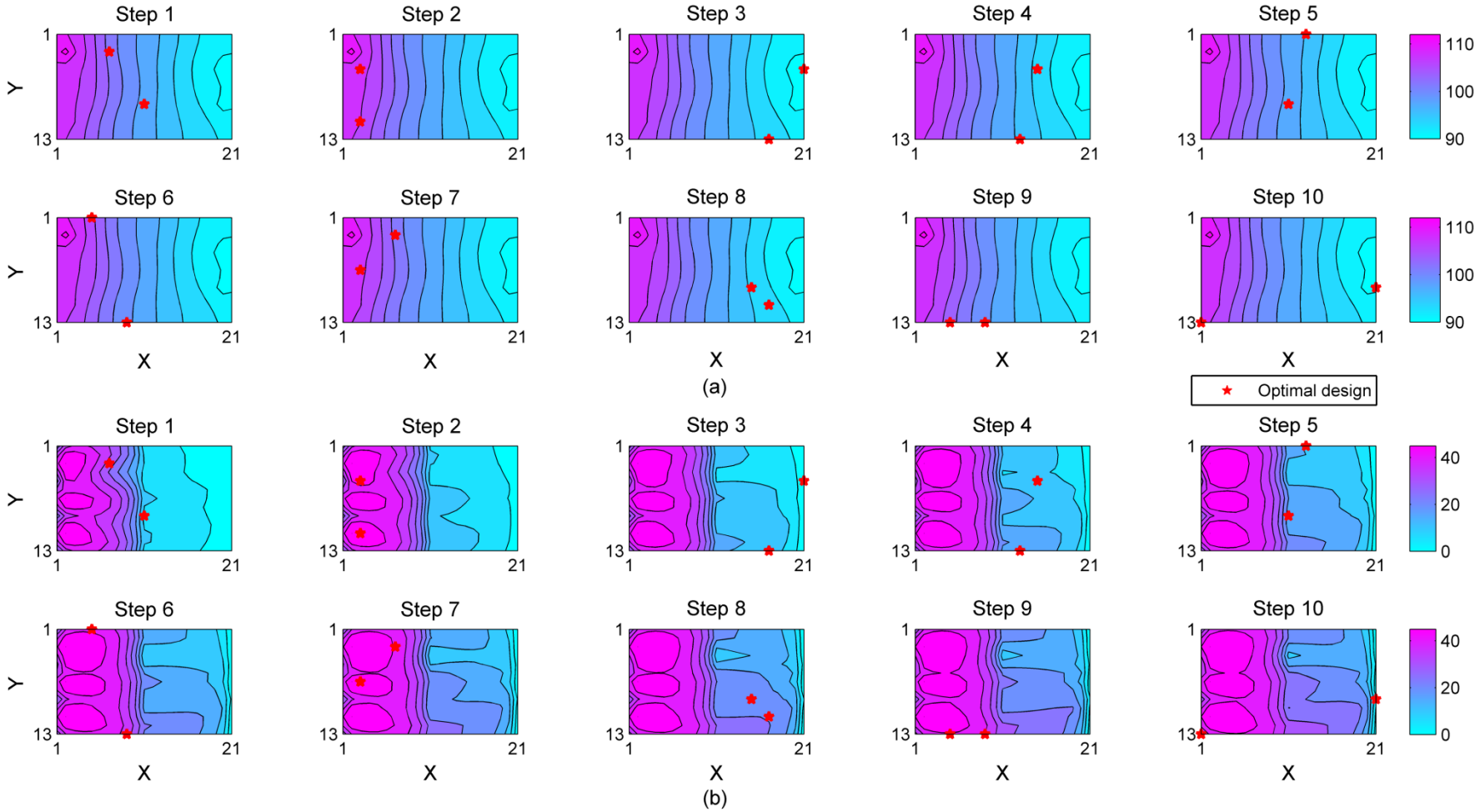


Fig. 8 The optimal sampling locations (red stars) at every assimilation step of Case 2. (a) and (b) are the contour maps of the flow filed and the concentration field, respectively.

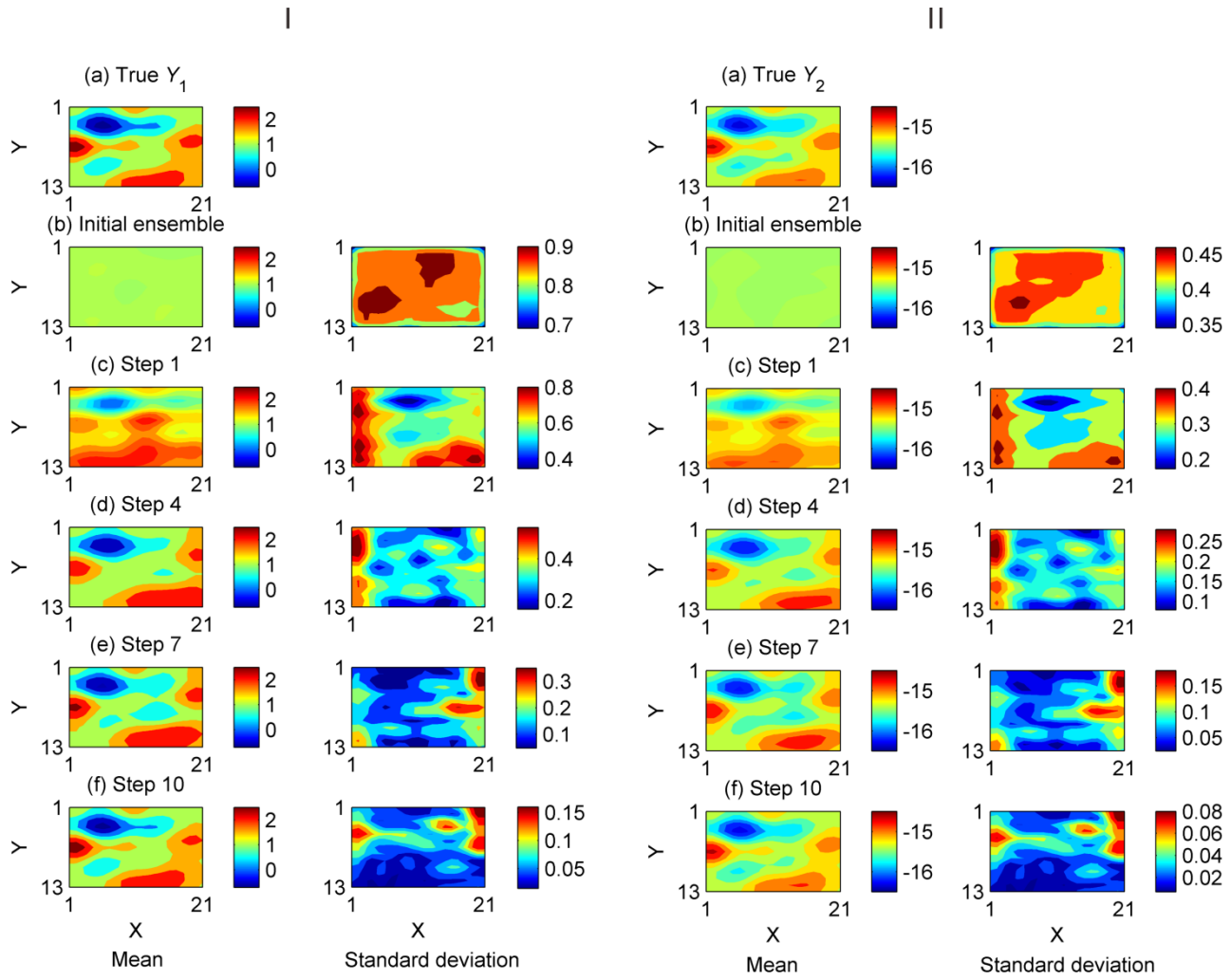


Fig. 9 The ensemble mean and the standard deviation of field Y_1 and Y_2 in Case 2. I for field Y_1, II for field Y_2.

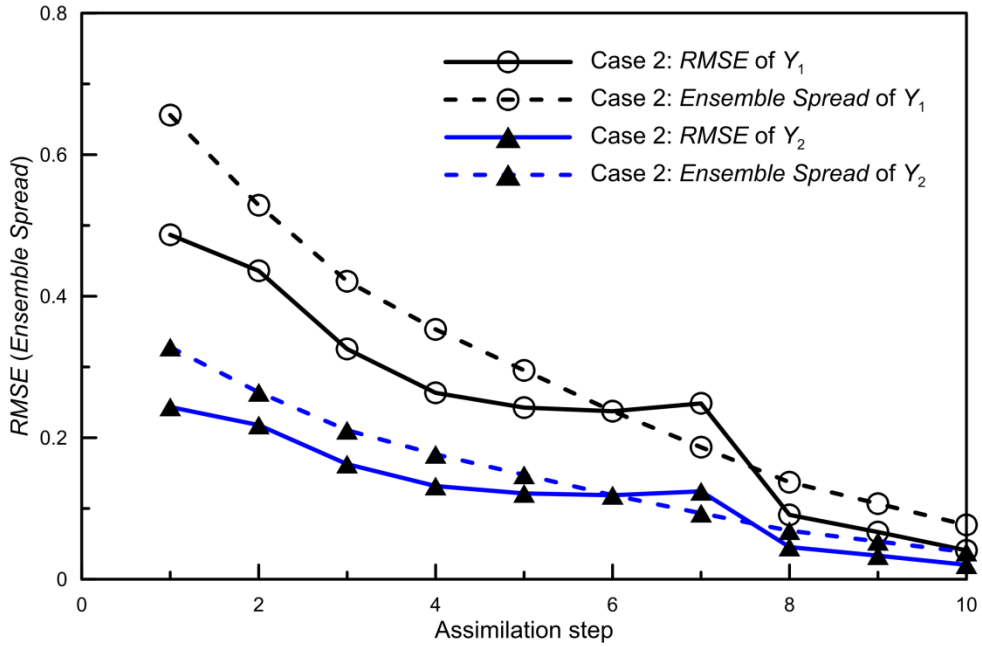


Fig. 10 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for field Y_1 and Y_2 in Case 2.

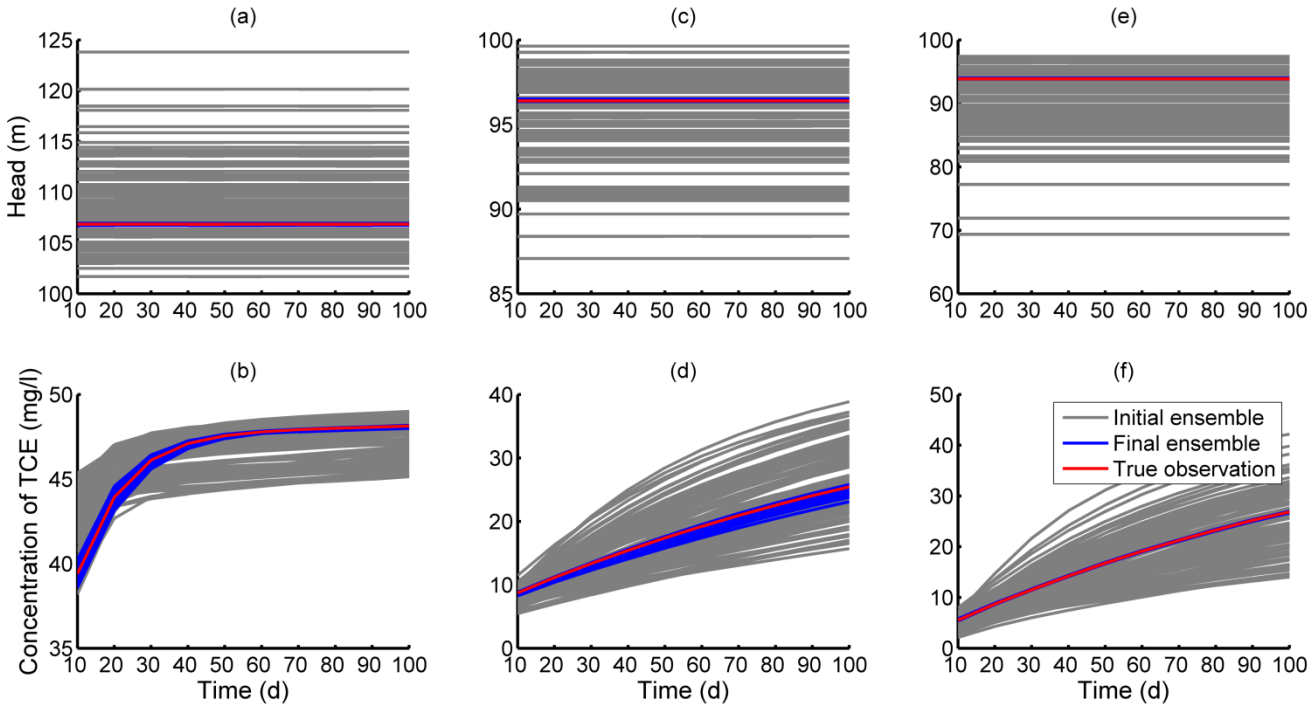


Fig. 11 The performance of data match. (a), (c) and (e) show the data match of the head of three selected wells respectively, while (b), (d) and (f) show the data match of the TCE concentration data of three selected wells respectively.

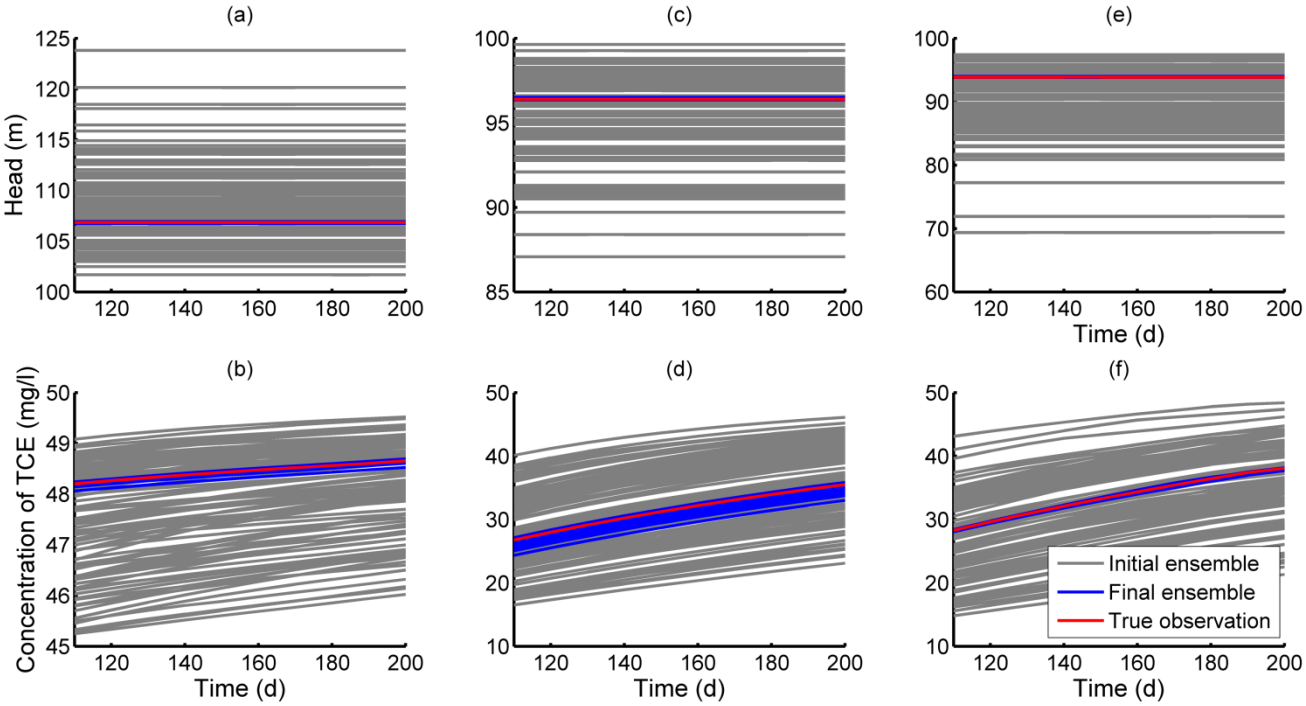


Fig. 12 The performance of model prediction. (a), (c) and (e) show the prediction of the head of three selected wells respectively, while (b), (d) and (f) show the prediction of the TCE concentration data of three selected wells respectively.

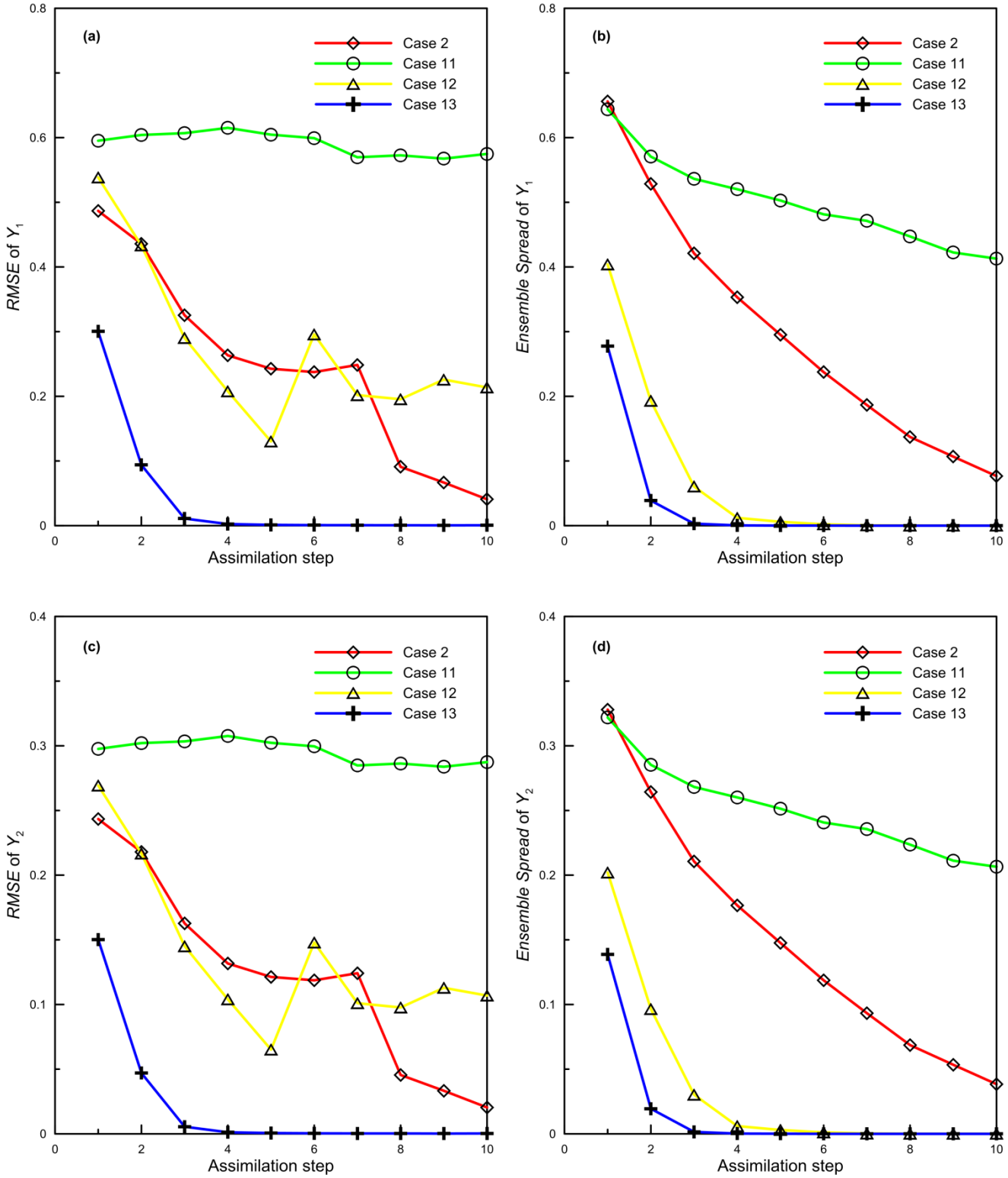


Fig. 13 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different sampling strategies.

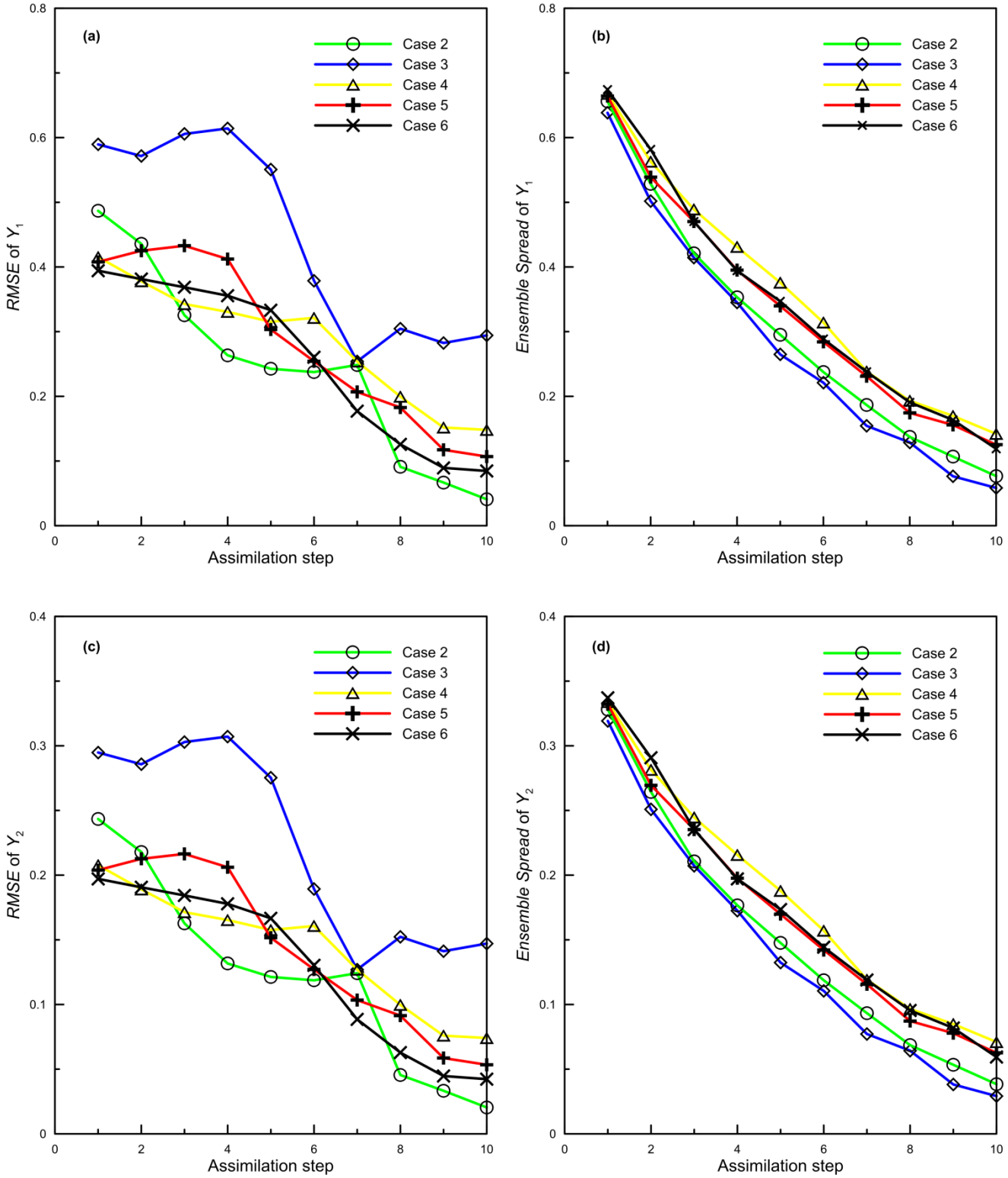


Fig. ~~13~~ 14 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different ensemble sizes.

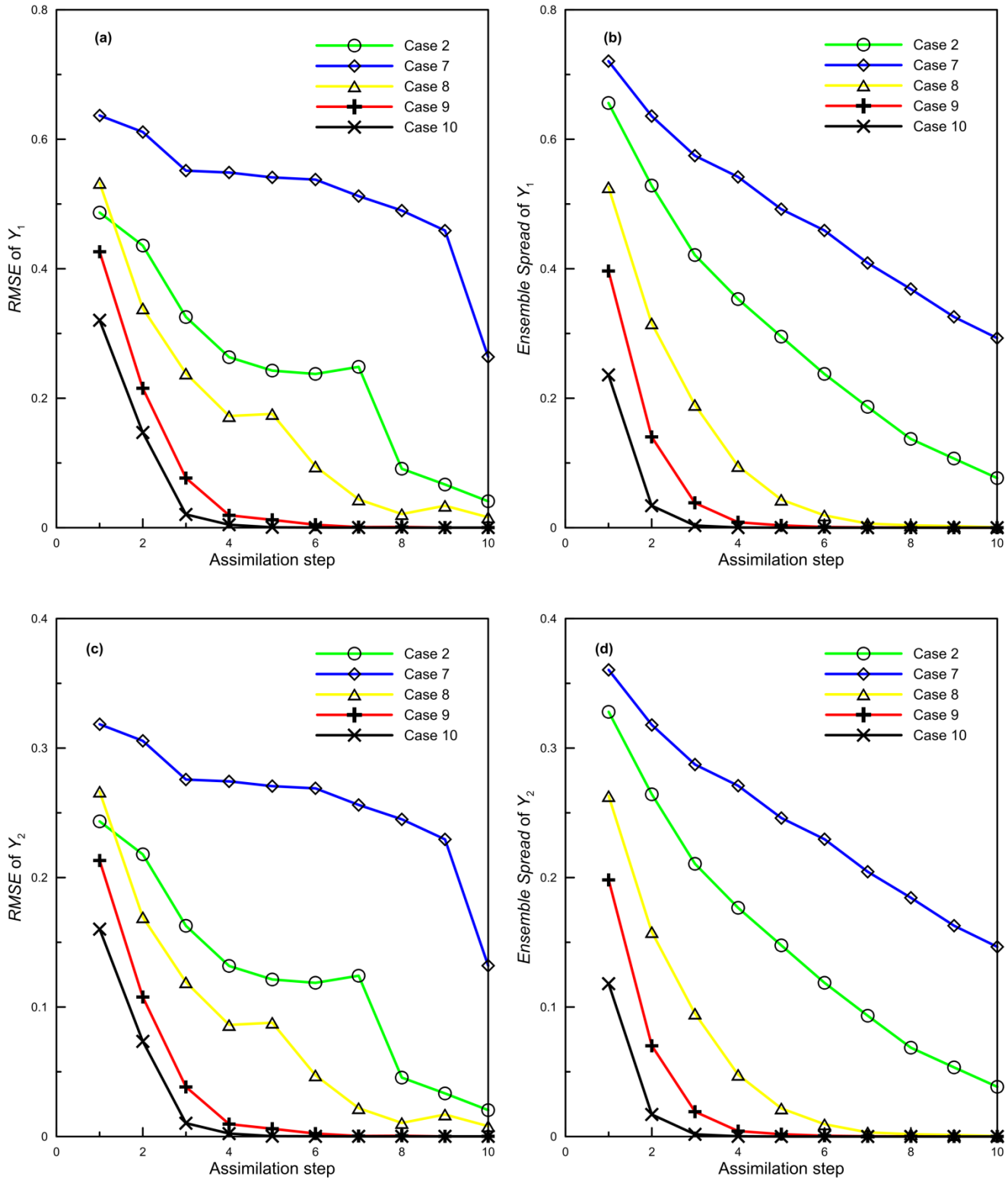
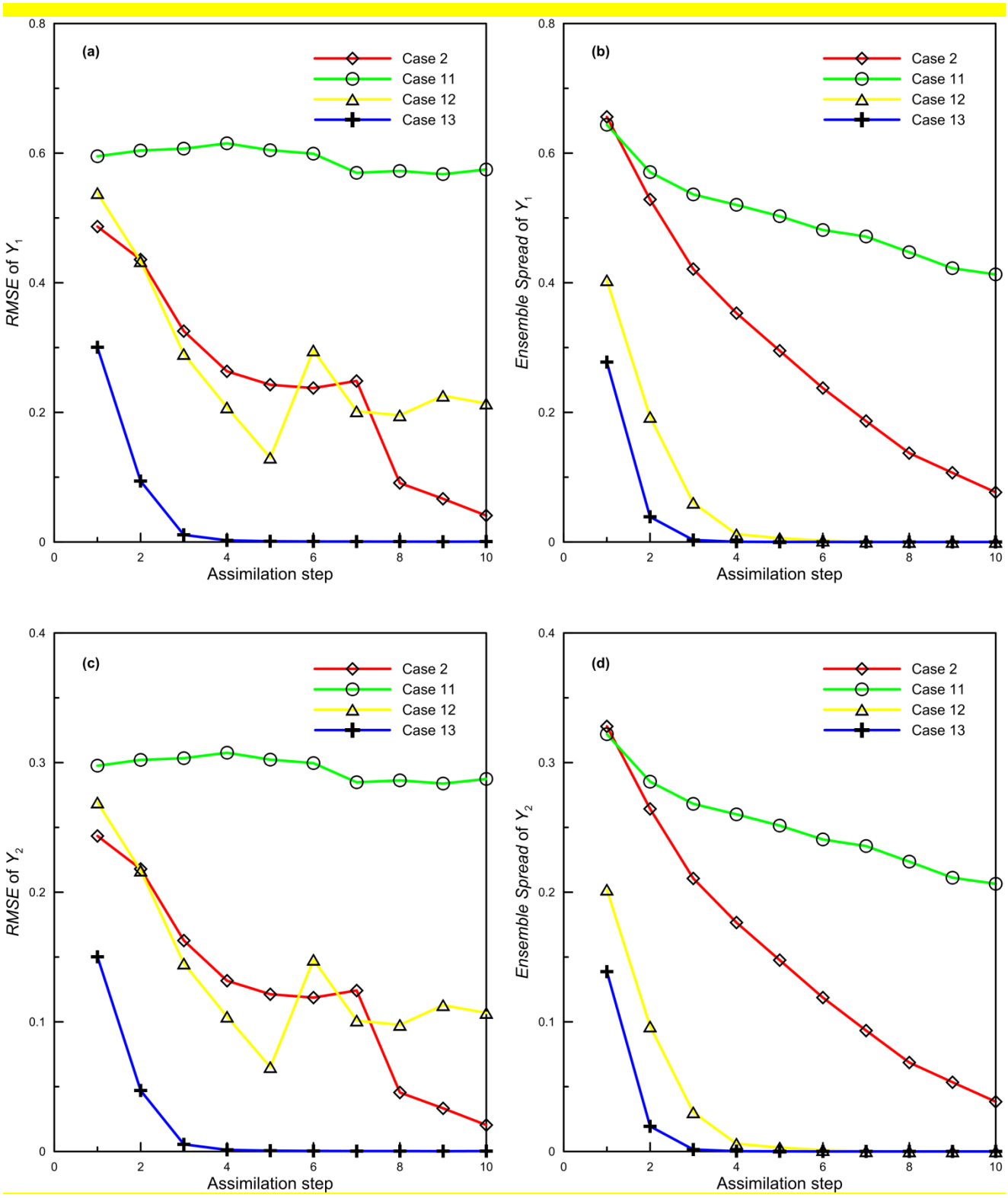


Fig. ~~14~~ 15 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different numbers of optimal sampling locations.



~~Fig. 15~~ ~~The performance indicators (*RMSE* and *Ensemble Spread*) at each assimilation step for different sampling strategies.~~

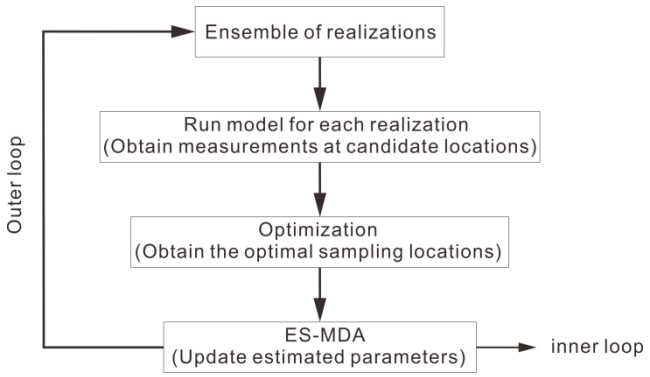


Fig. 16 The loop diagram of the improved SEOD method.

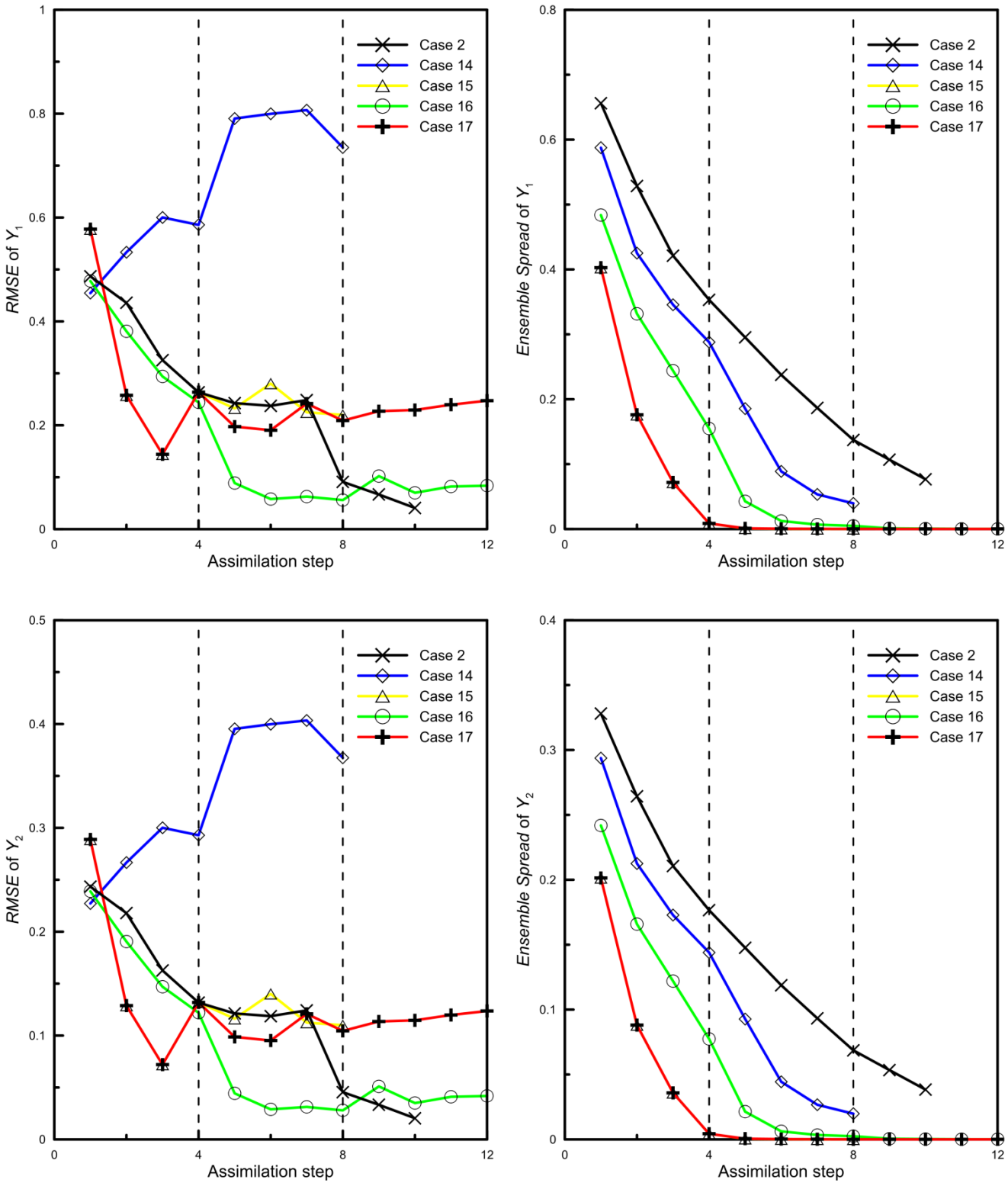
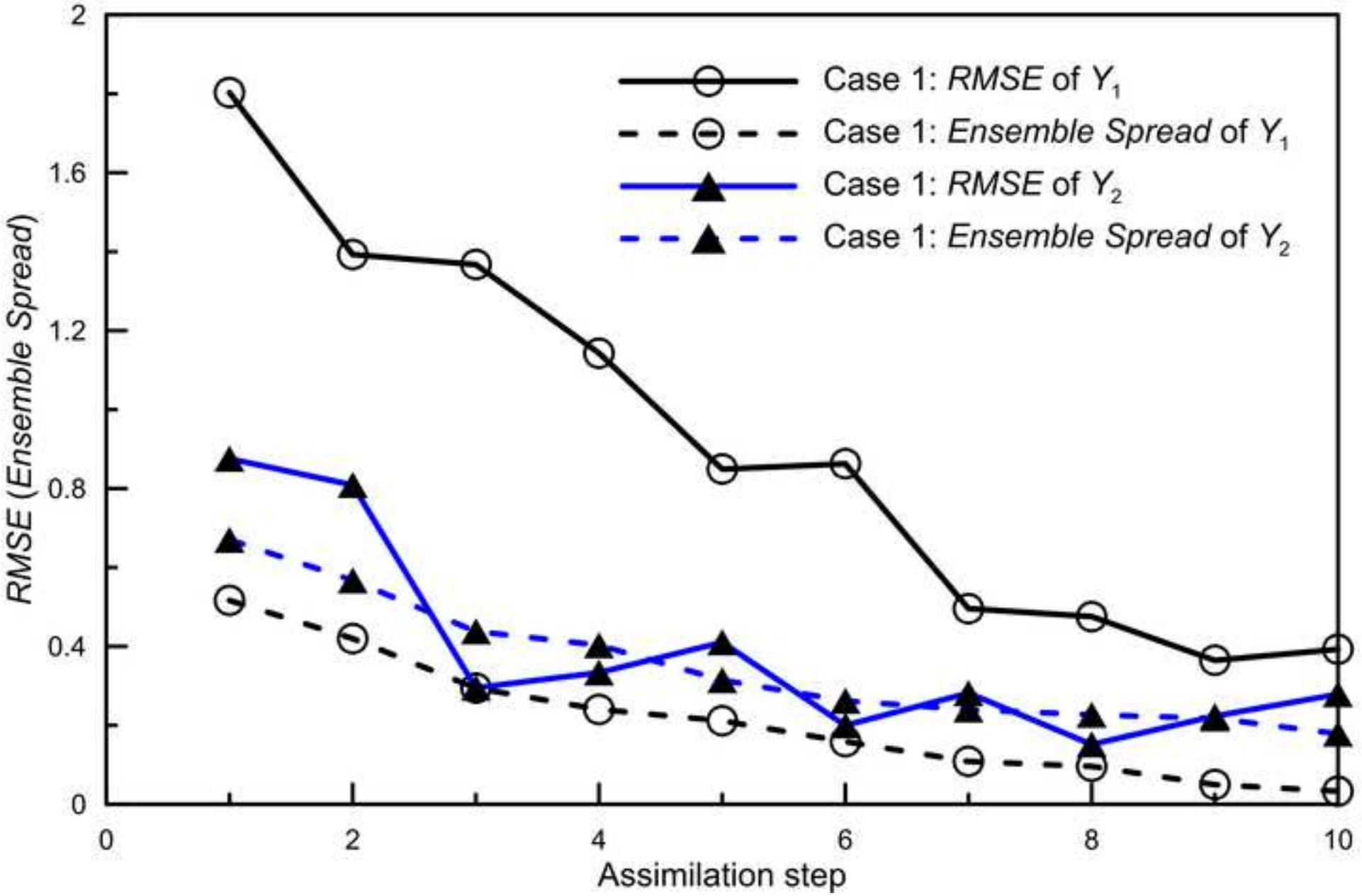


Fig. 17 The performance indicators (the *RMSE* and the *Ensemble Spread*) at each assimilation step for different cases using the improved SEOD method.

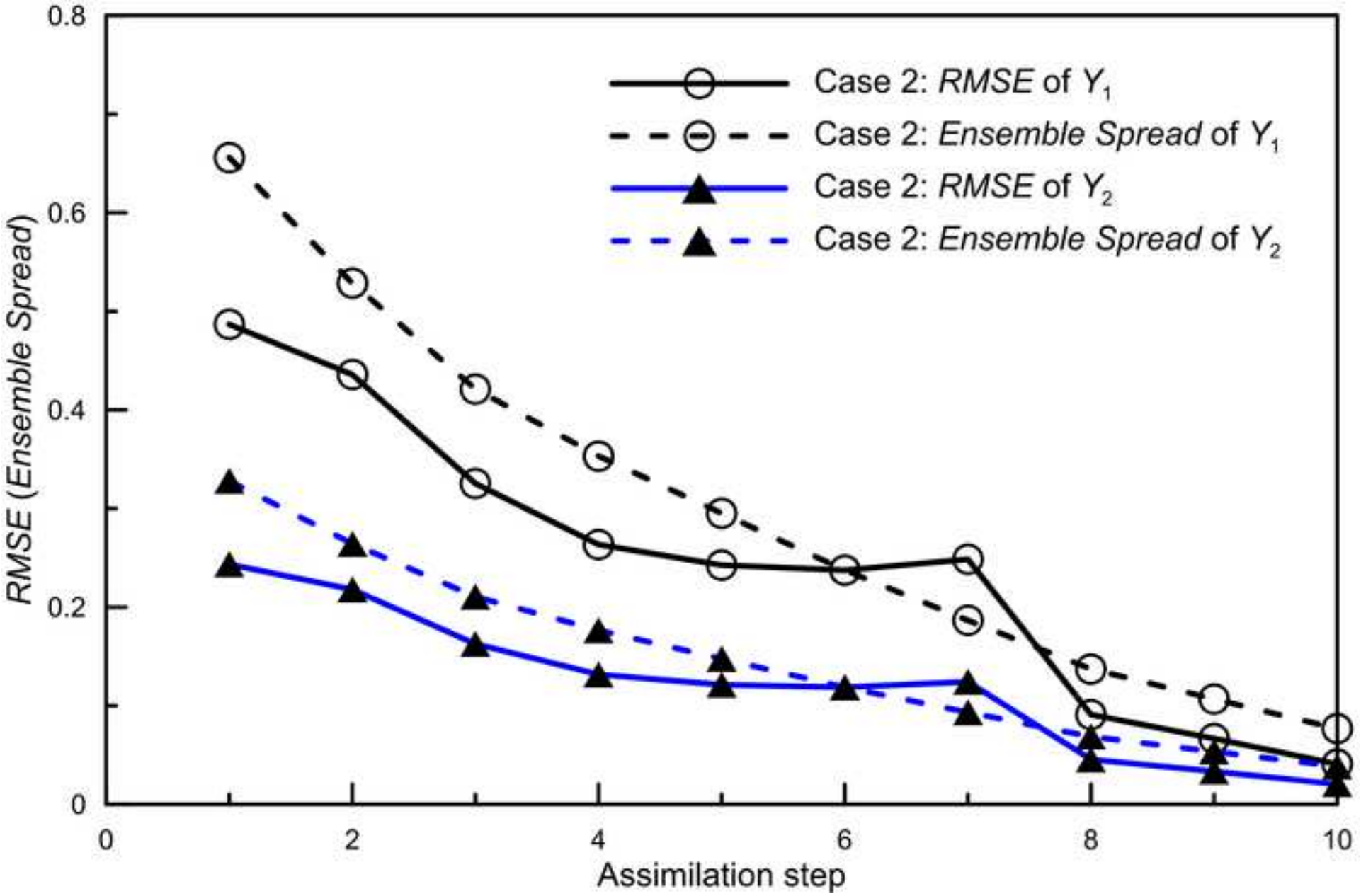
line figure

Click here to download line figure Fig 4.tif



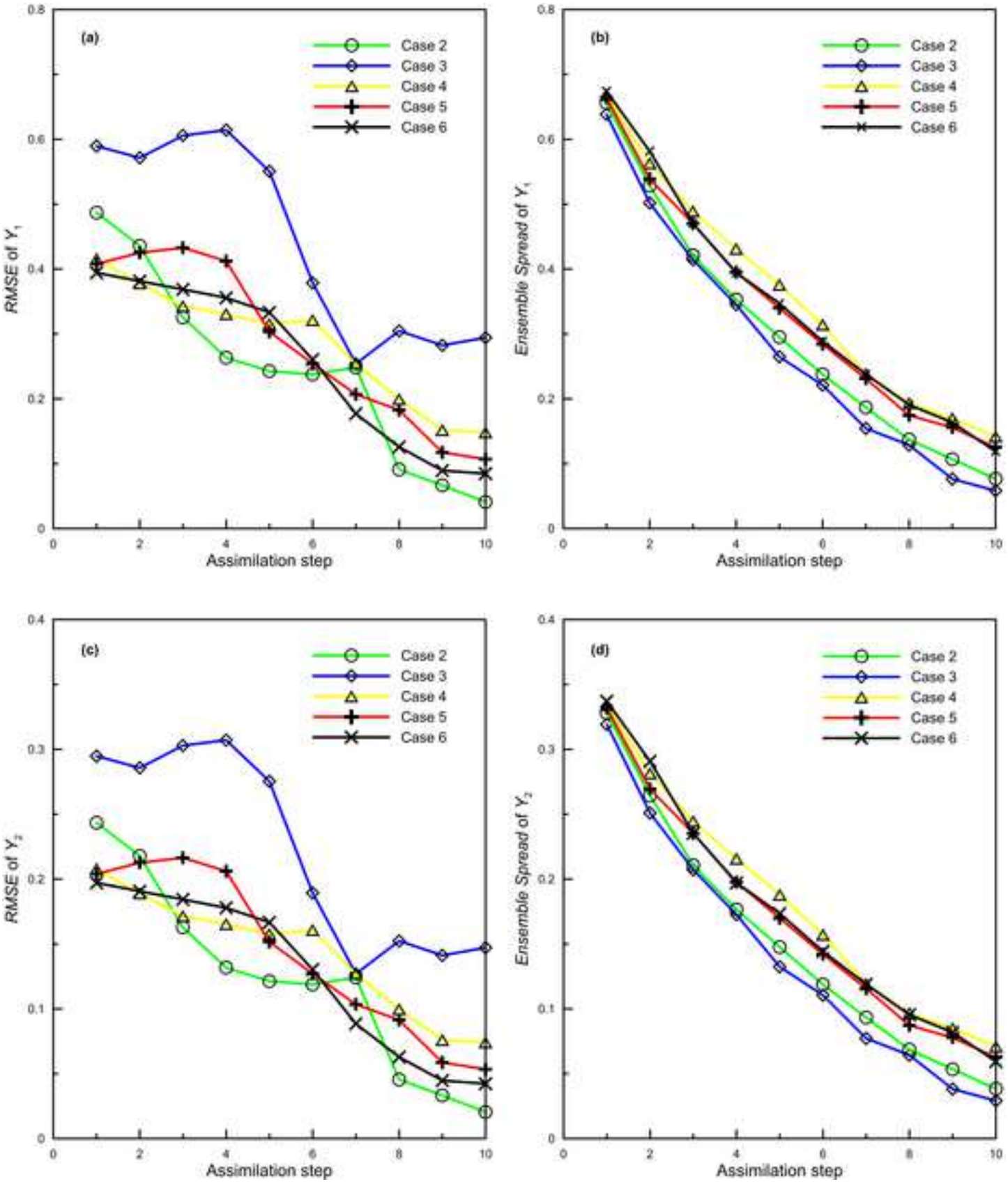
line figure

Click here to download line figure Fig 10.tif



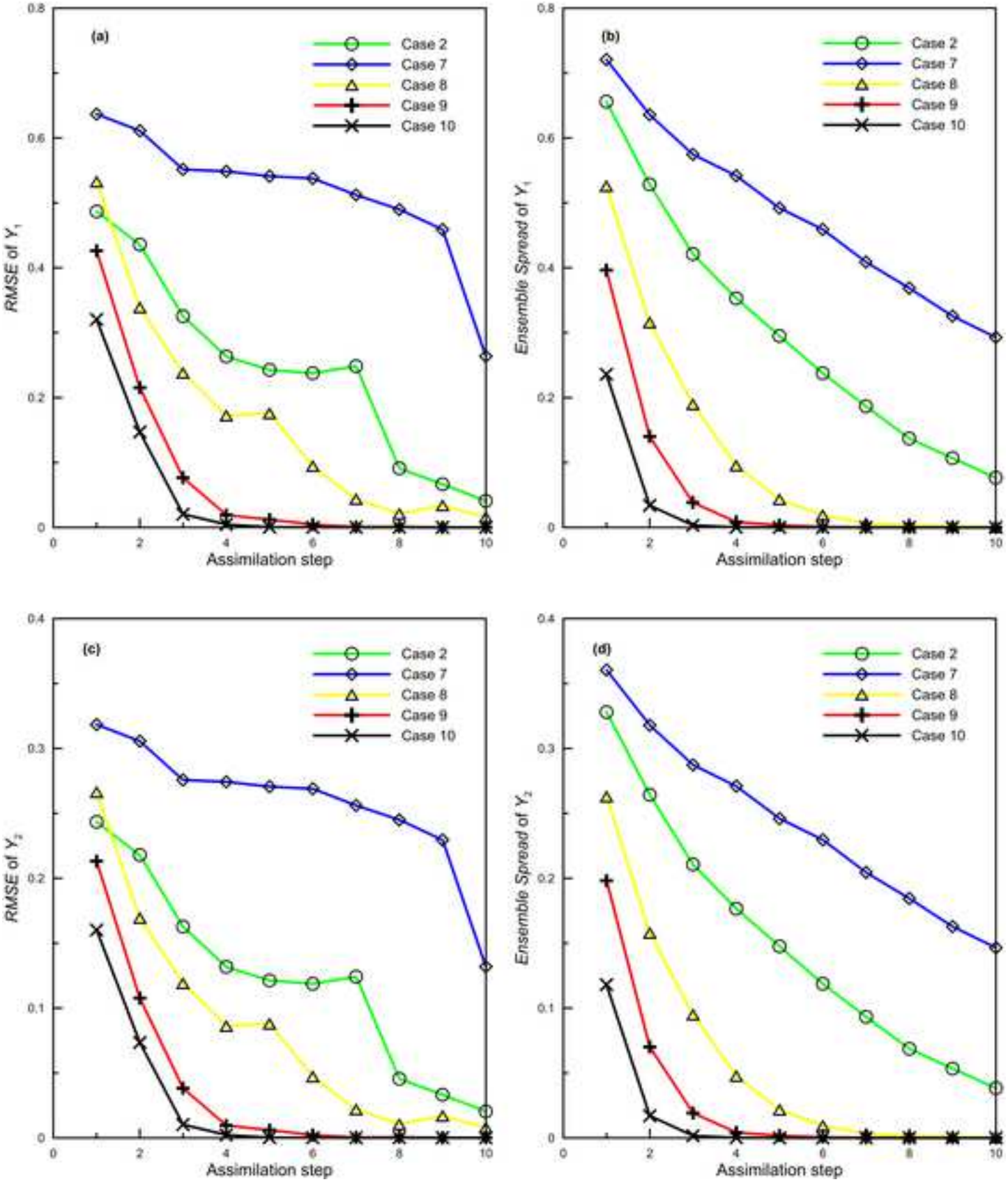
line figure

Click here to download line figure Fig 13.tif



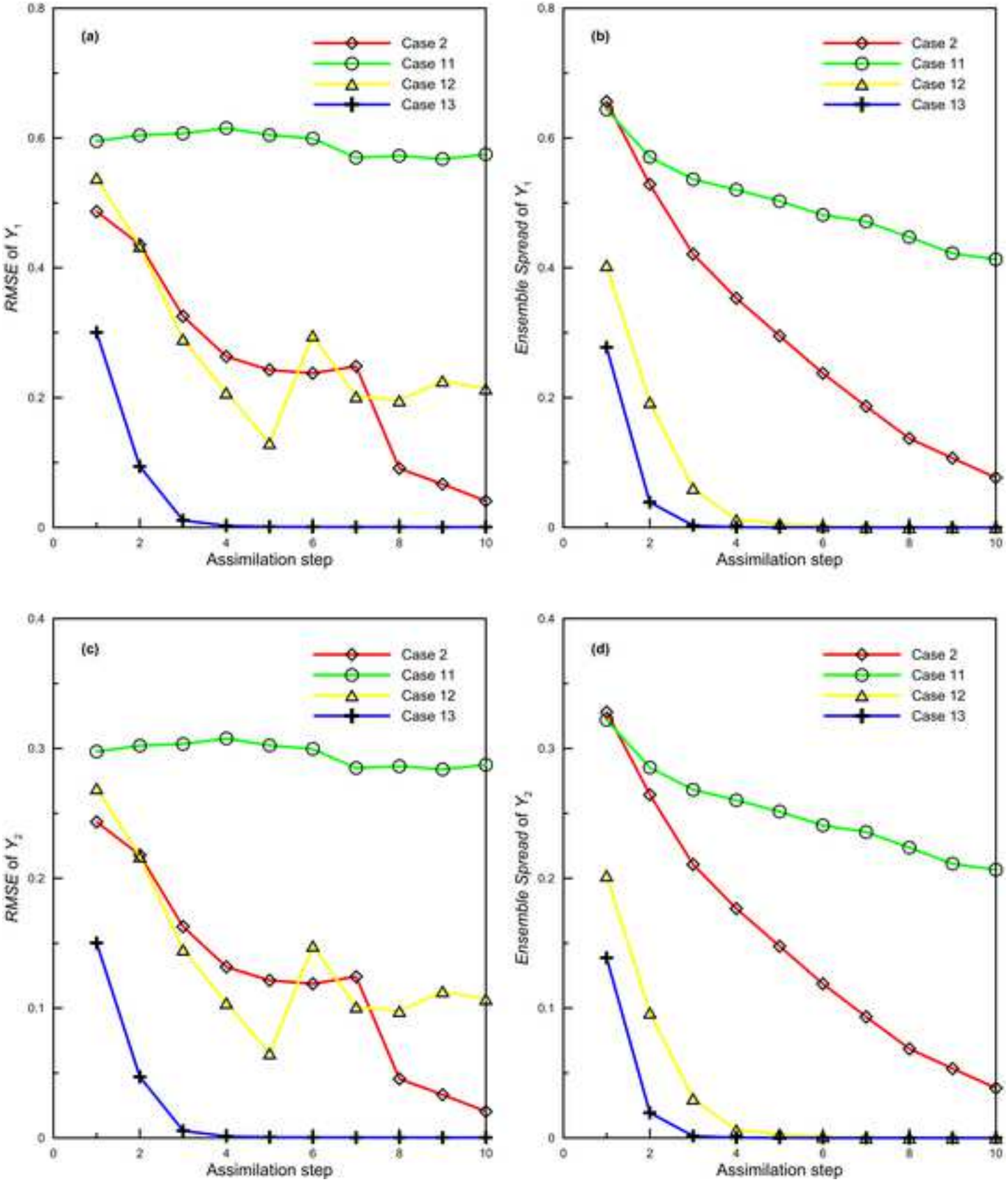
line figure

Click here to download line figure Fig 14.tif



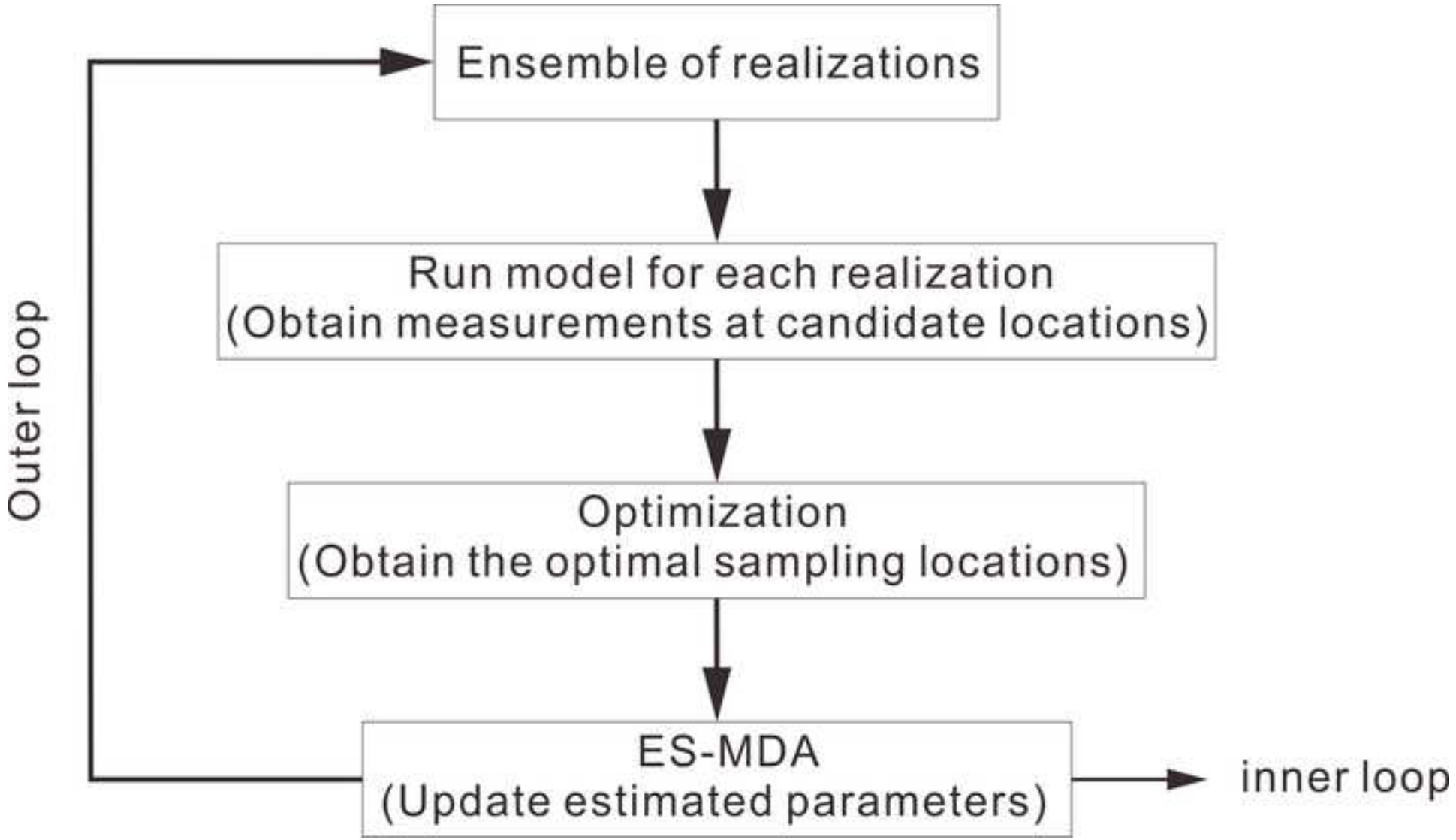
line figure

Click here to download line figure Fig 15.tif



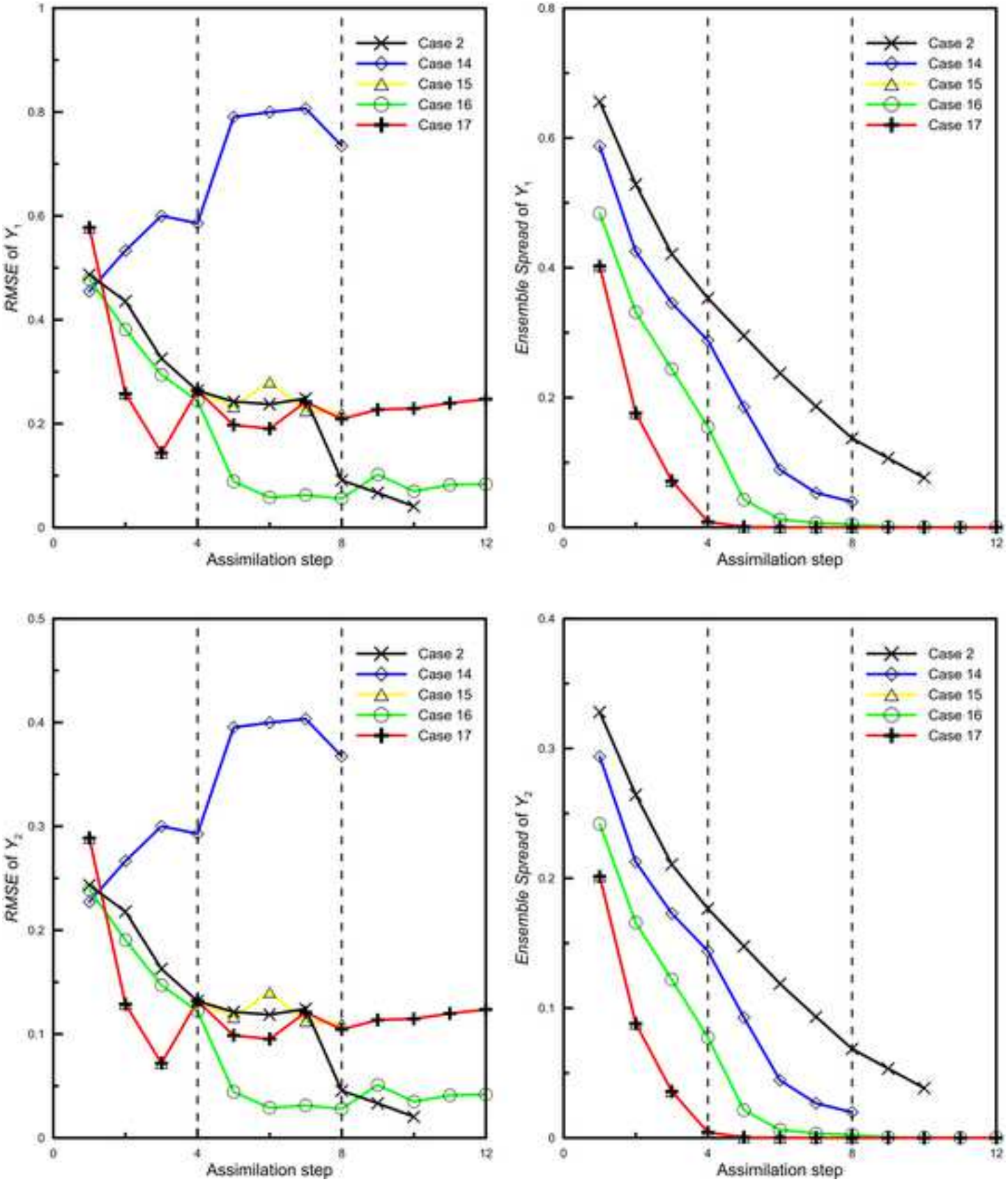
line figure

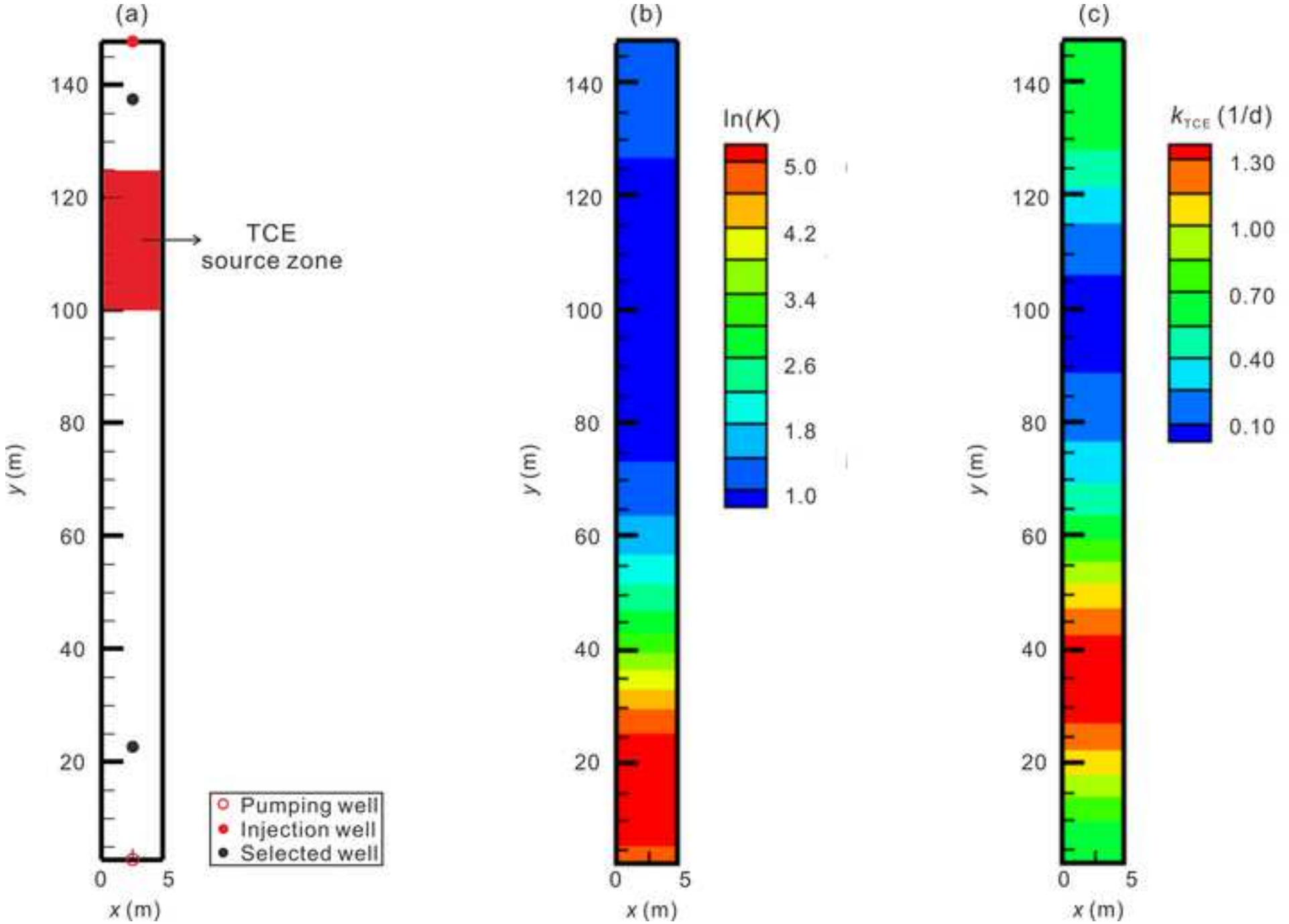
Click here to download line figure Fig 16.tif



line figure

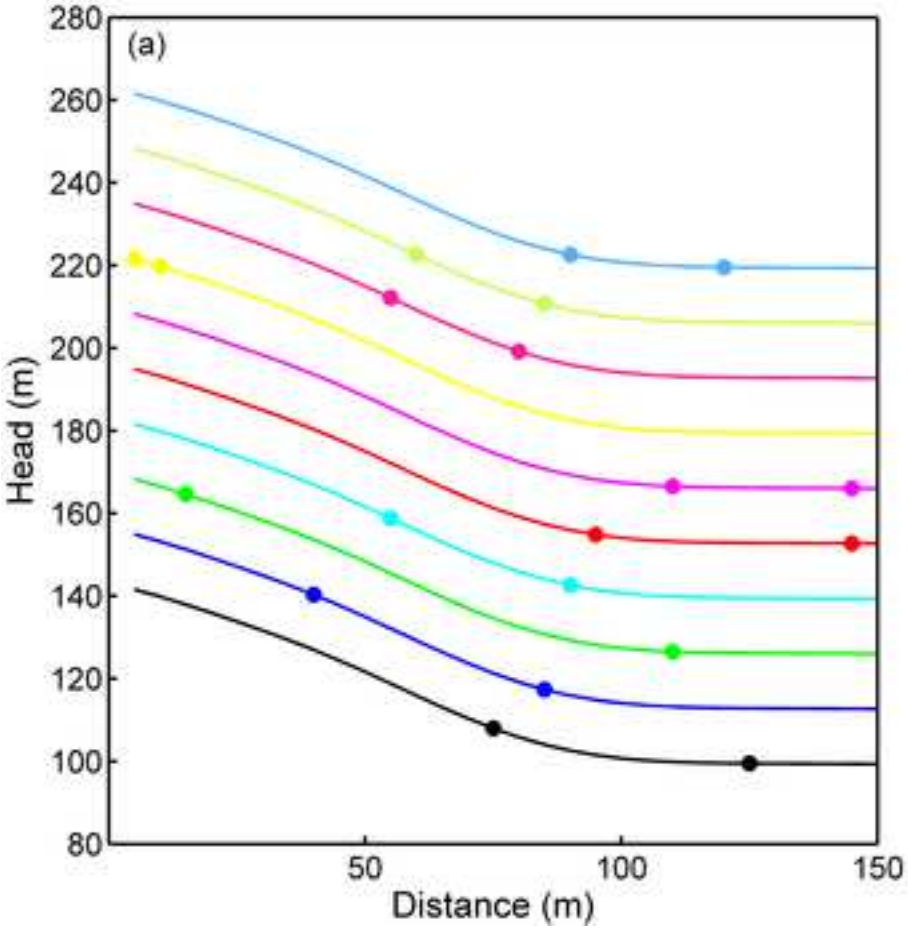
Click here to download line figure Fig 17.tif

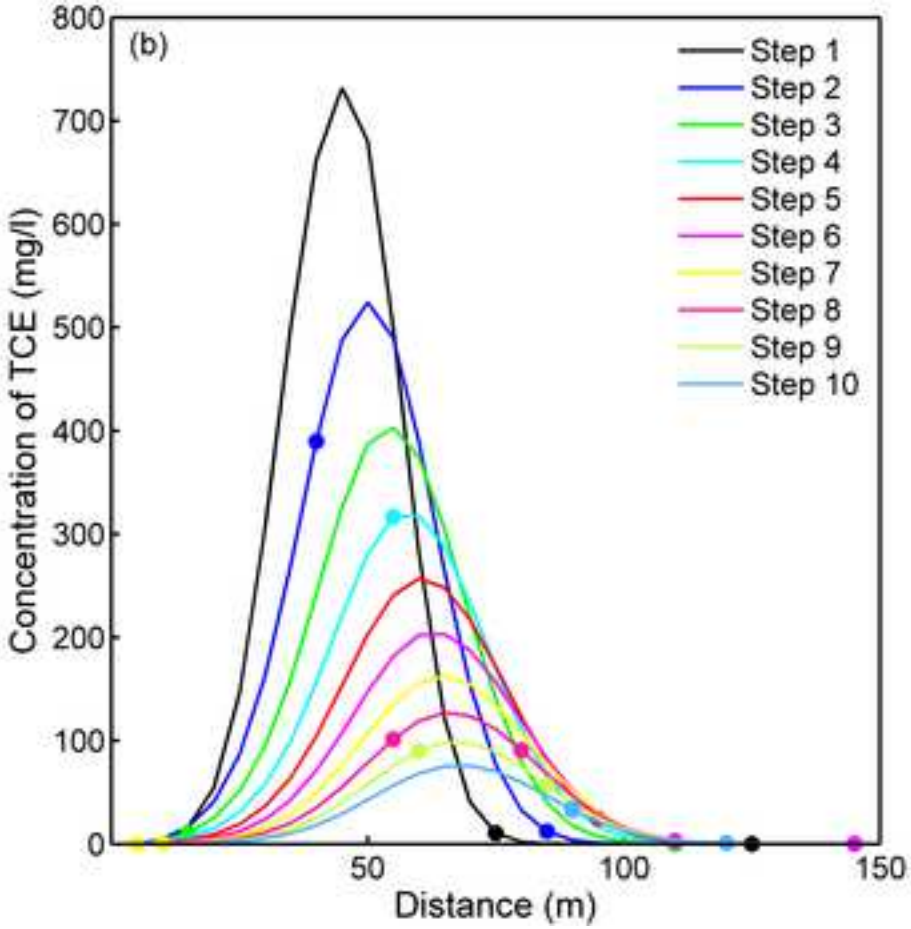




colour figure

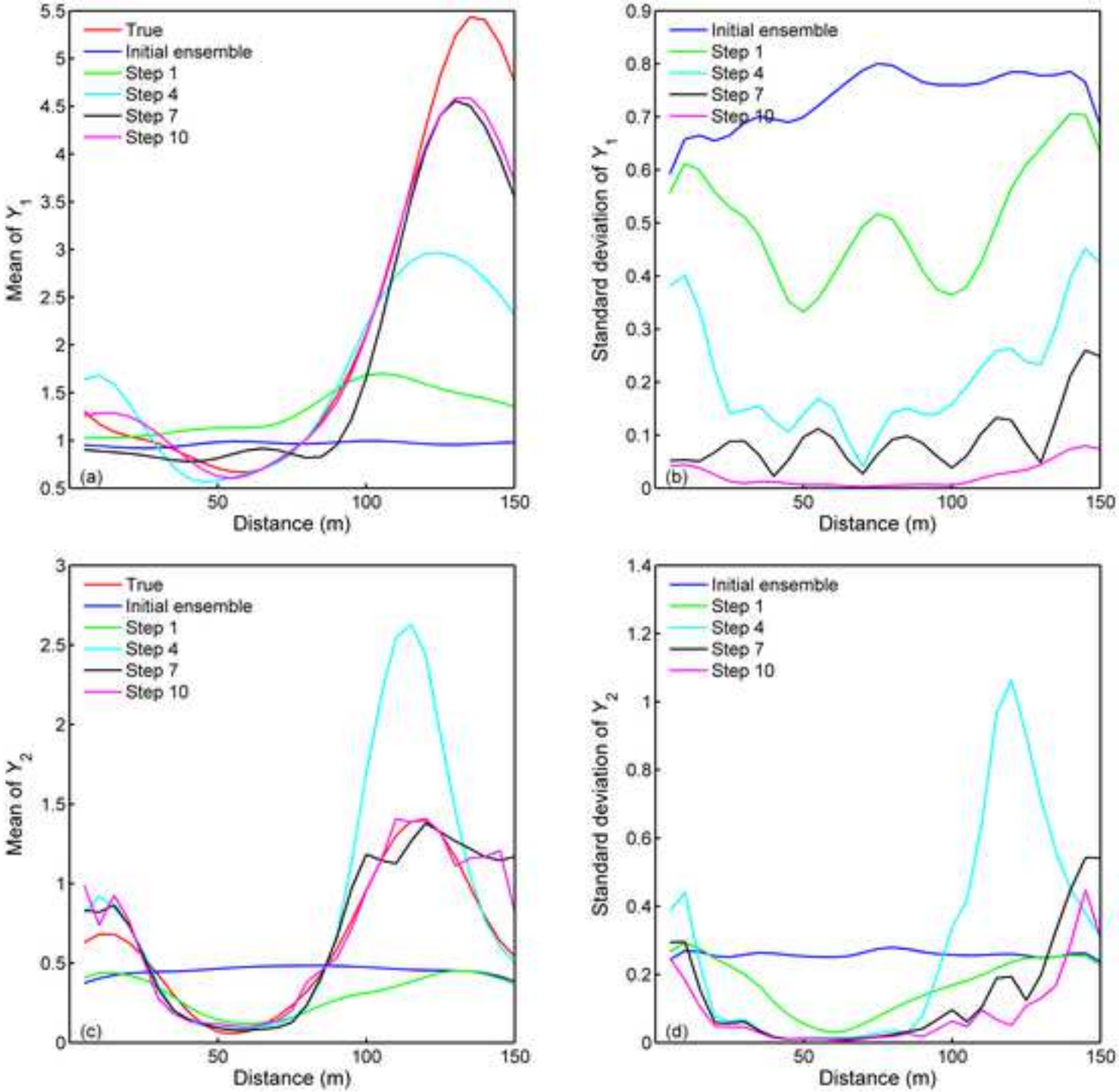
Click here to download colour figure Fig 2.tif





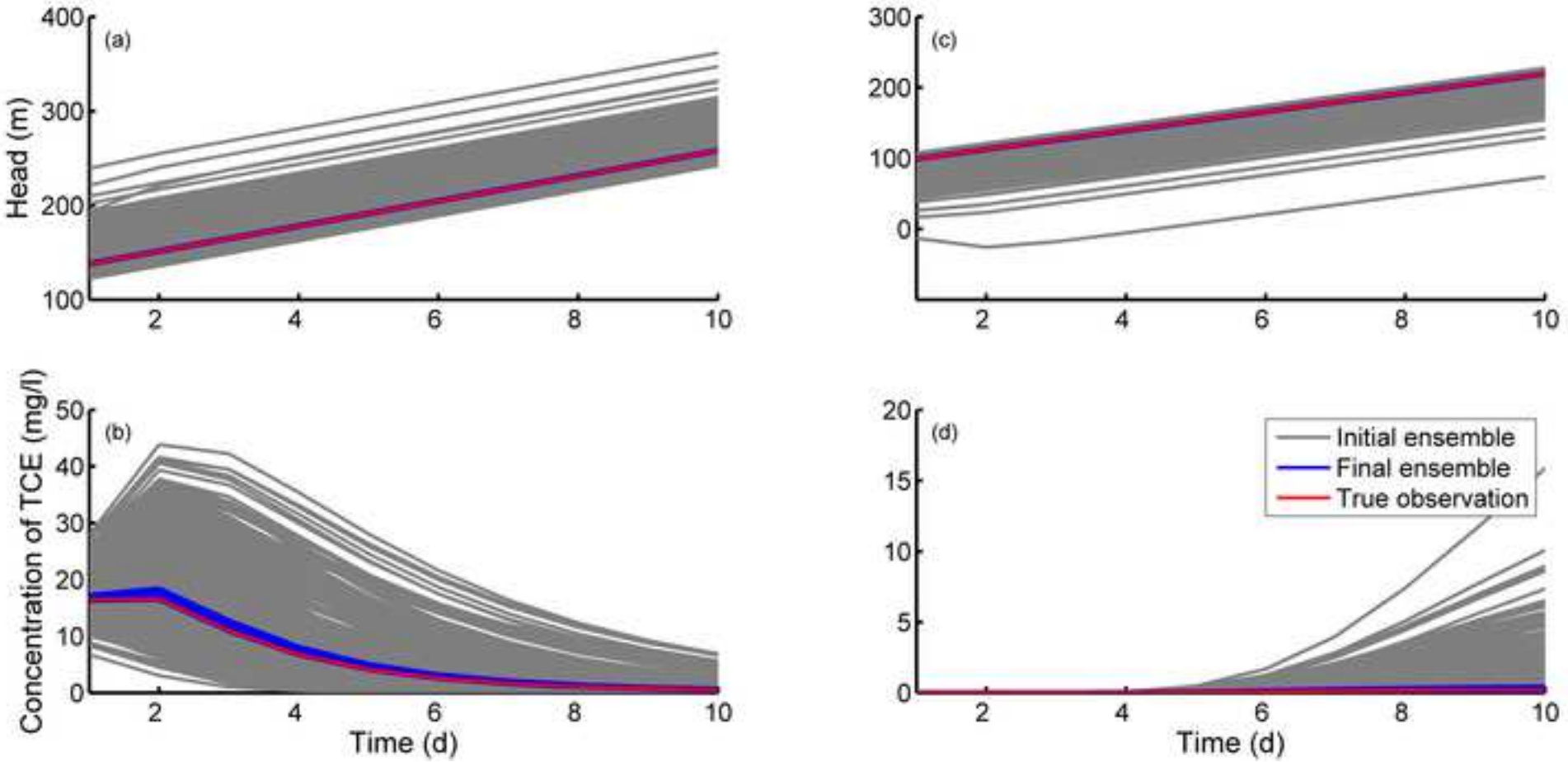
colour figure

Click here to download colour figure Fig 3.tif



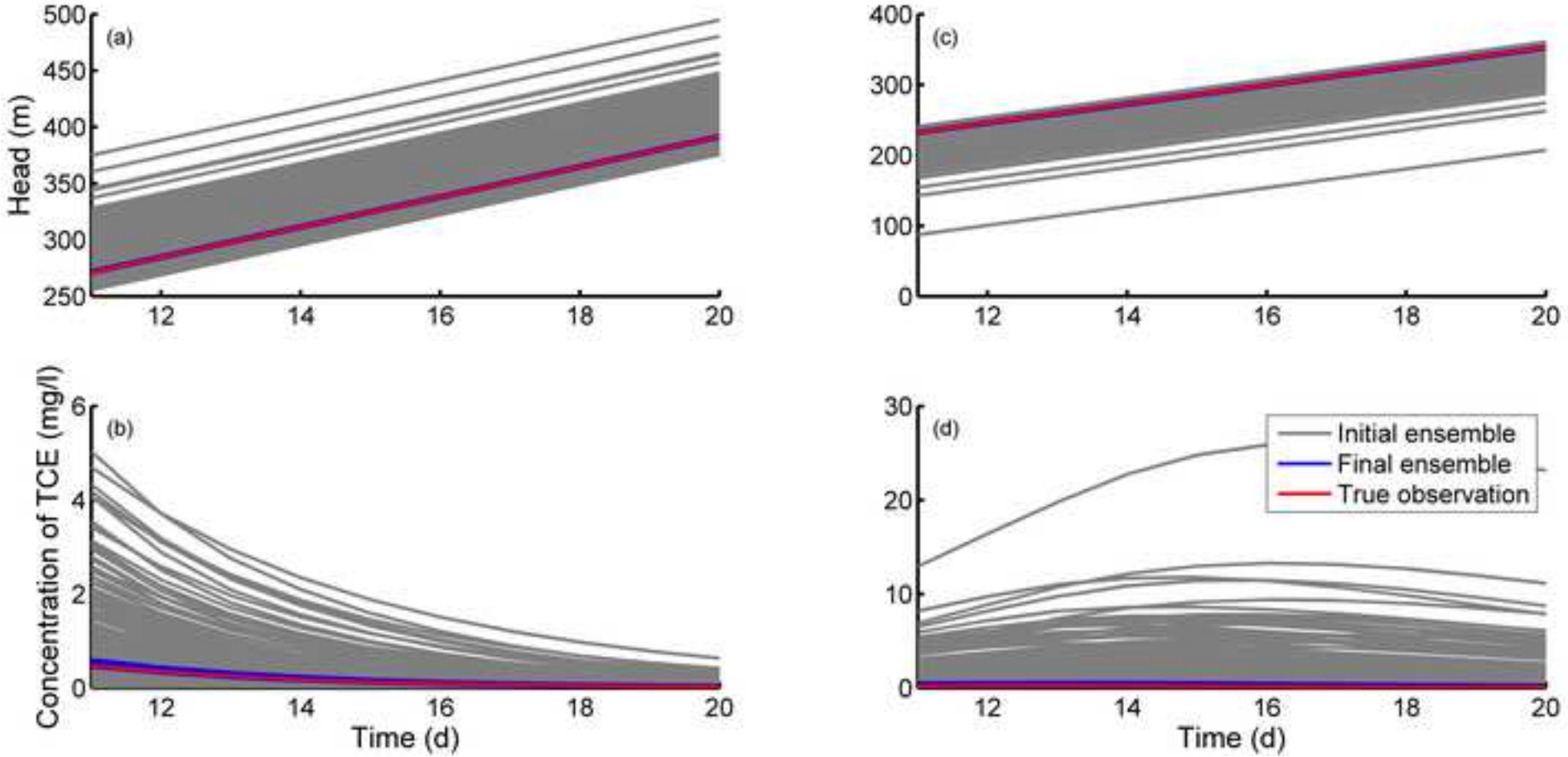
colour figure

Click here to download colour figure Fig 5.tif



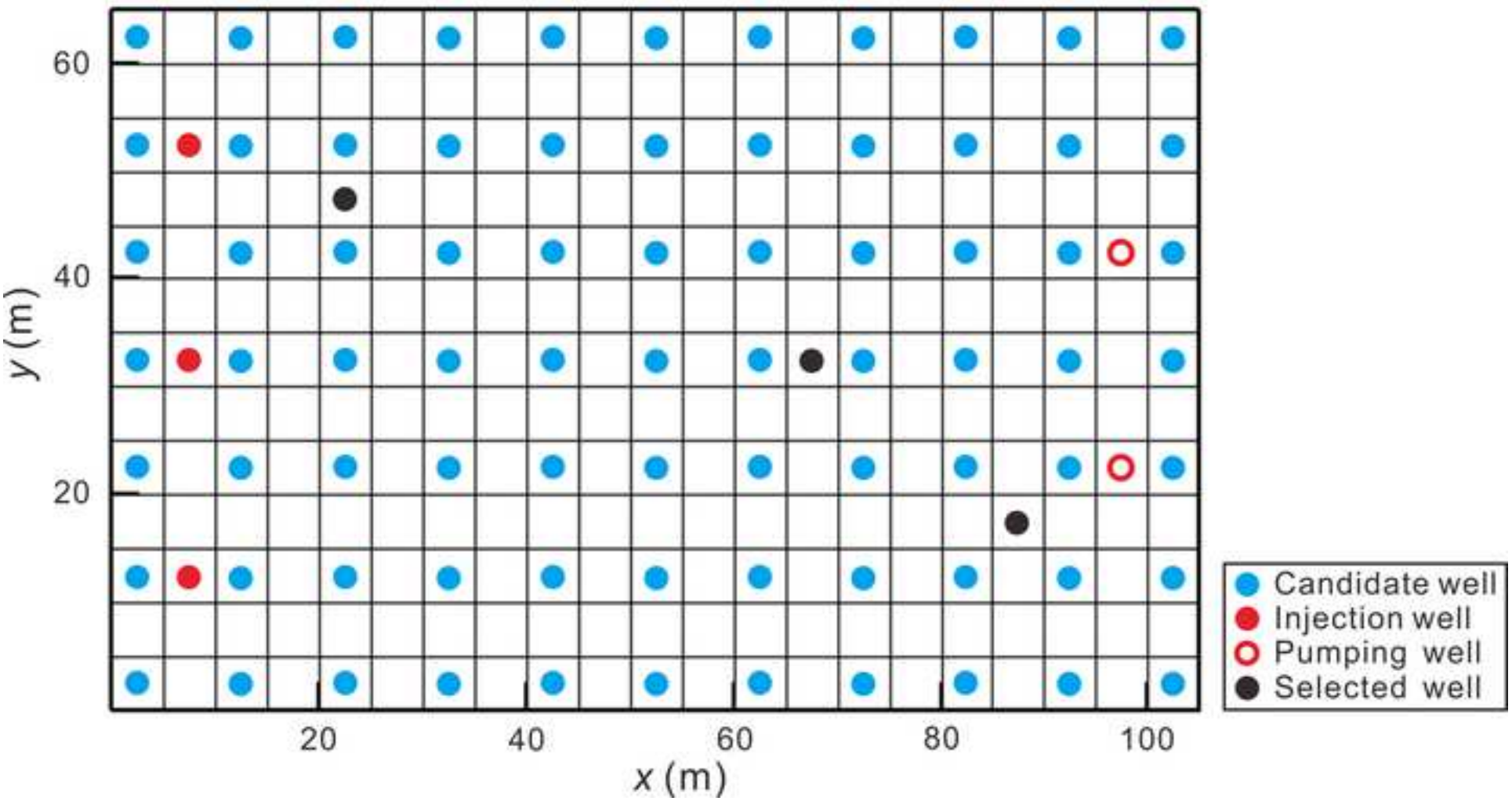
colour figure

Click here to download colour figure Fig 6.tif



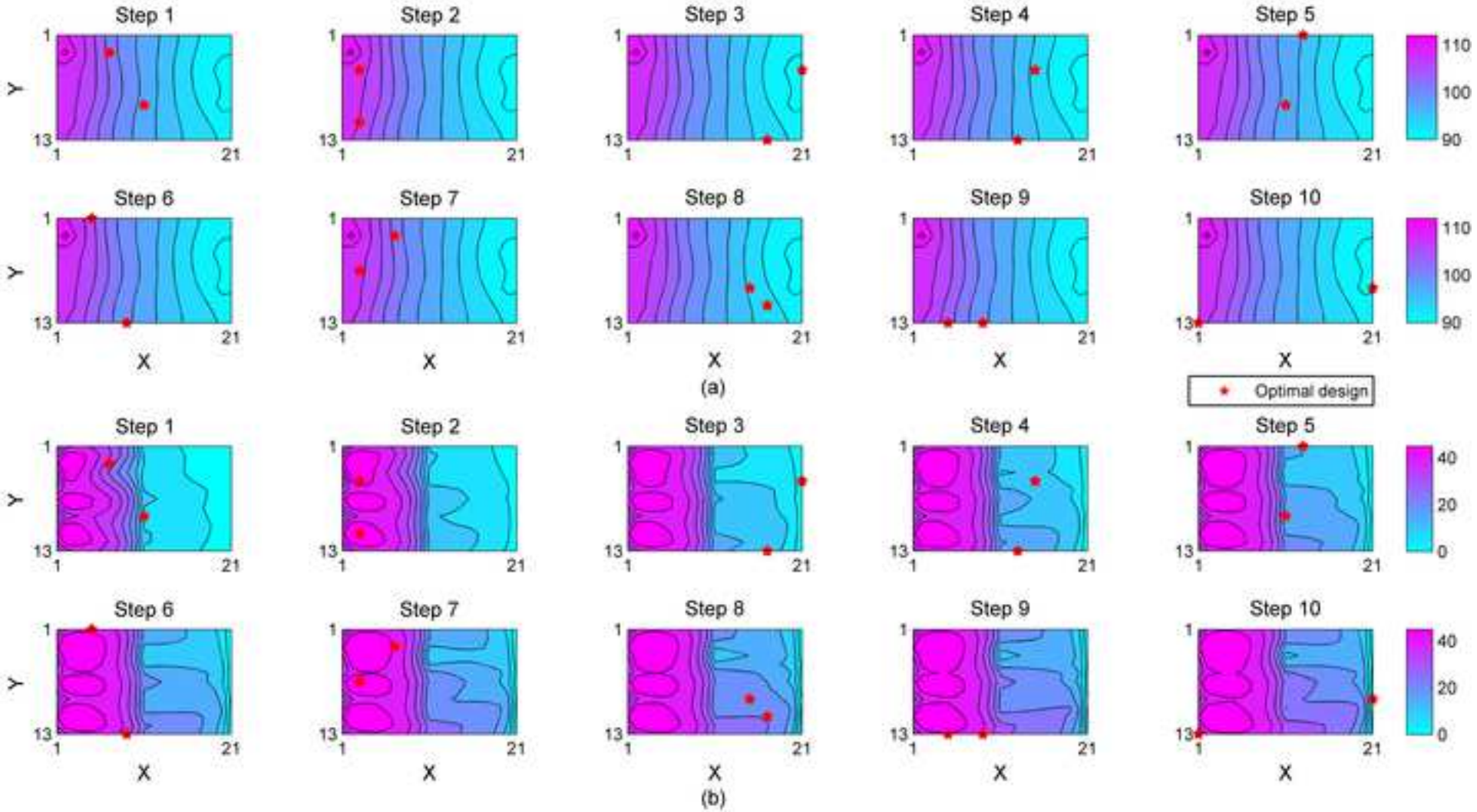
colour figure

Click here to download colour figure Fig 7.tif



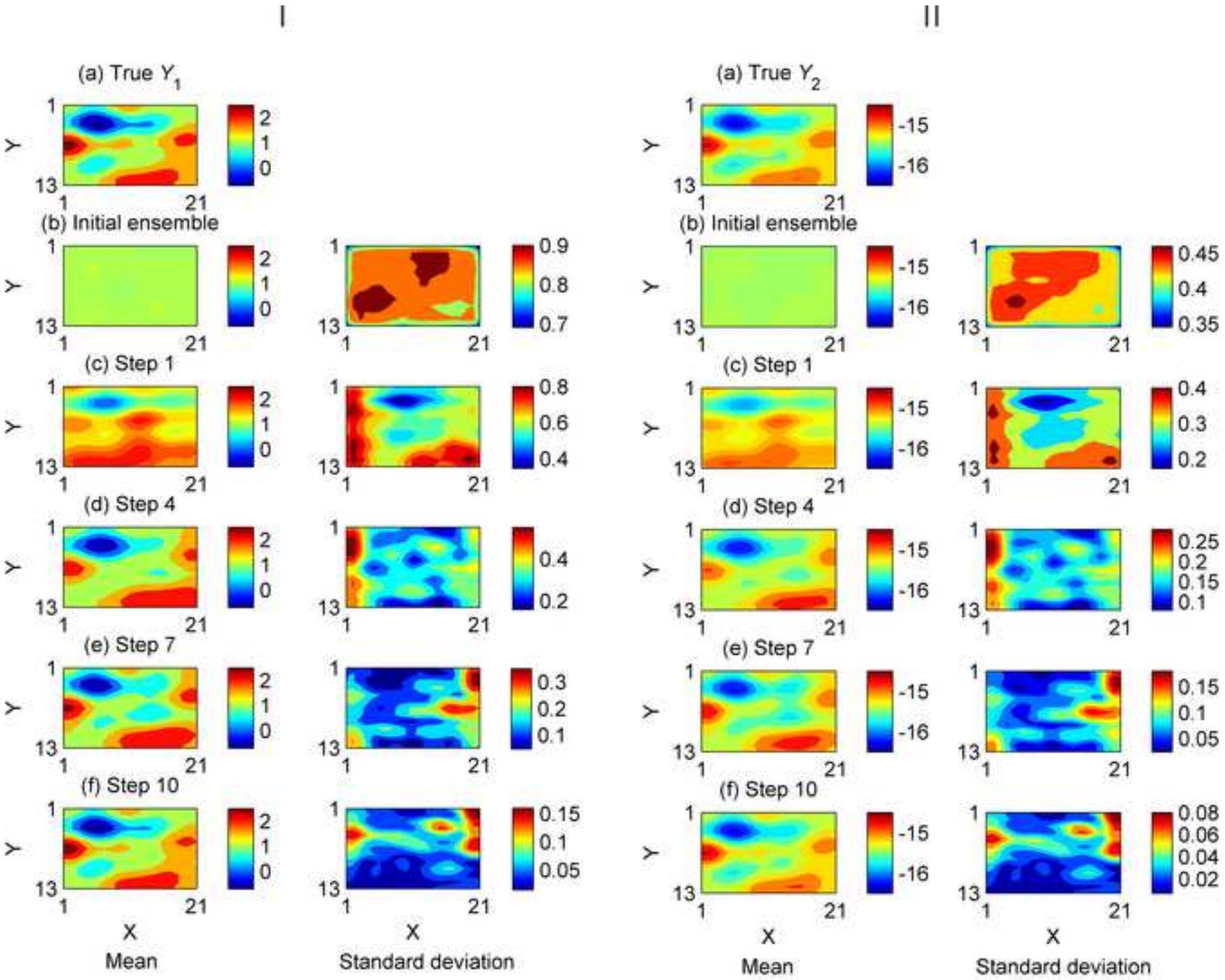
colour figure

Click here to download colour figure Fig 8.tif



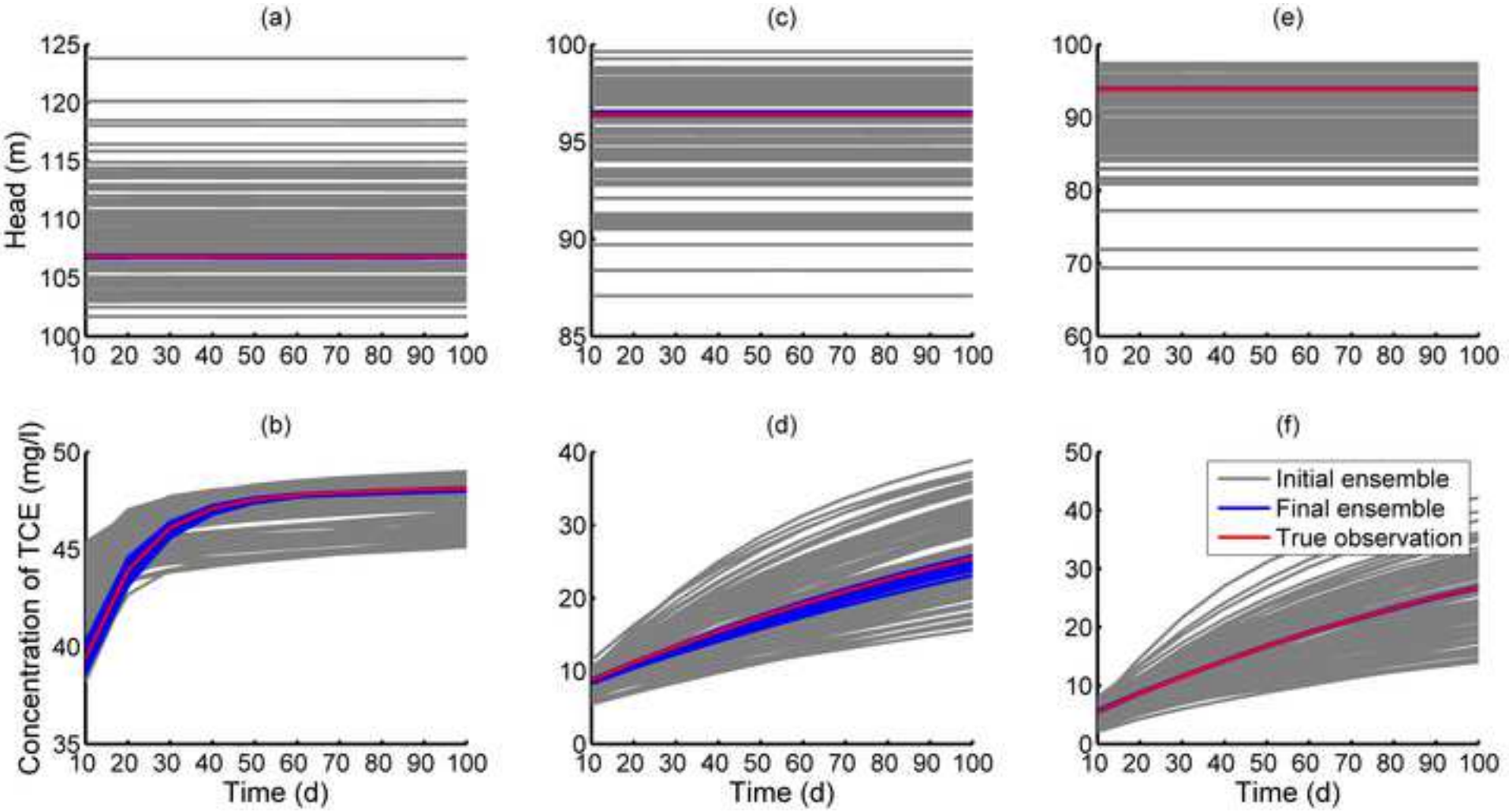
colour figure

Click here to download colour figure FIG 9.tif



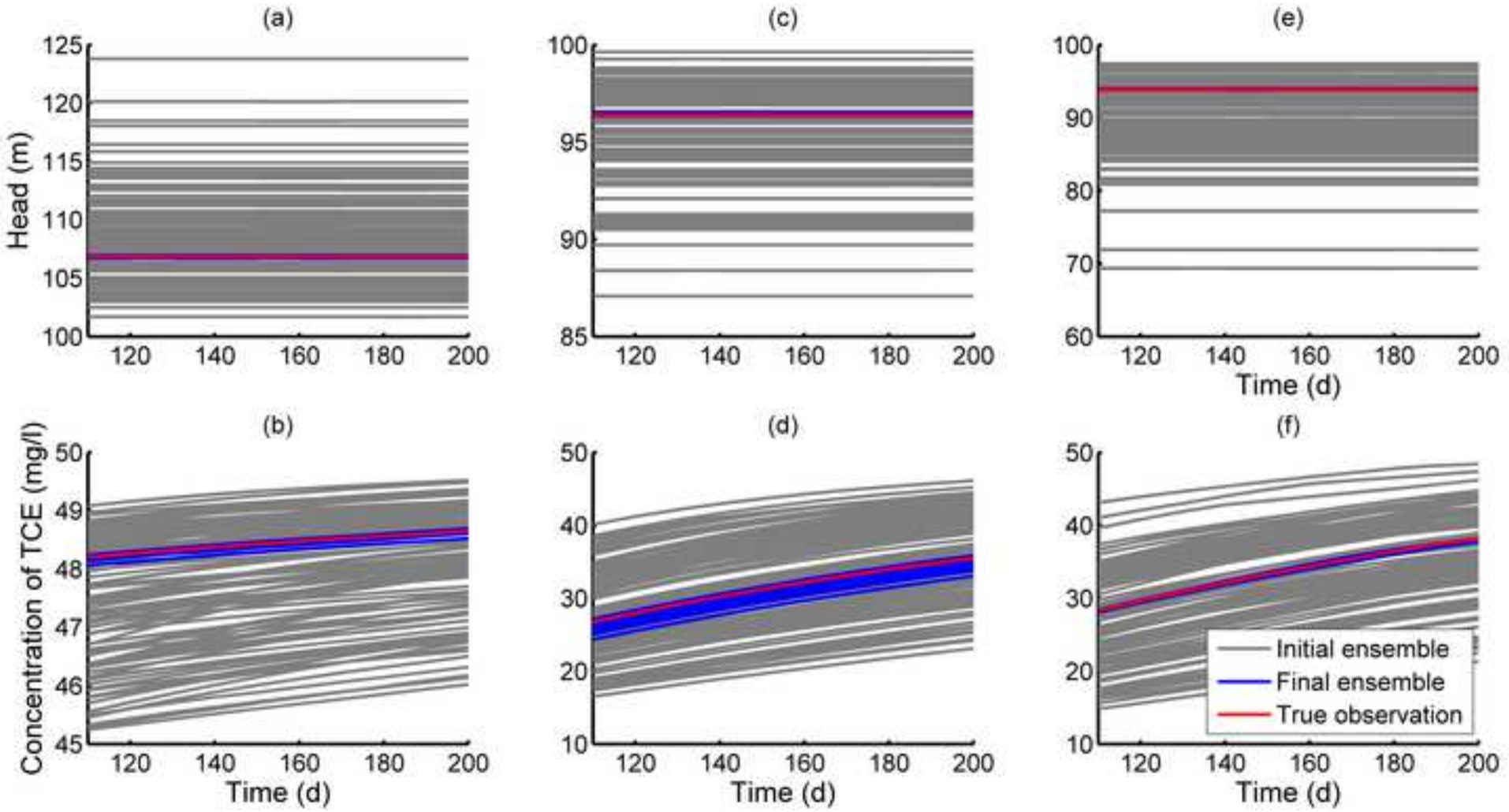
colour figure

Click here to download colour figure Fig 11.tif



colour figure

Click here to download colour figure Fig 12.tif



Tables

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/d)	50
Pumping rate per well (m^3/d)	45

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)*	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

* 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.

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/d)	80
Pumping rate per well (m^3/d)	120
TCE injection concentration per well (mg/L)	50

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