Journal of Hydrology 590 (2020) 125443

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Coupling ensemble smoother and deep learning with generative adversarial networks to deal with non-Gaussianity in flow and transport data assimilation

Jichao Bao, Liangping Li*, Fleford Redoloza

Department of Geology and Geological Engineering, South Dakota School of Mines and Technology, Rapid City 57701, USA

ARTICLE INFO

This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Vahid Nourani, Associate Editor

Keywords:
Generative adversarial networks
Ensemble smoother
Non-Gaussianity
Data assimilation
Deep learning

ABSTRACT

Dynamic data such as hydraulic head and concentration data can be integrated into the groundwater flow and contaminant transport model to improve its predictive ability for groundwater resource management and aquifer remediation. Ensemble Smoother with Multiple Data Assimilation (ES-MDA) has gained popularity for data assimilation in the field of hydrogeology, where aquifer parameters such as hydraulic conductivity are calibrated by conditioning on observed dynamic data. The ES-MDA has an optimal solution if aquifer parameters follow a multi-Gaussian distribution. However, fluvial deposits commonly exhibit a strong heterogeneity with channels (i.e., connectivity). In other words, the hydraulic conductivity does not follow the multi-Gaussian distribution. To deal with data assimilation in channelized aquifers, we propose to couple ES-MDA with deep learning. Specifically, Generative Adversarial Networks (GAN), a deep learning algorithm, are used to re-parameterize the channelized aquifer with a low-dimension latent variable. The ES-MDA is then used to update the latent variable by assimilating dynamic data into the groundwater model. Synthetic studies of groundwater flow and contaminant transport models are used to demonstrate the proposed method. The results illustrate that the coupling of GAN and ES-MDA is able to reconstruct the channel structures and reduce the uncertainty of hydraulic head and contaminant concentration predictions.

1. Introduction

Groundwater plays an important role in water supply in many places due to the scarcity and pollution of the groundwater. However, with the increase in water consumption, the decline of groundwater level and groundwater pollution seriously affect the substantiality of water supply. It is of great significance for groundwater management and aquifer remediation to understand the groundwater flow and contaminant transport processes. The main challenge in predicting groundwater flow and contaminant transport is the strong heterogeneity of the aquifer parameters. Data assimilation techniques provide an avenue to integrate dynamic data such as hydraulic head and concentration into the groundwater flow and contaminant transport model to estimate those aquifer parameters and reduce its uncertainty. Therefore the updated parameters can be used to improve the predictive ability of the flow and transport model. In particular, the Ensemble Kalman Filter (EnKF) proposed by Evensen (1994) is one of the most efficient data assimilation methods and has been successfully applied in many fields such as hydrogeology and petroleum engineering. For example, Chen and Zhang (2006) implemented the EnKF to update the hydraulic conductivity by assimilating hydraulic head data, and the influence of the ensemble size and observation timings on the results were also evaluated. Liu et al. (2008) applied the EnKF to investigate the flow and transport process at the Macro-Dispersion Experiment (MADE) site by simultaneously estimating multiple parameters such as hydraulic conductivities and dispersivities, and the results shown that the EnKF can be used to solve large-scale and non-linear inverse problems. Li et al. (2012) applied the EnKF in a transient groundwater flow and transport model to jointly estimate the hydraulic conductivity and porosity using hydraulic head and concentration data, and the results demonstrated that different types of data would help characterize aquifer parameters and improve the flow and contaminant transport predictions.

One drawback of the EnKF is that the recursive updates require restarting the simulation at each time step, which makes it time-consuming. An alternative method to the EnKF is the Ensemble Smoother (ES), which was proposed by Van and Evensen (1996). In this method, all the observation data are assimilated simultaneously, which greatly

---

* Corresponding author.
E-mail address: Liangping.Li@sdsmt.edu (L. Li).

https://doi.org/10.1016/j.jhydrol.2020.125443
Received 4 May 2020; Received in revised form 17 July 2020; Accepted 19 August 2020

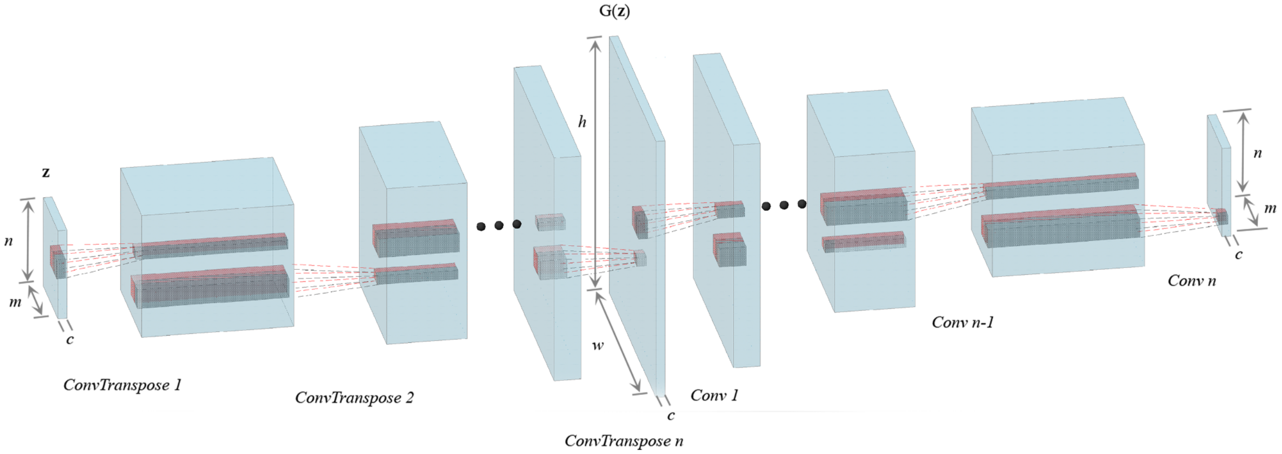


Fig. 1. The stucture of GAN. **z** represents the latent space variables, $G(\mathbf{z})$ indicates the image generated by the generator, *ConvTranspose* is the transposed convolutional layer, and *Conv* represents the convolutional layer.

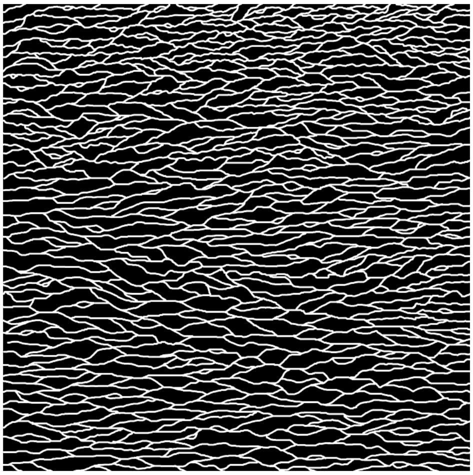


Fig. 2. Training image.

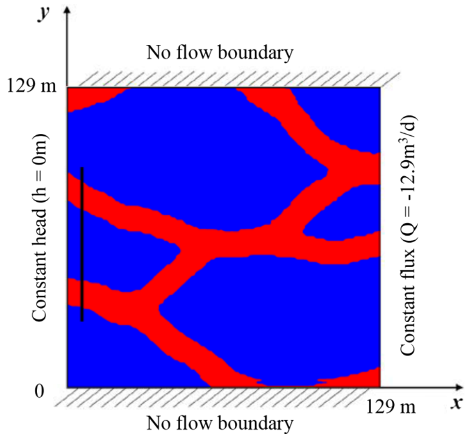


Fig. 3. The reference hydraulic conductivities with boundary conditions. No flow boundary on north and south, constant head on the west boundary and constant flux on the east boundary. The red areas represent the channels with high hydraulic conductivity and the blue background indicates low hydraulic conductivity materials. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reduces the computational burden. Bailey and Baù (2010) applied the ES to estimate the hydraulic conductivity by assimilating the hydraulic head and groundwater return flow volume. Skjervheim et al. (2011) employed the ES for history matching in petroleum engineering, and the results indicated that the ES was able to significantly reduce the simulation time and provide a flexible parameterization solution. However, the ES only assimilates data once to achieve a globe update, which might not be able to obtain a promising result that honors all the observation data. Consequently, a series of iterative ensemble methods have been proposed, for example, the iterative EnKF (Gu and Oliver, 2007) and the iterative ES (Chen and Oliver, 2012). Emerick and Reynolds (2013) proposed a new form of the Ensemble Soother with Multiple Data Assimilation (ES-MDA), in which the same measured data can be assimilated multiple times to reach a better result. Zhao et al. (2017) coupled the ES-MDA with the common basis Discrete Cosine Transform (DCT) to solve the history matching problem of the multi-facies channelized reservoirs, and this study illustrated that the proposed method was able to preserve the key geological features of prior

Table 1
The architectures of generator and discriminator.

Generator	Discriminator
Input: 5 × 5 latent space **z**	Input: 129 × 129 image
Layer 1: ConvT2d, 5 k 2s 2p 1d, InsNorm2d, ReLU	Layer 1: Conv2d, 5 k 2s 2p 1d, InsNorm2d, LeakyReLU
Layer 2: ConvT2d, 5 k 2s 2p 1d, InsNorm2d, ReLU	Layer 2: Conv2d, 5 k 2s 2p 1d, InsNorm2d, LeakyReLU
Layer 3: ConvT2d, 5 k 2s 2p 1d, InsNorm2d, ReLU	Layer 3: Conv2d, 5 k 2s 2p 1d, InsNorm2d, LeakyReLU
Layer 4: ConvT2d, 5 k 2s 2p 1d, InsNorm2d, ReLU	Layer 4: Conv2d, 5 k 2s 2p 1d, InsNorm2d, LeakyReLU
Layer 5: ConvT2d, 5 k 2s 2p 1d, InsNorm2d, ReLU	Layer 5: Conv2d, 5 k 2s 2p 1d, Sigmoid
Layer 6: ConvT2d, 5 k 1s 6p 3d, InsNorm2d, ReLU	Output: 5 × 5 array
Layer 7: ConvT2d, 5 k 1s 10p 5d, Tanh	
Output: 129 × 129 image	

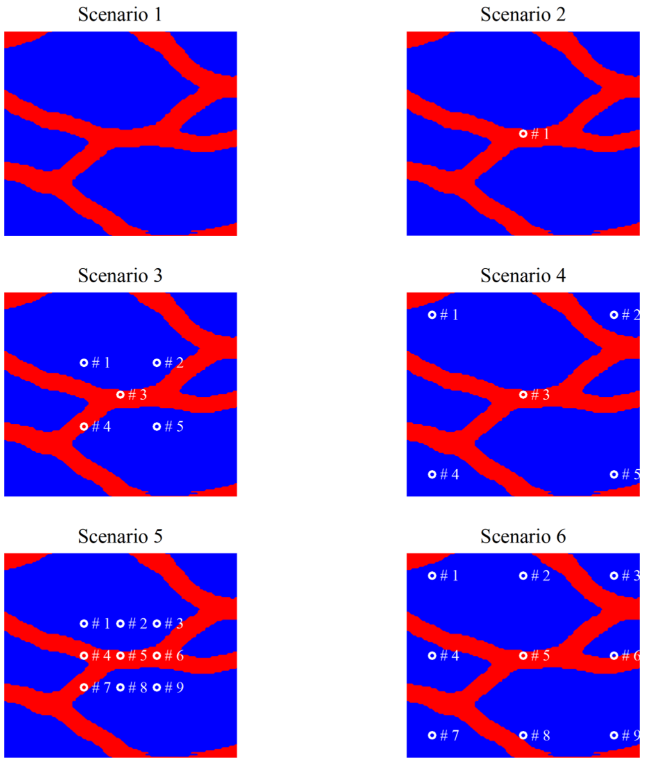


Fig. 4. Well locations for 6 Scenarios. The circles represent the well locations.

Table 2
Scenario studies.

Scenario	Ensemble size	Conditioned heads	Observation wells
1	100	×	0
2	100	✓	1
3	100	✓	5
4	100	✓	5 (Larger spacing)
5	100	✓	9
6	100	✓	9 (Larger spacing)

models and reduce the uncertainty of reservoir description. Fokker et al. (2016) applied ES-MDA to investigate the gas reservoir by assimilating the line-of-sight radar measurements, and the results demonstrated that ES-MDA could significantly reduce the uncertainty of parameter estimation. Evensen (2018) compared IES and ES-MDA with standard ES, based on Bayes' theorem, and this study revealed that these three methods could produce the same result for linear models, but iterative methods behave better in nonlinear situations. Also, increasing the assimilation times will lead to a better result for ES-MDA.

Although ES-MDA is an efficient method for data assimilation, the same problem still exists as the EnKF method. In other words, it has an optimal solution if parameters and state variables follow a multi-Gaussian distribution (Zhou et al., 2014). Many parameters and state variables often do not have a multi-Gaussian distribution such as hydraulic conductivities in fluvial deposits. Inverse methods have been proposed to deal with non-Gaussianity, such as the Markov Chain Monte Carlo method (MCMC) (Oliver et al., 1997) and Gradual Deformation Method (GDM) (Hu, 2000). Nevertheless, their drawbacks might limit the implementation of these methods. For instance, the MCMC method is computationally expensive because of a large number of evaluation for the forward modeling (Kuczera et al., 2010). The structure preservation capability of the GDM method is strongly influenced by the number of realization chains (Le and Nœtinger, 2002). In addition, if the prior is wrong, a reasonable structure might not be able to be retrieved.

Attempts based on ES-MDA have been proposed to deal with the issue of non-Gaussianity in data assimilation. Le et al. (2015) coupled ES-MDA with multi-point statistics simulation (MPS) to estimate non-Gaussian facies. In this study, ES-MDA was applied to update the permeability, and then the reservoir facies was regenerated by MPS, using the average of the updated permeability through ensemble as soft data. The results illustrated that the non-Gaussian facies could be reconstructed by the proposed procedure. Zhao et al. (2017) combined ES-MDA with a common basis discrete cosine transform and a regularization post-processing technique to deal with the three facies history matching problem. The results indicated that the proposed approach could significantly reduce the uncertainty of reservoir characterization and preserve the key geological features. Kim et al. (2018) proposed a combination of ES-MDA, DCT and iterative K-singular value decomposition (K-SVD) to estimate the permeability in a channelized aquifer using synthetic examples, and the proposed algorithm could capture the complex geological features of the gas reservoirs and provide plausible models to match the measured gas rate and water rate.

In recent years, the successful applications of deep learning techniques in many fields have proven its ability to handle multiple scales and heterogeneous information (Marçais and de Dreuzy, 2017), which provides an opportunity to cope with complex problems such as non-Gaussianity in hydrogeological sciences. Mo et al. (2019) proposed an autoregressive framework coupling convolutional encoder-decoder network with iterative local updating ensemble smoother (ILUES) (Zhang et al., 2018) to address the high-dimensional contaminant transport inversion problem. Their results showed that this method is able to identify the contaminant source and provide an accurate hydraulic conductivity field. Zhang et al. (2020) introduced a strategy to use the deep learning technique to modify the updating process of ensemble smoother, and the results illustrated that the proposed method performed much better than the standard ES method in non-Gaussian aquifers.

Deep learning with Generative Adversarial Networks (GAN), which was introduced by Goodfellow et al. (2014), has been gaining popularity in recent years. A variety of GANs have been proposed (e.g., Mirza and Osindero, 2014; Radford et al., 2015; Arjovsky et al., 2017). For instance, Laloy et al. (2018) proposed a spatial GAN (SGAN) to address the high dimensional inversion problems, and their work showed that the inversion framework which combined SGAN and MCMC could recover the complex geological features close to the true models. Sun (2018) presented a state-parameter identification GAN (SPID-GAN) which contains two pairs of generators and discriminators, the results demonstrated that GAN performed well in identifying the state-parameter bidirectional mappings. Janssens et al. (2020) applied GAN to improve the resolution of Computed Tomography (CT) images, and using the improved CT images as the input could lead to a better result of fluid flow simulation and more realistic pore distribution.

In this paper, we propose to combine ES-MDA and deep learning with GAN to deal with non-Gaussianity in flow and transport data assimilation. Specifically, GAN will be used to reparameterize the high dimensional non-Gaussian distributed hydraulic conductivity using a low dimension latent variable. ES-MDA is then used to update the variable in the latent space by assimilating dynamic data into groundwater models. Besides the hydraulic head data that are commonly used to characterize hydraulic conductivities, concentration data also are jointly employed to identify the non-Gaussian parameters in this study. The results demonstrate that the coupling of GAN and ES-MDA can reconstruct the channel structures and reduce the uncertainty of simulated hydraulic heads and contaminant plume. The rest of this paper is organized as follows: the methods are explained in Section 2. Several

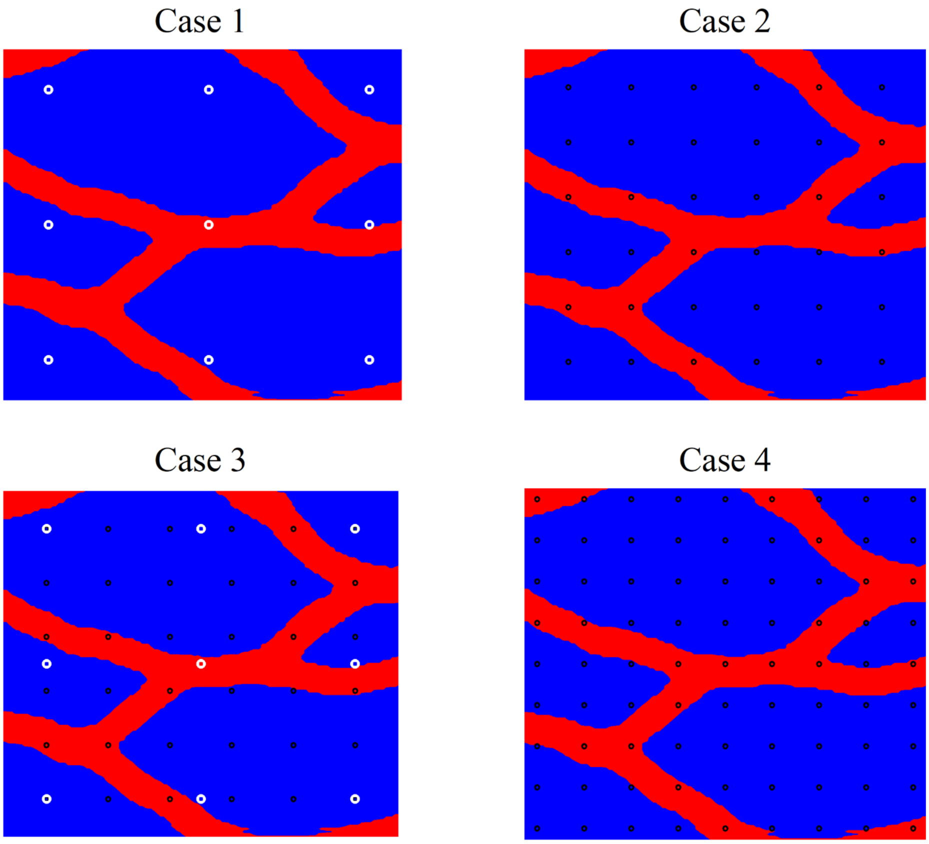


Fig. 5. The locations of observation wells. The hydraulic head observation wells are shown in white circles, and the concentration observation wells are shown in black circles.

Table 3
Solute transport.

Case	Ensemble size	Conditioned heads	Conditioned concentrations
1	100	9	0
2	100	0	36
3	100	9	36
4	100	0	81

synthetic experiments are conducted in Section 3. Then, analyses and discussions about the results are shown in Section 4. Finally, conclusions are presented in Section 5.

2. Methodology

2.1. Forward modeling

The single phase groundwater flow equation can be expressed as follows (Fetter, 2018):

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_{zz}\frac{\partial h}{\partial z}\right)+W=S_s\frac{\partial h}{\partial t} \tag{1}$$

where K_{xx}, K_{yy}, K_{zz} are the hydraulic conductivities along the x, y, and z directions; h is the hydraulic head; W represents source and sink; S_s is the specific storage; x, y, and z represent the coordinates; t is time. This equation is solved using the finite-difference method in which the aquifer is divided into a number of cells and the head is calculated at each node (Harbaugh et al., 2000).

The governing equation for solute transport can be described as (Huang et al., 1998; Zheng and Wang, 1999):

$$\begin{aligned}\frac{\partial M}{\partial t}= & \frac{\partial}{\partial x}\left(\theta D_x\frac{\partial C^k}{\partial x}\right)+\frac{\partial}{\partial y}\left(\theta D_y\frac{\partial C^k}{\partial y}\right)+\frac{\partial}{\partial z}\left(\theta D_z\frac{\partial C^k}{\partial z}\right)\\ & -\frac{\partial}{\partial x}\left(q_x C^k\right)-\frac{\partial}{\partial y}\left(q_y C^k\right)-\frac{\partial}{\partial z}\left(q_z C^k\right)+q_s C_s^k+\sum R_n\end{aligned} \tag{2}$$

where M is the solute content; t is time; θ is the porosity of the medium; C^k is the dissolved concentration of species k;D_x, D_y, and D_z are the hydrodynamic dispersion coefficients along the x, y, and z directions; q_x, q_y, and q_z are the volumetric flow rates per unit of aquifer in three directions; q_s is the volumetric flow rate per unit volume of aquifer representing fluid source and sinks; C_s^k is the concentration of the source or sink flux for species k;$\sum R_n$ is the chemical reaction term.

2.2. Generative adversarial networks

The structure of GAN is illustrated in Fig. 1, and the latent space $\mathbf{z}$ is sampled from a uniform distribution ($\mathbf{z}\sim U(-1, 1)$) as the input of the generator. Then through a series of transposed convolutional layers, the output image $G(\mathbf{z})$ is generated. The generated image and a training image are then sent to the discriminator. The discriminator consists of several convolutional layers, and the output arrays will be labeled as true or false. The task of the discriminator is to look at a given image and return true if the image is sampled from the training image (the image is “real”), or return false if the image was produced by the generator (the image is “fake”). Based on the images of the generator and the judgment of discriminator, the performance of the generator

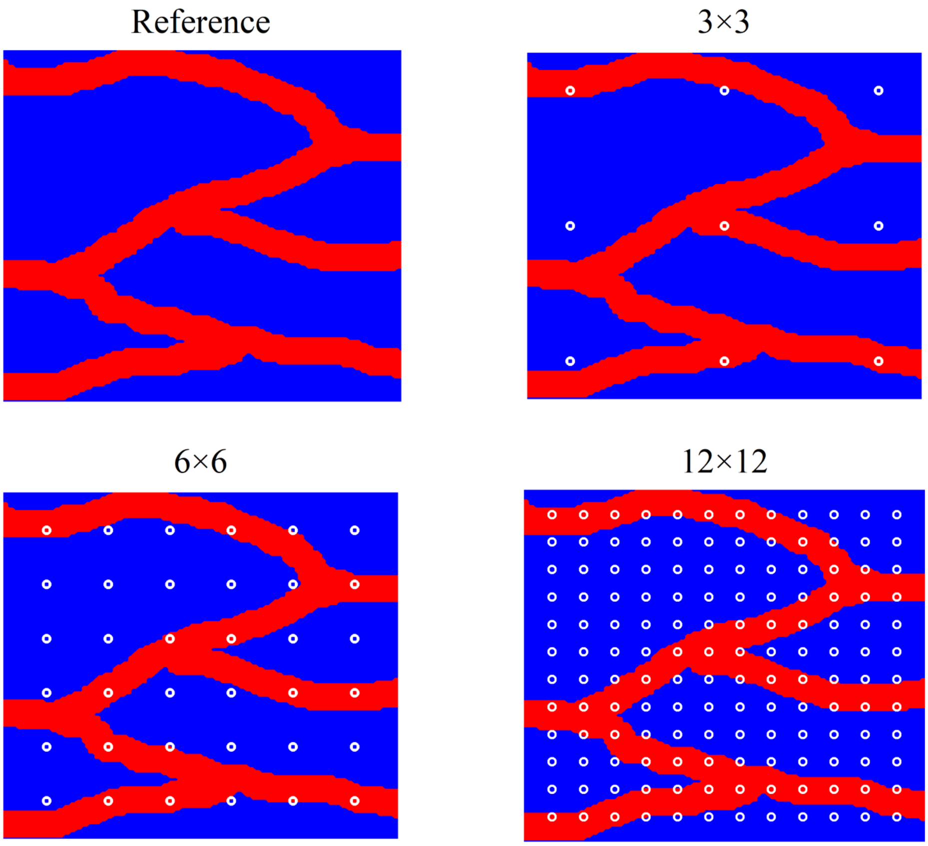


Fig. 6. The reference conductivity field cut from training image and locations of observation wells. The well locations are represented by the white circles.

and discriminator are improved using backpropagation. During training, the generator and the discriminator are competing against each other. The goal of the generator is to generate an image that can fool the discriminator, and the discriminator needs to do its best to distinguish whether the image is from real data or generated by the generator. The data generated by the generator is $x \sim P_G(x;\theta)$, where θ indicates the generator's network parameters. The likelihood function can be represented as:

$$L(\theta) = \prod_{i=1}^{n} P_G\left(x_i;\theta\right) \tag{3}$$

The generator aims at make the distribution $P_G(x;\theta)$ close to the real data distribution $P_R(x)$, and the solution is to maximize the likelihood function. If we use θ^* to indicate the values of θ corresponding to the maximum likelihood, then θ^* can be represented as:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} P_G\left(x_i;\theta\right) \tag{4}$$

This equals to minimize the Kullback–Leibler (KL) divergence between $P_R(x)$ and $P_G(x)$ (Arjovsky and Bottou, 2017):

$$KL\left(P_R \middle| P_G\right) = \int_x P_R(x) log \frac{P_R(x)}{P_G(x)} dx \tag{5}$$

Generally, the methods of GAN are to minimize the KL divergence or the Jensen-Shannon (JS) divergence, and the governing equation of GAN can be represented as the value function $V(G, D)$ (Goodfellow et al., 2014):

$$\underset{G}{\operatorname{argmin}}\,\underset{D}{\max}\, V\left(G, D\right) = E_{x \sim P_R}\left[logD(x)\right] + E_{\mathbf{z} \sim P_G}\left[log\left(1 - D\left(G(\mathbf{z})\right)\right)\right] \tag{6}$$

This function is solved through consecutively optimizing the discriminator and generator by minimizing the following equations:

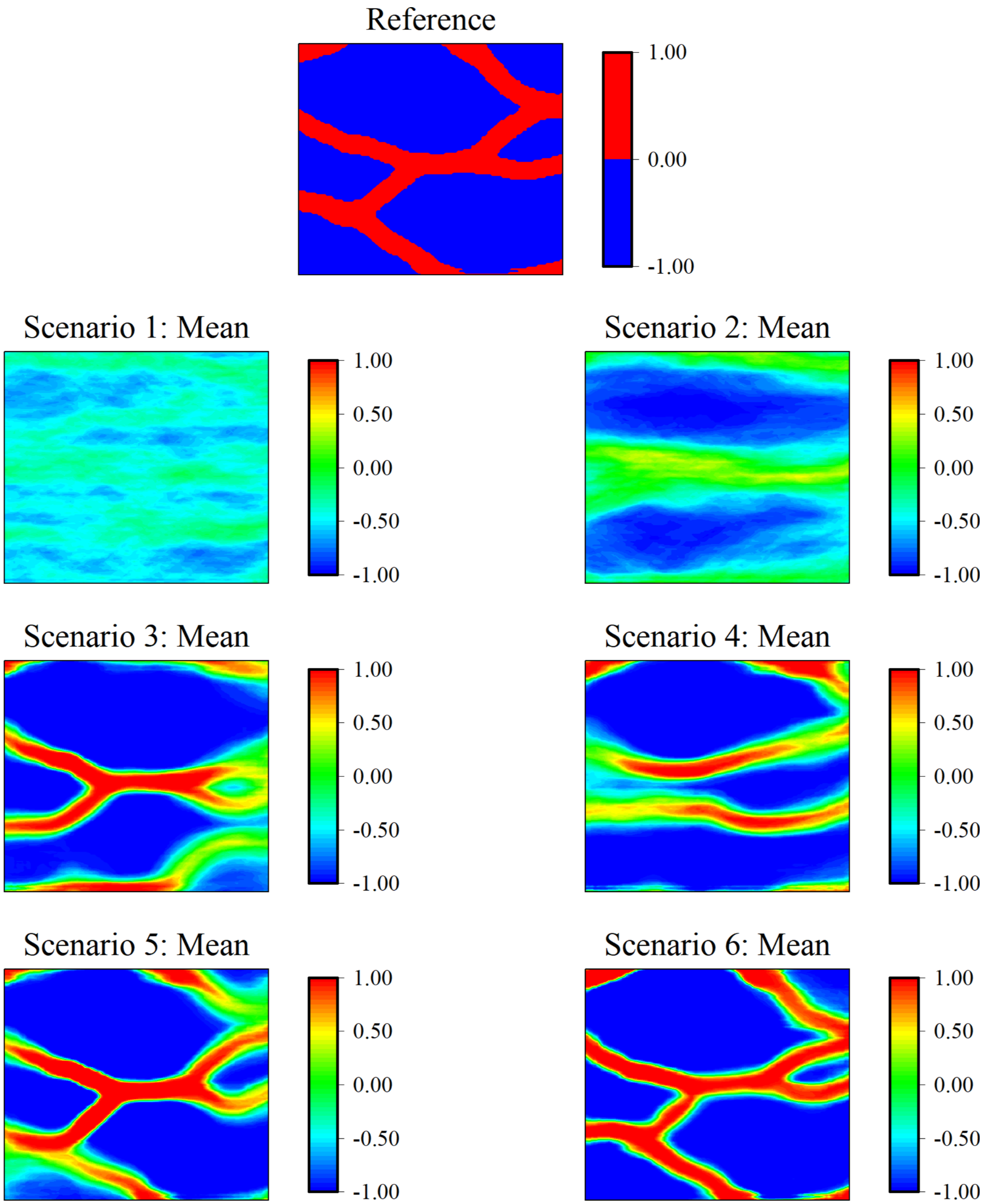
$$L^D = -E_{x \sim P_R}[logD(x)] - E_{\mathbf{z} \sim P_G}[log(1 - D(G(\mathbf{z})))] \tag{7}$$

$$L^G = E_{\mathbf{z} \sim P_G}[log(1 - D(G(\mathbf{z})))] \tag{8}$$

The structure of GAN used in this paper is similar to that used by Laloy et al. (2018). The training data is extracted from the training image shown in Fig. 2 (Laloy et al., 2018), and the architectures of generator and discriminator are shown in Table 1. The batch size is 16, and the learning rate is 0.001, the networks are trained for 100 epochs. In the table, k represents the kernel size, s represents the stride, p represents the zero-paddings, and d represents the dilation. The generator is trained using 2D transposed convolution, and the activation functions of the first 6 layers are ReLU. The final layer is another activation function called Tanh, and InsNorm2d represents the instance normalization. The discriminator is trained using 2D convolution, the activation functions of the first 4 layers are LeakyReLU(0.2), and the final layer is a Sigmoid function.

2.3. Ensemble smoother with multiple data assimilation

In the ES-MDA, we mainly focus on the parameter estimation, and the ES-MDA is used to update the parameters of the latent space $\mathbf{z}$. The parameter matrix at the ith step can be expressed as follows:



Reference

Scenario 1: Mean

Scenario 2: Mean

Scenario 3: Mean

Scenario 4: Mean

Scenario 5: Mean

Scenario 6: Mean

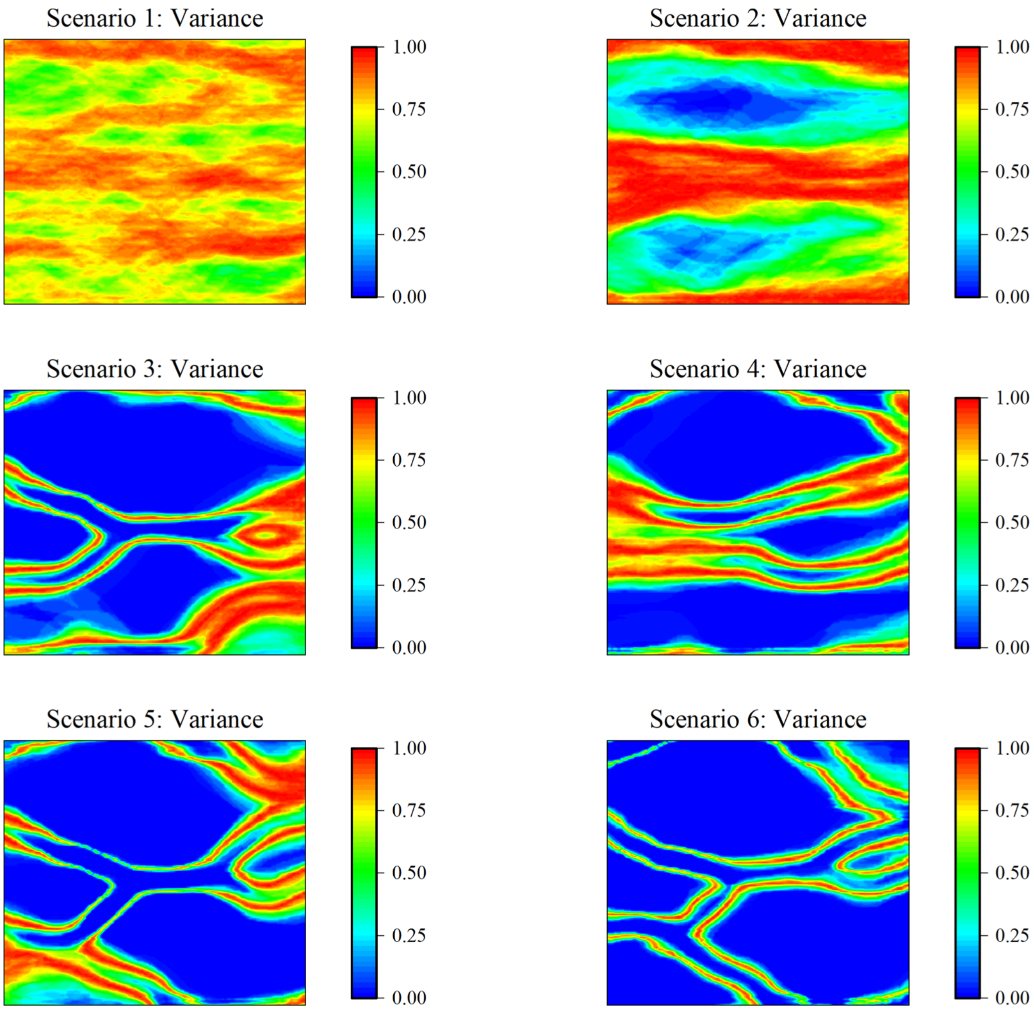
Fig. 7. The mean (logarithm) of conductivity for different scenarios.

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}_{11} & \mathbf{z}_{12} & \cdots & \mathbf{z}_{1n} \\ \mathbf{z}_{21} & \mathbf{z}_{22} & \cdots & \mathbf{z}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}_{m1} & \mathbf{z}_{m2} & \cdots & \mathbf{z}_{mn} \end{bmatrix} \tag{9}$$

the subscript m is the number of parameters of the latent space, and n is the number of realizations. In data assimilation, the forward model can be defined as the following form:

$$\mathbf{d} = F(\mathbf{k}) + \epsilon \tag{10}$$

where $\mathbf{d}$ is a vector containing N_d simulated data, such as the hydraulic head and concentration data, $F(.)$ is the forward operator, such as MODFLOW-2000 (Harbaugh et al., 2000), $\mathbf{k}$ is a vector containing N_k model parameter, and ϵ is the model error. The objective is to estimate the parameters that best reflect the observation data $\mathbf{d}_{obs}$, and the perturbed observation data is:


Scenario 1: Variance

Scenario 2: Variance

Scenario 3: Variance

Scenario 4: Variance

Scenario 5: Variance

Scenario 6: Variance

Fig. 8. The variance of conductivity for different scenarios.

$$\mathbf{d}_{uc} = \mathbf{d}_{obs} + \sqrt{\alpha_i}\mathbf{C}_D^{1/2}\mathbf{I}_d \tag{11}$$

where α_i is the inflation coefficient of the ith iteration ($\sum_{i=1}^{N_a}\frac{1}{\alpha_i} = 1$, N_a is the number of iterations), $\mathbf{C}_D$ is the observation error covariance, and $\mathbf{I}_d \sim N(0, \mathbf{I}_{Nd})$. The updating process in ES-MDA is represented as the following equation:

$$\mathbf{z}_j^{i+1} = \mathbf{z}_j^i + \mathbf{C}_{ZD}^i(\mathbf{C}_{DD}^i + \alpha_i\mathbf{C}_D)^{-1}(\mathbf{d}_{uc,j}^i - \mathbf{d}_j^i), \quad (j = 1, \ldots, N_r) \tag{12}$$

where $\mathbf{C}_{ZD}^i$ is the cross-covariance between the model parameters and the simulated data, $\mathbf{C}_{DD}^i$ is the auto-covariance of the simulated data, and N_r is the number of realizations. The procedure of coupling GAN with ES-MDA is shown in Algorithm 1. According to Emerick and Reynolds (2013), if the values of α_i are in decreasing order, a better result can be achieved. Therefore, α_i is set as follows:

$$\alpha_i = 2^{N_a - p}, \quad (p = 0, \ldots, N_a - 1) \tag{13}$$

In the updating equation, the matrix $\mathbf{C} = \mathbf{C}_{DD}^i + \alpha_i\mathbf{C}_D$ needs to be inverted. The matrix may be singular and small singular values may cause instability and error, so a Truncated Singular Value Decomposition (TSVD) is applied to solve the pseudo-inverse, and the measurement error covariance $\mathbf{C}_D$ is rescaled wtih the Cholesky decomposition $\mathbf{C}_D = \mathbf{C}_D^{1/2}(\mathbf{C}_D^{1/2})^T$ (Emerick, 2012). Apply TSVD to matrix $\mathbf{C}$ to obtain matrix $\widehat{\mathbf{C}}$:

$$\widehat{\mathbf{C}} = \mathbf{U_n}\Lambda_\mathbf{n}\mathbf{V_n}^T \tag{14}$$

where $\Lambda_\mathbf{n}$ is a diagonal matrix containing N_n largest singular values and N_n is defined, according to:

$$\widehat{\mathbf{C}} = \frac{\sum_{i=1}^{N_n}\lambda_i}{\sum_{i=1}^{N_t}\lambda_i} \leqslant E \tag{15}$$

where λ_i are the singular values sorted in a decreasing order, N_t is the total number of singular values, and E is the energy of the singular values retained, typically between 0.9 and 1.0. Thus, N_n is the number that makes the ratio of the sum of the N_n largest singular values to the sum of the total singular values less than or equal to E. The pseudo-inverse of matrix $\mathbf{C}$ can be solved as follows:

$$\mathbf{C}^+ = \mathbf{V_n}\Lambda_\mathbf{n}^+\mathbf{U_n}^T \tag{16}$$

where $\Lambda_\mathbf{n}^+$ is the pseudo-inverse of $\Lambda_\mathbf{n}$.

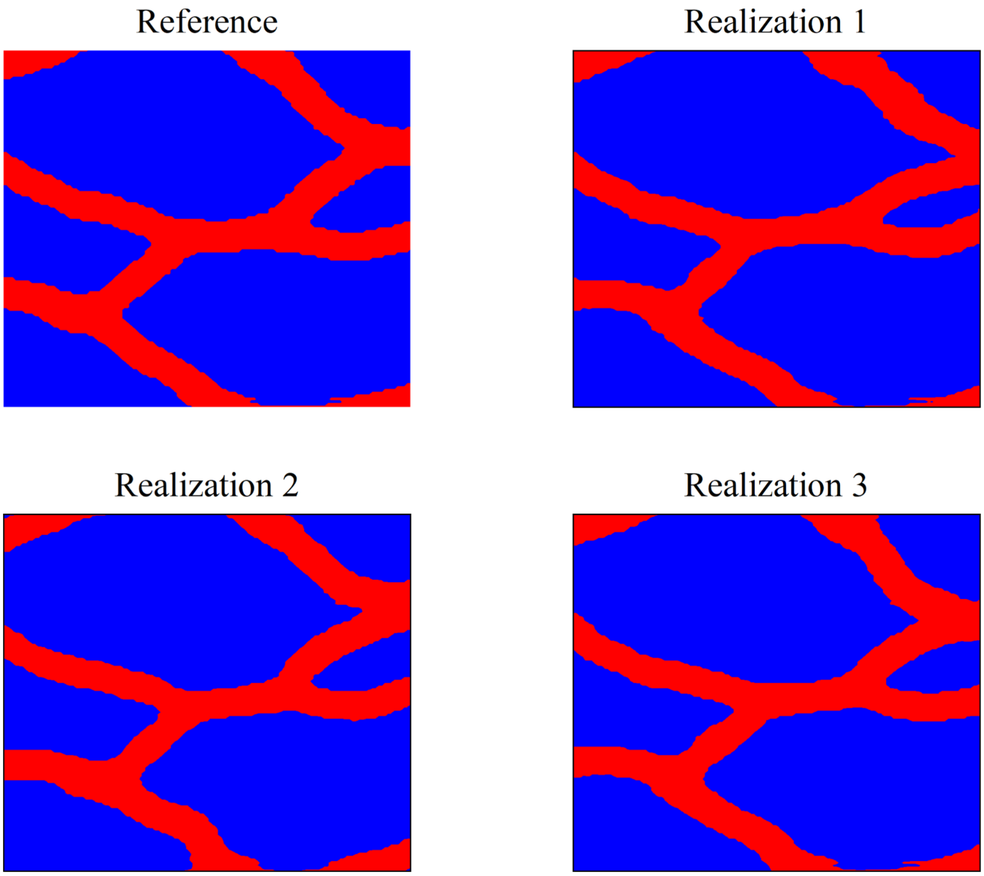


Fig. 9. Individual realizations for Scenario 6.

```
Set: $N_a$ = The number of iterations
Set: $\mathbf{d}_{obs}$ = Observation data (hydraulic head or concentration data)
Set: $N_r$ = The number of realizations
begin
    Sample initial $\mathbf{z}$ from the uniform distribution $\mathbf{z} \sim U(-1,1)$
    for $i = 1, 2, \cdots, N_a$ do
        $\alpha_i = 2^{N_a - i}$
        for $j = 1, 2, \cdots, N_r$ do
            Generate the hydraulic conductivity field with the trained generator: $\mathbf{K}_j^i = G(\mathbf{z}_j^i)$
            Run forward modeling to obtain the hydraulic head or concentration data $\mathbf{d}_j^i$
            Perturb the observation data: $\mathbf{d}_{uc,j}^i = \mathbf{d}_{obs} + \sqrt{\alpha_i}\mathbf{C}_D^{1/2}\mathbf{I}_d$
        Calculate: $\mathbf{C}_{ZD}^i = \frac{1}{N_r - 1}\sum_{j=1}^{N_r}(\mathbf{z}_j^i - \bar{\mathbf{z}}^i)(\mathbf{d}_j^i - \bar{\mathbf{d}}^i)^{\mathrm{T}}$
        Calculate: $\mathbf{C}_{DD}^i = \frac{1}{N_r - 1}\sum_{j=1}^{N_r}(\mathbf{d}_j^i - \bar{\mathbf{d}}^i)(\mathbf{d}_j^i - \bar{\mathbf{d}}^i)^{\mathrm{T}}$
        Update: $\mathbf{z}_j^{i+1} = \mathbf{z}_j^i + \mathbf{C}_{ZD}^i(\mathbf{C}_{DD}^i + \alpha_i\mathbf{C}_D)^{-1}(\mathbf{d}_{uc,j}^i - \mathbf{d}_j^i), j = 1, ..., N_r$
end
```

3. Synthetic examples

3.1. Model setup

Synthetic examples are presented to evaluate the performance of coupling GAN with ES-MDA. A reference hydraulic conductivity field and its boundary conditions are shown in Fig. 3. The confined aquifer has a size of 129 m × 129 m × 1 m, which is discretized into 129 columns by 129 rows by 1 layer. The reference hydraulic conductivity field is an image generated by the trained generator via GAN, using the training image in Fig. 2. Sandy channels represent high conductivity conduits, and the silt floodplain deposits have low hydraulic conductivities. The north and south boundaries of the aquifer are assumed to be no flow boundary conditions, and the west side is a constant head boundary ($h = 0$ m). The east side is considered as a constant flow

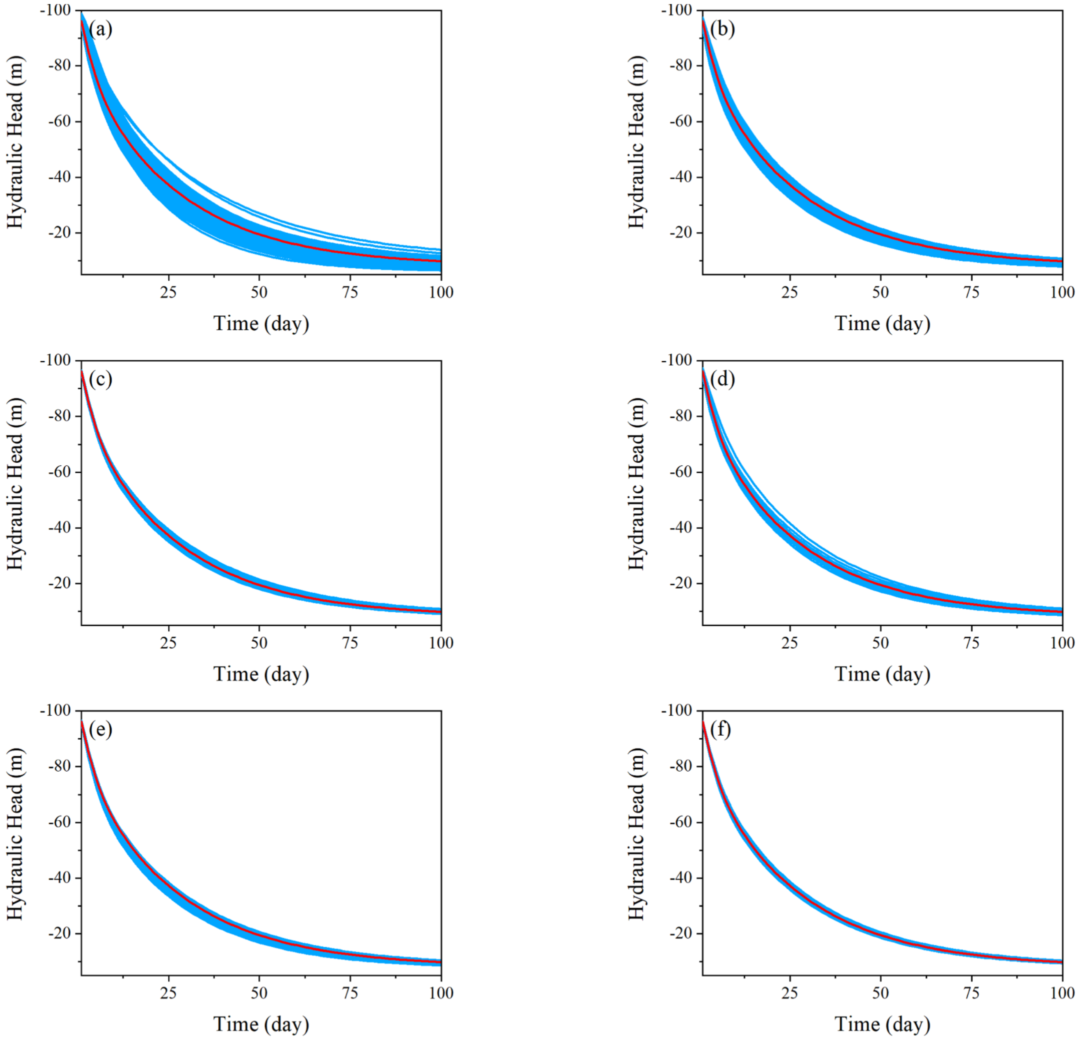


Fig. 10. The simulated hydraulic head for different scenarios. (a) # 1 of Scenarios 1; (b) # 3 of Scenarios 2; (c) # 3 of Scenarios 3; (d) # 5 of Scenarios 4; (e) # 5 of Scenarios 5; (g) # 6 of Scenarios 5. The red line is the reference hydraulic head, and the blue lines indicate the simulated hydraulic head of each realization. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

boundary and the prescribed flow rate is -12.9 m^3/d. The initial hydraulic head is equal to the depth of the aquifer (-100 m) over the simulated domain. The specific storage of the aquifer is assumed to be a constant value of 0.003 m^{-1}. The porosity is set to 0.3 everywhere. The MODFLOW-2000 (Harbaugh et al., 2000) is used to model the groundwater flow in 100 days. The MT3DMS (Zheng and Wang, 1999) is used to simulate solute transport. The black line shown in Fig. 3 represents the contaminant source, and its concentration is assumed to be a constant value of 10 mg/L. Only advection and dispersion are considered for solute transport. The longitudinal dispersivity is set to 10 m over the simulated area, and the horizontal transverse dispersivity is 1 m.

3.2. Scenario studies

Six scenarios with different numbers and different locations of the observation wells are conducted to evaluate the influence of information on the results (see Fig. 4). The settings of each scenario are shown in Table 2. In Scenario 1, no conditioning data are considered. The initial ensemble of hydraulic conductivities are generated using the trained generater as the reference field. Scenario 2 has only one observation well located at the center of the aquifer. Scenario 3 and Scenario 4 increase the number of the observation wells to 5, but the spacing between each two wells in Scenario 4 is larger than that of Scenario 3. Scenario 5 and Scenario 6 have 9 observation wells. The distance between each two wells in Scenario 5 is 20 m, while the distance in Scenario 6 is 50 m. Note that the wells at the corner and center in Scenario 5 and Scenario 6 are at the same places as that of Scenario 3 and Scenario 4. Fig. 4 shows the well locations in space for each Scenario. The hydraulic head data are used for data assimilation. The reference hydraulic head is obtained by running the groundwater flow model using the reference conductivity field, and have a measurement error with a mean of 0 and a standard deviation of 0.5. The ensemble size is 100, and the number of iterations in the ES-MDA is set to 8, which has shown to be sufficient to achieve a reason0able result in previous studies (e.g., Li et al., 2018).

3.3. Solute transport

Concentration data also can be used to identify hydraulic conductivities. Solute transport experiments are conducted to simulate the migration of contaminants in the aquifer. Concentration data are sampled at each time step at different locations, and the measured concentrations are added with Gaussian noise with a mean of 0 and a standard deviation of 0.5 as the observation error. Four Cases are compared in this set of experiments. Case 1, Case 2, and Case 3 are used to evaluate the effects of different types of observation data. Case 2 and Case 4 are compared to evaluate the results when more concentration

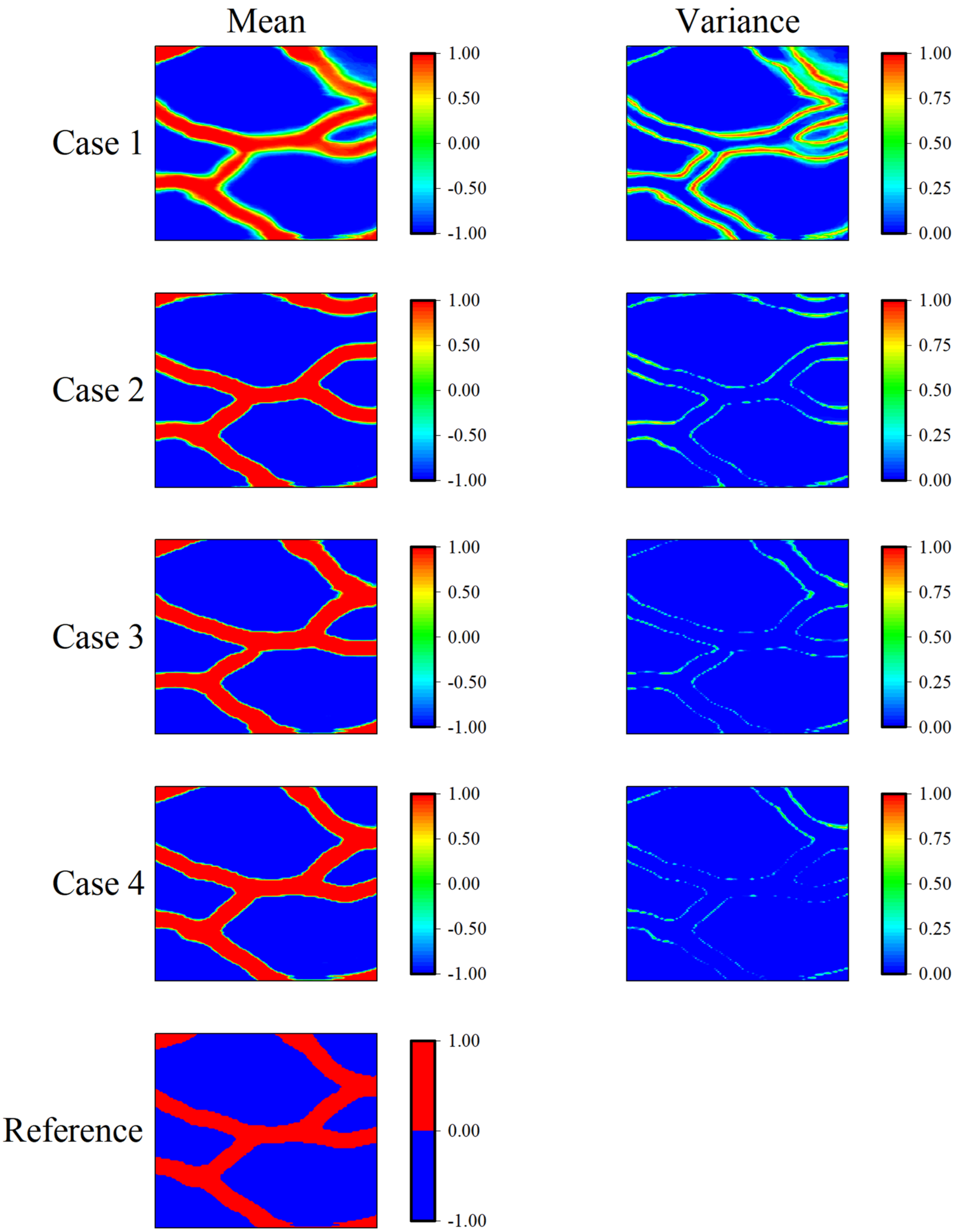


Fig. 11. The mean (logarithm) and variance of conductivity for each case.

data are considered. Note that Case 1 is the same as Scenario 5 in the last section, and only hydraulic head collected from 9 wells will be used to identify hydraulic conductivities. In Case 2, only concentration data will be used for data assimilation, and there are 36 observation wells (6 × 6) which are uniformly distributed over the aquifer for sampling the concentration, and Case 3 uses both the hydraulic head of Case 1 and concentration data of Case 2 for data assimilation. Case 4 only considers concentration data for data assimilation but increases the number of measurements to 81 (9 × 9). Fig. 5 shows the location of wells for sampling head and concentration data for each Case, and the settings of each case are shown in Table 3.

RMSE and *Spread* are used to evaluate the result. These two metrics have already been used in other studies to assess the results in data assimilation (e.g., Chen and Zhang, 2006; Franssen et al., 2008; Li et al., 2018). *RMSE* is the root mean square error, which indicates the bias and can be used to measure the accuracy of estimation, and *Spread* measures the uncertainty of ensemble retaliations:

$$RMSE = \sqrt{\frac{1}{N_k}\sum_{j=1}^{N_k}(\overline{\mathbf{K}_j} - \mathbf{K}_{ref})^2} \tag{17}$$

$$Spread = \sqrt{\frac{1}{N_k}\sum_{j=1}^{N_k} Var(\mathbf{K}_j)} \tag{18}$$

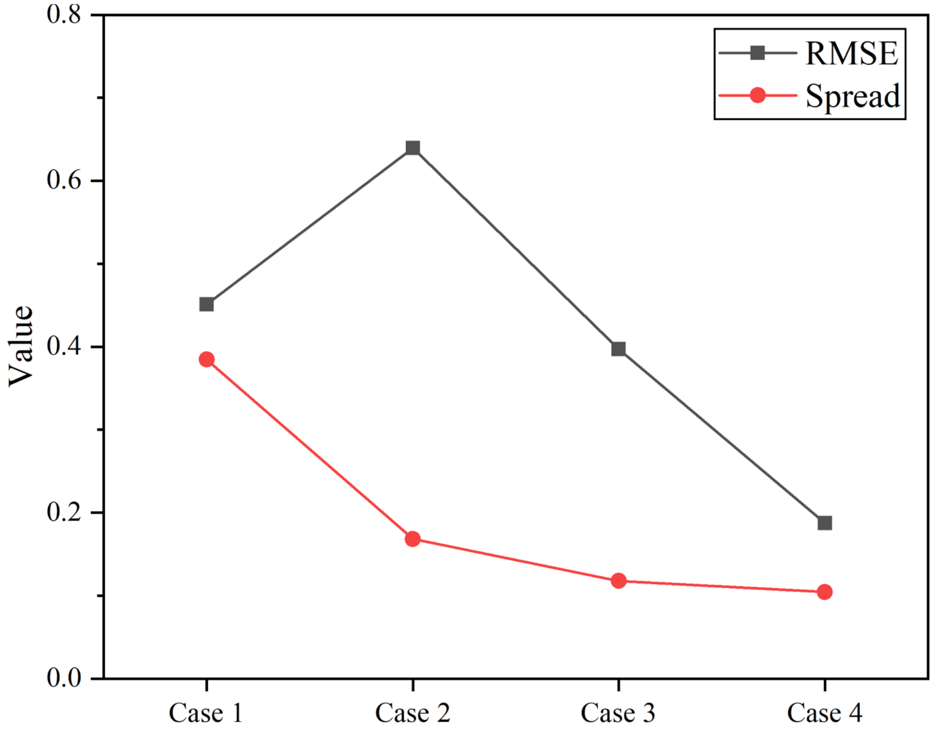


Fig. 12. The hydraulic conductivity estimation RMSE and Spread of each case. The black line represents the RMSE and the red line represents the Spread. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where N_k is the number of parameters, $\overline{\mathbf{K}_j}$ is the mean of estimated parameters, $\mathbf{K}_{ref}$ is the reference parameters, and $Var(\mathbf{K}_j)$ is the variance at each point.

3.4. Random image for reference field

In previous examples, the reference conductivity field is generated through a trained generator using the GAN, and multiple data are assimilated into the groundwater flow and contaminant transport model to recover the reference image by coupling the trained generator and the ES-MDA. It is more challenging if the reference image is directly extracted from the training image instead of the one generated by deep learning, which is more close to the real case in the field. In other words, the reference image (i.e., the random image here) is not included in the training data sets for the GAN. Here, an image randomly cut from the training image in Fig. 2 is used as the reference hydraulic conductivity field to test our approach. Note that, although this random image was a random cut from training image in Fig. 2, it is not used in the training of GAN. The hydraulic head collected from 9 wells (3 × 3), 36 wells (6 × 6), and 144 wells (12 × 12) are used for data assimilation respectively. The reference conductivity field and the observation well locations are shown in Fig. 6. The observation wells are uniformly distributed over the domain, and the boundary conditions are the same as shown in Fig. 3.

4. Results

4.1. Effects of well locations

The mean maps of log-conductivities over 100 realizations for 6 Scenarios are shown in Fig. 7, and the variance maps are shown in Fig. 8. Note that the unit of the conductivities is m/d. We can see that: (1) When the number of wells is increased from 1 of Scenario 1 to 9 of Scenario 6, the characterization of hydraulic conductivity is clearly improved; (2) When more wells locate in or near the channels, a better result can be obtained, as shown in Scenario 3 and 4, because water moves faster in high conductivity areas, this can provide more information for updating process; (3) When the number of wells is same as in Scenario 5 and 6, it is evident that the larger spacing between wells shows a better result because hydraulic head data could provide more information about the channel structures in space; (4) Scenario 6 has the best result, in which the mean of log-conductivity is close to the reference image in Fig. 3; and (5) the uncertainty (i.e., variance) is reduced, when more data are considered, in particular for Scenario 6.

Fig. 9 shows the reference image and three individual realizations from Scenario 6. It clearly shows that the reference image has similar channel structures as those in the training image in Fig. 2, which means that GAN is an efficient approach to model the non-Gaussian aquifer. The low dimension variable in latent space is able to represent the high dimension hydraulic conductivity field. This is consistent with the findings from Laloy et al. (2018), where a detailed discussion about the effectiveness of GAN for modeling channelized aquifers is listed. Also, the individual realizations after assimilating hydraulic head data show the similar geological structures (i.e., connectivity) as the reference field, which plays a critical role for the flow and transport predictions.

In order to compare the simulated hydraulic head, the well in the middle of each Scenario is presented in Fig. 10. The red line is the reference hydraulic head, and the blue lines indicate the result of each realization. As expected, the result is better as the number of observation wells increases. The simulated hydraulic head of Scenario 5 (Fig. 10(f)) are much closer to the reference value because of the best characterization of conductivity field (i.e., connectivity). In addition, since the hydraulic conductivity estimation of Scenario 4 is not as good as the one in Scenario 3, the uncertainty of simulated hydraulic head is larger (Fig. 10(d)).

4.2. Contaminant transport prediction

The mean and variance of each case are shown in Fig. 11, and the values of *RMSE* and *Spread* are shown in Fig. 12. The variance and *Spread* illustrate that as more data are available, the uncertainty of estimation can be significantly reduced. In Case 1, only the hydraulic

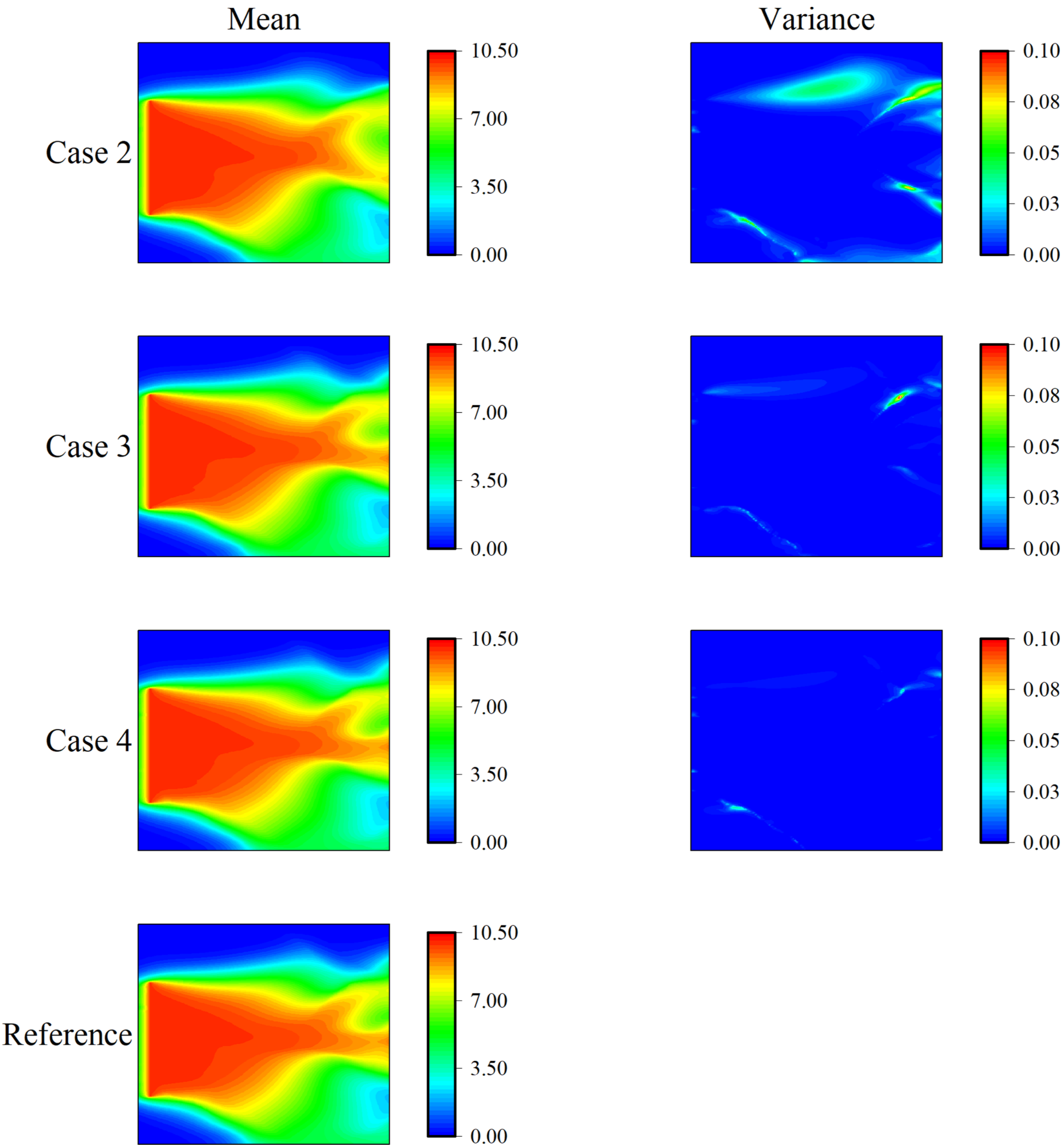


Fig. 13. The contaminant plume at 100 days. The first column indicates the mean concentration (mg/L) for each case, and the second column is the variance. The last image is the reference concentration.

head from 9 wells are used for data assimilation, and the boundaries of the channels are not as clear as the other Cases. In Case 2, although the channel boundaries are much clearer, some channels are at the wrong locations, like the one at the upper right corner, which causes the *RMSE* value of Case 2 to be the largest. If the information of hydraulic head and contaminant concentration are combined together as in Case 3, the shape of the channels can be accurately reproduced. By comparing Case 2 and Case 4, a better result can be achieved if more concentration data are available. This experiment illustrates that combining different types of data or increasing the number of observation wells can lead to a better result.

The results for contaminant transport predictions are shown in Fig. 13. The contaminant plume of Case 2, Case 3, and Case 4 look similar in general, but the concentration near the right boundary of Case 3 and Case 4 are closer to the reference plume. Additionally, the variance of Case 2 is larger than that of Case 3 and Case 4. In order to quantify the bias and uncertainty of each Case, *RMSE* and *Spread* are calculated and plotted in Fig. 14. Both the *RMSE* and *Spread* show a downward trend, which means that the error and uncertainty are significantly reduced. The result of Case 4 is the best since both hydraulic head and concentration data are used for conditioning.

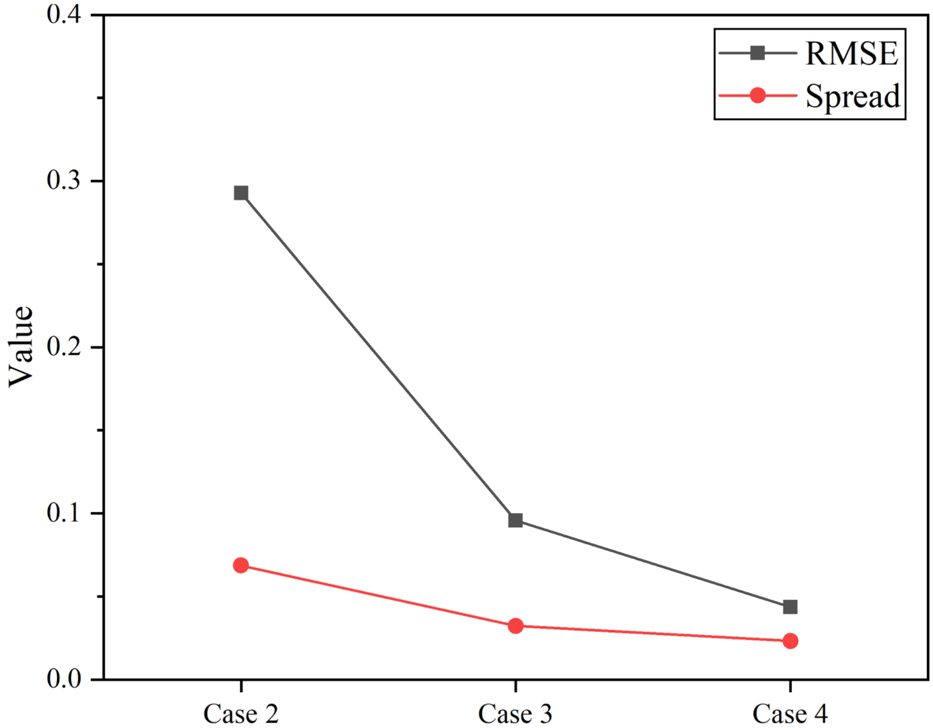


Fig. 14. The RMSE and Spread of concentration. The black line represents the RMSE of Case 2, Case 3, and Case 4. The red line represents the Spread of Case 2, Case 3, and Case 4. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.3. Random image for reference field

A random image that is not generated by the generator or extracted from the training set is more reasonable for a real-world application. The mean and variance of ensemble hydraulic conductivities for the cases of 9 wells (3 × 3), 36 wells (6 × 6), and 144 wells (12 × 12) are shown in Fig. 15. The mean is closer to the reference as the number of observation wells increase, and the variance (i.e., uncertainty) is smaller, accordingly. Fig. 16 shows the *RMSE* and *Spread*, and both lines decrease as more observation data are considered. This demonstrates that the estimation error and uncertainty are reduced as the number of information is increased for assimilation. Meanwhile, as shown in Figs. 7 and 8, only 9 wells are able to obtain a good result when the reference conductivity is generated by the trained generator. However, the result of 9 wells, in this case, cannot accurately reconstruct the hydraulic conductivity field or preserve the channel structures, when the reference field is a random image from the Fig. 2. In addition, as mentioned above, if more observation data become available, such as for the case of 144 wells, the reference field still can be reproduced using the proposed method.

5. Discussion

In this paper, for the sake of simplicity, only 2D cases are presented, but the proposed method could be also extended in 3D. Studies have proved the ability of GAN to model fluvial deposits and meandering channels in 3D cases (e.g., Mosser et al., 2017; Laloy et al., 2018; Zhang et al., 2019). One of the challenges for the GAN is to condition on hard data (i.e., measured conductivities). Approaches have been proposed to deal with this issue. For example, Zhang et al. (2019) developed a semantic in-painting approach with a new loss function to make the images generated by GAN honoring the measured facies at well locations. Ruffino et al. (2020) proposed a framework by adding an explicit cost term to the GAN loss function to enforce pixel-wise conditioning. As a result, the hard data can be constrained in the generated images. In addition, studies have shown that GAN is able to reconstruct different structures (Chan and Elsheikh, 2017; Gao et al., 2020) or the porous media with different scales (e.g., Guan, 2018;Hsu et al., 2020;Feng et al., 2019). Although conditioning on hard data using GAN has been studied in the literature, how to jointly condition on both hard data and dynamic data has not been explored yet, and a future study will be conducted.

We proposed to couple the ES-MDA with deep learning to deal with non-Gaussianity in data assimilation, which is different with the past approaches such as the pattern-match method (Zhou et al., 2012; Li et al., 2013), which relies on searching patterns through the ensemble training images consisting of both conductivity and head in order to preserve the channel structures. The searching procedure is time-consuming and computationally expensive. When deep learning is introduced into data assimilation, the trained generator is much faster to generate the channelized aquifer by changing the low-dimension latent variable, which will then be updated by conditioning on dynamic data through the ES-MDA.

Since the locations of channels are unknown in the field, an even distribution of wells would be placed at first, as the synthetic examples shown here. Depending on the responses of wells (i.e., measured hydraulic head data), future well locations could be adjusted to capture dynamic data which could bring most information for the parameter estimation.

6. Conclusion

In this paper, a new approach of coupling deep learning and ES-MDA is proposed to deal with the data assimilation problem of the non-Gaussian channelized aquifer. The ES-MDA method is applied to update the parameters of the latent space by assimilating both the hydraulic head and contaminant concentration data. Then the trained generator through GAN is able to reproduce the channelized structures with much fewer parameters (i.e, latent variable). The results demonstrate that the coupling of GAN and ES-MDA can accurately reconstruct the channelized aquifer and reduce the uncertainty of hydraulic head and contaminant concentration prediction. A couple of synthetic examples are

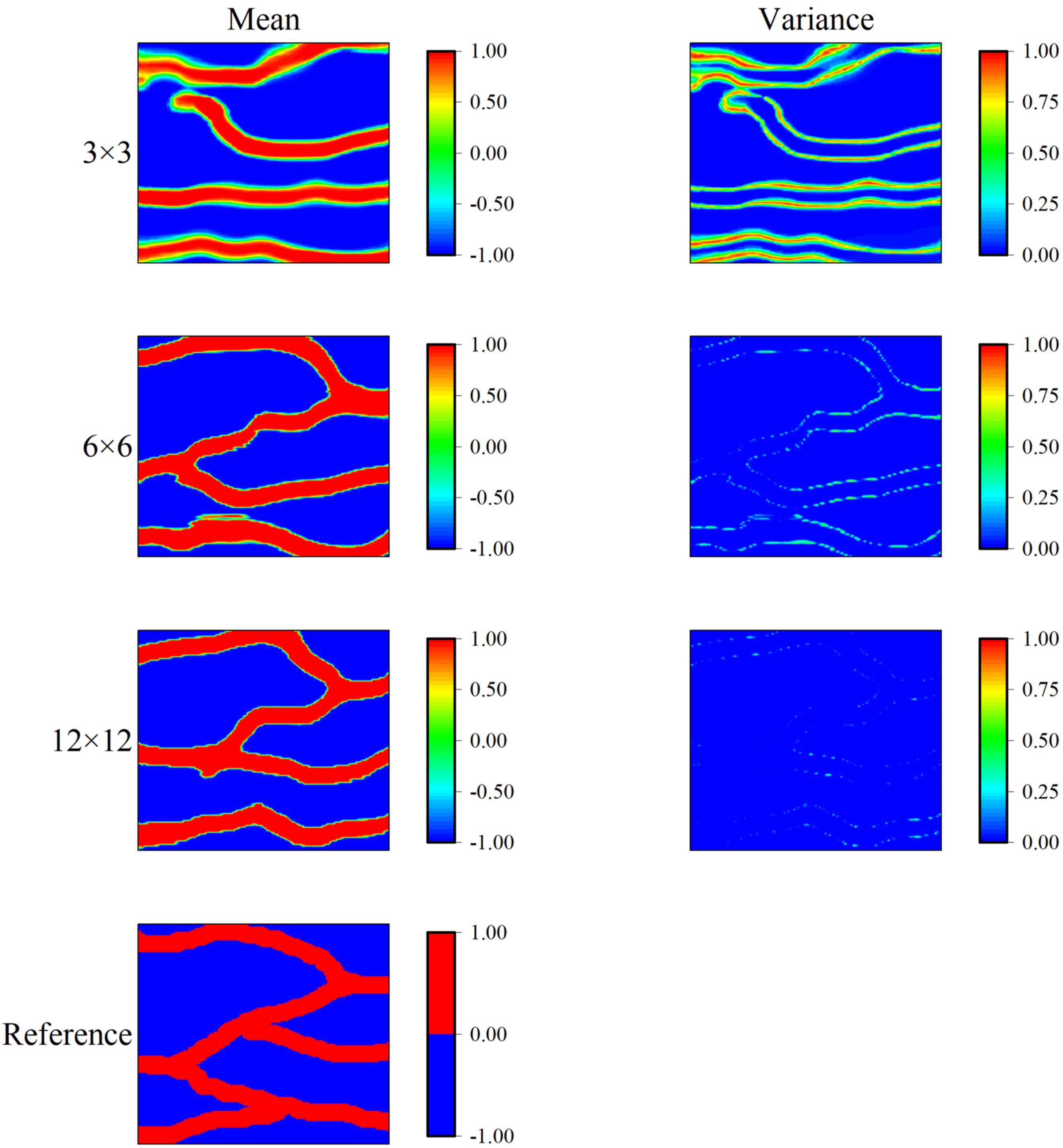


Fig. 15. The mean (logarithm) and variance of the random image test.

used to test the proposed method, where the number of parameters to be updated is reduced from 129 by 129 of conductivities to 5 by 5 of the latent space variable for each realization. The results indicate: (1) increasing the number of observation wells can lead to a better characterization of channel structures and reduce the uncertainty of estimation; (2) when additional concentration data are considered, the characterization of channel structures are the best; (3) more data are needed for reproducing the channel structure, if the reference conductivity field is not generated by the trained generated through deep learning which often occurs in the field; and (4) the coupling of deep learning with GAN and ES-MDA is an efficient and effective data assimilation method for identifying non-Gaussian channel structures.

CRediT authorship contribution statement

Jichao Bao: Software, Formal analysis, Investigation, Writing - original draft. **Liangping Li:** Supervision, Conceptualization, Methodology. **Fleford Redoloza:** Software, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

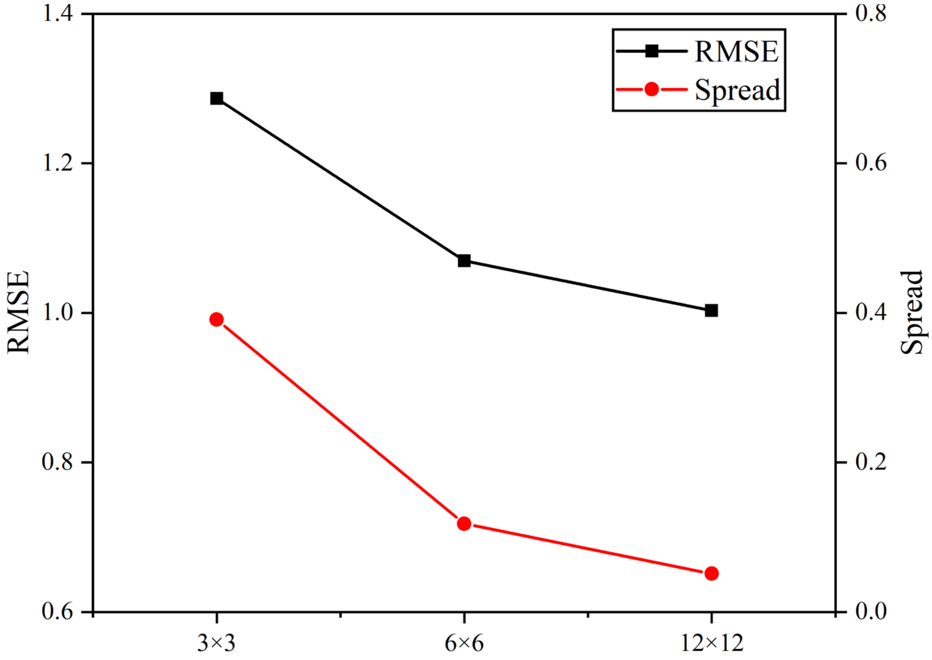


Fig. 16. The RMSE and Spread of the random image test. The black line represents the RMSE of different observation wells, and the red line shows the Spread.

Acknowledgement

The authors acknowledges the financial support of the South Dakota Board of Regents through a Competitive Research Grant. This work also has been supported through a grant from the National Science Foundation (OIA-1833069). The authors wish to thank the associate editor and five anonymous reviewers for their comments, which substantially helped to improve the final version of the manuscript.

References

Arjovsky, M., Bottou, L., 2017. Towards principled methods for training generative adversarial networks. arxiv e-prints, art. arXiv preprint arXiv:1701.04862.
Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein gan. arXiv preprint arXiv:1701.07875.
Bailey, R., Baù, D., 2010. Ensemble smoother assimilation of hydraulic head and return flow data to estimate hydraulic conductivity distribution. Water Resour. Res. 46.
Chan, S., Elsheikh, A.H., 2017. Parametrization and generation of geological models with generative adversarial networks. arXiv preprint arXiv:1708.01810.
Chen, Y., Oliver, D.S., 2012. Ensemble randomized maximum likelihood method as an iterative ensemble smoother. Math. Geosci. 44, 1–26.
Chen, Y., Zhang, D., 2006. Data assimilation for transient flow in geologic formations via ensemble kalman filter. Adv. Water Resour. 29, 1107–1122.
Emerick, A., 2012. History Matching and Uncertainty Characterization: Using Ensemble-based Methods. LAP LAMBERT Academic Publishing.
Emerick, A.A., Reynolds, A.C., 2013. Ensemble smoother with multiple data assimilation. Comput. Geosci. 55, 3–15.
Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. J. Geophys. Res. Oceans 99, 10143–10162.
Evensen, G., 2018. Analysis of iterative ensemble smoothers for solving inverse problems. Comput. Geosci. 22, 885–908.
Feng, J., He, X., Teng, Q., Ren, C., Chen, H., Li, Y., 2019. Reconstruction of porous media from extremely limited information using conditional generative adversarial networks. Phys. Rev. E 100 033308.
Fetter, C.W., 2018. Applied Hydrogeology. Waveland Press.
Fokker, P., Wassing, B., Van Leijen, F., Hanssen, R., Nieuwland, D., 2016. Application of an ensemble smoother with multiple data assimilation to the bergermeer gas field, using ps-insar. Geomech. Energy Environ. 5, 16–28.
Gao, X., He, W., Hu, Y., 2020. Modeling of meandering river deltas based on the conditional generative adversarial network. J. Petrol. Sci. Eng 107352.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 2672–2680.
Gu, Y., Oliver, D., 2007. An iterative ensemble kalman filter for multiphase fluid flow data assimilation. SPE J. 12, 438–446.
Guan, K., 2018. Reconstructing pore networks using generative adversarial networks.
Harbaugh, A.W., Banta, E.R., Hill, M.C., McDonald, M.G., 2000. Modflow-2000, the u. s. geological survey modular ground-water model-user guide to modularization concepts and the ground-water flow process. Open-file Report. U. S. Geological Survey 134.
Hendricks Franssen, H., 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.
Hsu, T., Epting, W.K., Kim, H., Abernathy, H.W., Hackett, G.A., Rollett, A.D., Salvador, P. A., Holm, E.A., 2020. Microstructure generation via generative adversarial network for heterogeneous, topologically complex 3d materials. arXiv preprint arXiv:2006.13886.
Hu, L.Y., 2000. Gradual deformation and iterative calibration of gaussian-related stochastic models. Math. Geol. 32, 87–108.
Huang, K., Mohanty, B.P., Leij, F.J., van Genuchten, M.T., 1998. Solution of the nonlinear transport equation using modified picard iteration. Adv. Water Resour. 21, 237–249.
Janssens, N., Huysmans, M., Swennen, R., 2020. Computed tomography 3d super-resolution with generative adversarial neural networks: implications on unsaturated and two-phase fluid flow. Materials 13, 1397.
Kim, S., Min, B., Lee, K., Jeong, H., 2018. Integration of an iterative update of sparse geologic dictionaries with es-mda for history matching of channelized reservoirs. Geofluids 2018.
Kuczera, G., Kavetski, D., Renard, B., Thyer, M., 2010. A limited-memory acceleration strategy for mcmc sampling in hierarchical bayesian calibration of hydrological models. Water Resour. Res. 46.
Laloy, E., Hérault, R., Jacques, D., Linde, N., 2018. Training-image based geostatistical inversion using a spatial generative adversarial neural network. Water Resour. Res. 54, 381–406.
Le, D.H., Younis, R., Reynolds, A.C., et al., 2015. A history matching procedure for non-gaussian facies based on es-mda. In: SPE Reservoir Simulation Symposium. Society of Petroleum Engineers.
Le Ravalec-Dupin, M., Nœtinger, B., 2002. Optimization with the gradual deformation method. Math. Geol. 34, 125–142.
Li, L., Puzel, R., Davis, A., 2018. Data assimilation in groundwater modelling: ensemble kalman filter versus ensemble smoothers. Hydrol. Process. 32, 2020–2029.
Li, L., Srinivasan, S., Zhou, H., Gómez-Hernández, J.J., 2013. A pilot point guided pattern matching approach to integrate dynamic data into geological modeling. Adv. Water Resour. 62, 125–138.
Li, L., Stetler, L., Cao, Z., Davis, A., 2018. An iterative normal-score ensemble smoother for dealing with non-gaussianity in data assimilation. J. Hydrol. 567, 759–766.
Li, L., Zhou, H., Gómez-Hernández, J.J., Franssen, H.J.H., 2012. Jointly mapping hydraulic conductivity and porosity by assimilating concentration data via ensemble kalman filter. J. Hydrol. 428, 152–169.
Liu, G., Chen, Y., Zhang, D., 2008. Investigation of flow and transport processes at the made site using ensemble kalman filter. Adv. Water Resour. 31, 975–986.
Marçais, J., de Dreuzy, J.R., 2017. Prospective interest of deep learning for hydrological inference.
Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.
Mo, S., Zabaras, N., Shi, X., Wu, J., 2019. Deep autoregressive neural networks for high-dimensional inverse problems in groundwater contaminant source identification. Water Resour. Res. 55, 3856–3881.
Mosser, L., Dubrule, O., Blunt, M.J., 2017. Reconstruction of three-dimensional porous media using generative adversarial neural networks. Phys. Rev. E 96 043309.

Oliver, D.S., Cunha, L.B., Reynolds, A.C., 1997. Markov chain monte carlo methods for conditioning a permeability field to pressure data. Math. Geol. 29, 61–91.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Ruffino, C., Hérault, R., Laloy, E., Gasso, G., 2020. Pixel-wise conditioned generative adversarial networks for image synthesis and completion. Neurocomputing.

Skjervheim, J.A., Evensen, G., et al., 2011. An ensemble smoother for assisted history matching. In: SPE Reservoir Simulation Symposium. Society of Petroleum Engineers.

Sun, A.Y., 2018. Discovering state-parameter mappings in subsurface models using generative adversarial networks. Geophys. Res. Lett. 45, 11–137.

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

Zhang, J., Lin, G., Li, W., Wu, L., Zeng, L., 2018. An iterative local updating ensemble smoother for estimation and uncertainty assessment of hydrologic model parameters with multimodal distributions. Water Resour. Res. 54.

Zhang, J., Zheng, Q., Wu, L., Zeng, L., 2020. Using deep learning to improve ensemble smoother: Applications to subsurface characterization. arXiv preprint arXiv:2002.09100.

Zhang, T.F., Tilke, P., Dupont, E., Zhu, L.C., Liang, L., Bailey, W., 2019. Generating geologically realistic 3d reservoir facies models using deep learning of sedimentary architecture with generative adversarial networks. Petrol. Sci. 16, 541–549.

Zhao, Y., Forouzanfar, F., Reynolds, A.C., 2017. History matching of multi-facies channelized reservoirs using es-mda with common basis dct. Comput. Geosci. 21, 1343–1364.

Zheng, C., Wang, P.P., 1999. MT3DMS: a modular three-dimensional multispecies transport model for simulation of advection, dispersion, and chemical reactions of contaminants in groundwater systems; documentation and user's guide. Technical Report. Alabama Univ University.

Zhou, H., Gómez-Hernández, J.J., Li, L., 2012. A pattern-search-based inverse method. Water Resour. Res. 48.

Zhou, H., Gómez-Hernández, J.J., Li, L., 2014. Inverse methods in hydrogeology: evolution and recent trends. Adv. Water Resour. 63, 22–37.