A Novel Deep Learning Approach for Data Assimilation of Complex Hydrological Systems

Jiangjiang Zhang[1,2], Chenglong Cao[1,2], Tongchao Nan[1,2], Lei Ju[3], Hongxiang Zhou[4], and Lingzao Zeng[5]

[1]Yangtze Institute for Conservation and Development, Hohai University, Nanjing, China,
[2]The National Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing, China,
[3]National Demonstration Center for Environment and Planning, College of Geography and Environmental Science, Henan University, Kaifeng, China,
[4]College of Metrology and Measurement Engineering, China Jiliang University, Hangzhou, China,
[5]Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, China.

Key Points:

- Many popular data assimilation methods are constrained by the Gaussian assumption or suffer from low computational efficiency.
- We propose a novel data assimilation method called $\mathrm{DA}_{(\mathrm{DL})}$ based on deep learning.
- $\mathrm{DA}_{(\mathrm{DL})}$ shows promising performance in problems involving non-linearity, high-dimensionality and non-Gaussianity.

Corresponding author: Jiangjiang Zhang, zhangjiangjiang@hhu.edu.cn

Abstract

In hydrological research, data assimilation (DA) is widely used to fuse the information contained in process-based models and observational data to reduce simulation uncertainty. However, many popular DA methods are limited by low computational efficiency or their reliance on the Gaussian assumption. To address these limitations, we propose a novel DA method called $\mathrm{DA}_{(\mathrm{DL})}$, which leverages the capabilities of deep learning (DL) to model nonlinear relationships and recognize complex patterns. $\mathrm{DA}_{(\mathrm{DL})}$ first generates a large volume of training data from the prior ensemble, and then trains a DL model to update the system knowledge (e.g., model parameters in this study) from multiple predictors. For highly nonlinear models, an iterative form of $\mathrm{DA}_{(\mathrm{DL})}$ can be implemented. Additionally, strategies of data augmentation and local updating are proposed to enhance $\mathrm{DA}_{(\mathrm{DL})}$ for problems involving small ensemble size and the equifinality issue, respectively. In two hydrological DA cases involving Gaussian and non-Gaussian distributions, $\mathrm{DA}_{(\mathrm{DL})}$ shows promising performance compared to two ensemble smoother (ES) methods, i.e., $\mathrm{ES}_{(\mathrm{K})}$ and $\mathrm{ES}_{(\mathrm{DL})}$, which respectively apply the Kalman- and DL-based updates. Potential improvements to $\mathrm{DA}_{(\mathrm{DL})}$ can be made by designing better DL model architectures, imposing physical constraints to the training of the DL model, and further updating other important variables like model states, forcings and error terms.

1 Introduction

Effective prediction of hydrological systems generally requires two sources of information. The first source is the mathematical model (e.g., analytical, numerical or data-driven) that embodies our understanding of the hydrological process (Clark et al., 2015; Peel & Blöschl, 2011; Spieler et al., 2020). The second source is the observational data obtained at different spatial/temporal scales using various techniques (Etter et al., 2020; Slater & Binley, 2021; F. Zheng et al., 2018). Due to insufficient knowledge of the system dynamics and related parameters, the model may not adequately represent the hydrological reality, although it can provide spatially and temporally continuous predictions (Liu & Gupta, 2007; Vrugt et al., 2008). On the other hand, the observational data are more consistent with the system behavior (despite some discrepancies due to measurement errors), but are usually sparse in time and/or space and may lack the ability to explain and extrapolate. In general, either source of information is incomplete and should not be used alone to support hypothesis testing and decision-making.

To synergize scientific knowledge (i.e., model) with observations (i.e., data), two different strategies can be implemented. The first strategy uses the data to constrain the model in order to obtain more reliable spatial/temporal simulations and predictions of the hydrological system. This can be achieved by applying data assimilation (DA) to extract information from the observations and update the model state (Evensen, 2009), structural errors (Evensen, 2019; Smith et al., 2008), parameters (Chen & Zhang, 2006), and initial/boundary conditions (Dechant & Moradkhani, 2011), among others. The second strategy, on the other hand, utilizes scientific knowledge to improve data-driven prediction approaches (e.g., machine learning) to make them more explainable and extrapolative. This strategy has gained popularity in recent years across various research fields (Karpatne et al., 2022). Some successful applications include process-guided DL for lake temperature prediction (Read et al., 2019) and physics-informed neural networks for both forward and inverse problems (Q. He et al., 2020; Karniadakis et al., 2021). Our focus here is on the first strategy of hydrological DA. The second strategy is beyond the scope of the present work, interested readers can refer to a recent book on knowledge-guided machine learning edited by Karpatne et al. (2022).

In hydrological predictions, simulation uncertainties are inevitable and typically treated as random variables. DA can be viewed from the Bayesian perspective to quantify these uncertainties. The background knowledge is first represented by a prior distribution and

then updated to posterior probability by assimilating the information contained in the observational data (Carrassi et al., 2018; Law et al., 2015; Reich & Cotter, 2015). Compared to the prior distribution, the posterior is usually less uncertain and better reflects the underlying data-generating process. The quality of a DA approach is affected by two factors: (1) how informative are the data? and (2) how effective is the DA algorithm at extracting the information contained in the data? Under a limited budget, it is essential to make rational decisions about when, where, and what kind of data to collect. This can be achieved through Bayesian experimental design (Tarakanov & Elsheikh, 2020; Thibaut et al., 2022; J. Zhang et al., 2015) or data-worth analysis (Dausman et al., 2010; Wang et al., 2018; Xue et al., 2014). To handle nonlinear and non-Gaussian observation/system models, Markov chain Monte Carlo (MCMC) or particle filter (PF) methods can be used as the suitable DA methods to approximate the posterior, even when its exact form is unknown (Moradkhani et al., 2005; Shi et al., 2023; Vrugt, 2016). However, MCMC and PF can become computationally expensive when dealing with complex problems due to the curse of dimensionality, despite recent advances in improving their simulation efficiency (Pan et al., 2022; Pulido & van Leeuwen, 2019; Reuschen et al., 2021; J. Zhang, Vrugt, et al., 2020). Nevertheless, high-dimensional DA problems can be efficiently implemented if the models are approximately Gaussian. The ensemble Kalman filter (EnKF) is one such DA algorithm that has been extensively used in hydrological science (Evensen, 2009). EnKF is a Monte Carlo approach to the Kalman filter for sequential DA. It represents the system state distribution using an ensemble of state vectors and computes the mean and covariance matrix from this ensemble. When the ensemble size of EnKF is small, its robustness can be improved by artificially increasing the spread of the forecast ensemble through inflation (Bauser et al., 2018) or by considering the spatial decay of correlations through localization (Anderson, 2012). For efficiency and simplicity, all available data can be assimilated in a single global update using the ensemble smoother (ES; van Leeuwen & Evensen, 1996), a popular variant of EnKF. To deal with strongly non-linear models, iteration can be introduced to EnKF and ES, resulting in the iterative EnKF (Gu & Oliver, 2007; Lorentzen & Naevdal, 2011) and ES (Chen & Oliver, 2012; Emerick & Reynolds, 2013). Despite their popularity in various research fields, the performance of EnKF and its variants may still deteriorate when dealing with non-Gaussian problems. Currently, developing DA methods that are both general (i.e., suitable for non-linear, high-dimensional and non-Gaussian problems) and efficient (i.e., without requiring a massive amount of model runs) is an important need in hydrological science.

Essentially, EnKF and its variants work by updating the state from the innovation vector (i.e., the difference between perturbed observation and model prediction). The Kalman gain matrix, based on the first two statistical moments, defines a linear mapping from the innovation vector to the update vector (i.e., the difference between the updated state and the prior state). Thus, these Kalman-based DA methods are restricted by the Gaussian assumption. To relieve this constraint and formulate a more general DA method, we proposed to use deep learning (DL) to construct a non-linear mapping to replace the Kalman gain matrix (J. Zhang, Zheng, et al., 2020). A high volume of training data are generated from the prior ensemble to train the DL model and possible non-Gaussian patterns in the data can be automatically recognized. This new method, called $\text{ES}_{(\text{DL})}$, has shown promising performance in subsurface characterization problems involving high-dimensional and non-Gaussian variables. Latter, Man et al. (2022) applied $\text{ES}_{(\text{DL})}$ to characterize vapor intrusion sites. Xiao et al. (2023) used the DL-based DA method to update future state of geological CO_2 plumes from historical data. Godoy et al. (2022) adopted random forest instead of DL to construct a non-linear mapping to improve EnKF in the estimation of heterogeneous conductivity field. Wang and Yan (2022) introduced multi-fidelity simulation to further improve the efficiency of $\text{ES}_{(\text{DL})}$ for fast DA of subsurface flow problems.

Despite the improved performance of the DL-based DA method over its Kalman-based counterpart, there is still room for further enhancement. In this paper, we introduce a novel DA method called $\text{DA}_{(\text{DL})}$ based on DL. Unlike $\text{ES}_{(\text{DL})}$, which only uses the innovation vector as the predictor, $\text{DA}_{(\text{DL})}$ inputs the prior parameters, model outputs and the innovation

vector simultaneously as predictors of the DL model. This allows $\mathrm{DA}_{(\mathrm{DL})}$ to utilize more information and achieve better updating results than $\mathrm{ES}_{(\mathrm{DL})}$, as demonstrated later in this work. When running the system model takes a long time, only a small ensemble size is usually affordable. To ensure the robustness of $\mathrm{DA}_{(\mathrm{DL})}$ in this situation, we will introduce a simple but effective data argumentation method. In addition, to address the issue of equifinality commonly encountered in hydrological research, we will introduce the local updating approach developed in our previous work (J. Zhang et al., 2018) to $\mathrm{DA}_{(\mathrm{DL})}$.

The rest of this paper is organized as follows. Section 2 presents the theory and implementation details of the new $\mathrm{DA}_{(\mathrm{DL})}$ method. Then, two hydrological cases are used to demonstrate the performance of $\mathrm{DA}_{(\mathrm{DL})}$. Finally, conclusions and discussions are provided in the last section.

2 Methods

In this work, the hydrological system of concern is described as

$$\widetilde{\mathbf{y}} = \mathcal{H}(\mathbf{m}^*, \mathbf{U}) + \epsilon, \tag{1}$$

where $\widetilde{\mathbf{y}}$ signifies a vector of observational data obtained at different times and locations that summarize the responses of the hydrological system $\mathcal{H}$ to external forcings $\mathbf{U}$, $\mathbf{m}^*$ denotes the unknown parameters, and ϵ represents the measurement errors. When a simulator of the hydrological process is available, the data can be modeled as

$$\widetilde{\mathbf{y}} = \mathcal{F}\big(\mathbf{m}^*, \widetilde{\mathbf{U}}, \widetilde{\psi}_0\big) + \mathbf{E}, \tag{2}$$

where $\widetilde{\psi}_0$ signifies the initial states, and $\mathbf{E}$ represents additive errors originated from observational and modeling processes. In hydrological DA, the observational data can either be assimilated sequentially in a filtering problem or used in a batch update with a smoother. Moreover, DA can target not only model parameters but also model state, external forcings, and model errors. In this work, we focus on the parameter estimation problem. Adopting a Bayesian formalism, posterior distribution of the model parameters can be derived as

$$p(\mathbf{m}|\widetilde{\mathbf{y}}) = \frac{p(\mathbf{m})p(\widetilde{\mathbf{y}}|\mathbf{m})}{p(\widetilde{\mathbf{y}})}, \tag{3}$$

where $p(\mathbf{m})$ and $p(\mathbf{m}|\widetilde{\mathbf{y}})$ are the prior and posterior distribution of model parameters, respectively, $p(\widetilde{\mathbf{y}}|\mathbf{m})$ demotes the likelihood function, and $p(\widetilde{\mathbf{y}}) = \int p(\widetilde{\mathbf{y}}|\mathbf{m})p(\mathbf{m})d\mathbf{m}$ is the evidence. For complex hydrological system that involves non-linear processes, analytical form of $p(\mathbf{m}|\widetilde{\mathbf{y}})$ is not available, and Monte Carlo method can be used to provide an approximate estimate.

EnKF and its variants (e.g., ES) use an ensemble of parameters or states to represent uncertainties. From $p(\mathbf{m})$, N_e random samples can be drawn to form the prior parameter ensemble, $\mathbf{M}^0 = \{\mathbf{m}_1^0, ..., \mathbf{m}_{N_\mathrm{e}}^0\}$. Through running $\mathcal{F}(\cdot)$, the ensemble of prior state can be obtained, i.e., $\mathbf{Y}^0 = \{\mathbf{y}_1^0, ..., \mathbf{y}_{N_\mathrm{e}}^0\}$. The Kalman formula can be used to update each sample in $\mathbf{M}^0$ in the following way:

$$\mathbf{m}_i^1 = \mathbf{m}_i^0 + \mathbf{C}_{\mathrm{MY}}^0\big(\mathbf{C}_{\mathrm{YY}}^0 + \mathbf{R}\big)^{-1}\big(\widetilde{\mathbf{y}} + \epsilon_i - \mathbf{y}_i^0\big), \tag{4}$$

where $i = 1, ..., N_\mathrm{e}$, $\mathbf{M}^1 = \{\mathbf{m}_1^1, ..., \mathbf{m}_{N_\mathrm{e}}^1\}$ is the updated parameter ensemble, $\mathbf{C}_{\mathrm{MY}}^0$ is the the cross-covariance between $\mathbf{M}^0$ and $\mathbf{Y}^0$, $\mathbf{C}_{\mathrm{YY}}^0$ is the auto-covariance of $\mathbf{Y}^0$, $\mathbf{R}$ is the covariance of measurement errors, and ϵ_i is a random realization of measurement errors. Equation (4) describes an update from the innovation vector, $\Delta\mathbf{y}_i = \widetilde{\mathbf{y}} + \epsilon_i - \mathbf{y}_i^0$, to the update vector, $\Delta\mathbf{m}_i = \mathbf{m}_i^1 - \mathbf{m}_i^0$:

$$\Delta\mathbf{m}_i = \mathcal{M}_\mathrm{K}(\Delta\mathbf{y}_i), \tag{5}$$

where $\mathcal{M}_\mathrm{K}(\cdot)$ is a mapping defined by the Kalman gain matrix, $\mathbf{K} = \mathbf{C}_{\mathrm{MY}}^0(\mathbf{C}_{\mathrm{YY}}^0 + \mathbf{R})^{-1}$. It is clear that this mapping is linear and depends on the Gaussian assumption. As shown in our

previous work (J. Zhang, Zheng, et al., 2020), this Kalman-based DA method, called $\text{ES}_{(\text{K})}$ for convenience, cannot obtain reliable results in hydrological DA that involves non-Gaussian distributions.

To address this issue, we formulated a non-linear mapping with DL. From $\mathbf{M}^0$ and $\mathbf{Y}^0$, we can randomly select two samples without repetition and calculate the differences to obtain $\left\{\Delta\mathbf{m}_{ij} = \mathbf{m}_i^0 - \mathbf{m}_j^0, \Delta\mathbf{y}_{ij} = \mathbf{y}_i^0 + \boldsymbol{\epsilon}_{ij} - \mathbf{y}_j^0\right\}, i = 1, .., N_\text{e} - 1, i < j \leq N_\text{e}$. The number of combinations is $N_\text{e}(N_\text{e} - 1)/2$. By feeding these training data to a properly designed DL model, we can obtain a non-linear mapping

$$\Delta\mathbf{m}_i = \mathcal{M}_\text{DL}(\Delta\mathbf{y}_i), \tag{6}$$

and recognize complex patterns like non-Gaussian distribution contained in the data. This DL-based method, known as $\text{ES}_{(\text{DL})}$, performs similarly as $\text{ES}_{(\text{K})}$ under the Gaussian condition. However, in non-Gaussian cases, $\text{ES}_{(\text{DL})}$ can produce more reliable results (J. Zhang, Zheng, et al., 2020).

Nevertheless, in $\text{ES}_{(\text{DL})}$, the starting points of $\Delta\mathbf{m}$ and $\Delta\mathbf{y}$, i.e., $\mathbf{m}^0$ and $\mathbf{y}^0$, are not utilized in both training and inference of the DL model, leaving room for potential improvement. For a DL model, if more relevant predictors can be treated as the inputs, the target should be better identified. Based on this idea, we propose in this work a more powerful DA method than $\text{ES}_{(\text{DL})}$, which is called $\text{DA}_{(\text{DL})}$. This new method constructs a non-linear mapping with three predictors, i.e.,

$$\Delta\mathbf{m}_i = \mathcal{M}_\text{DL}\left(\Delta\mathbf{y}_i, \mathbf{m}_i^0, \mathbf{y}_i^0\right), \tag{7}$$

and each sample in the updated ensemble can be obtained as, $\mathbf{m}_i^1 = \Delta\mathbf{m}_i + \mathbf{m}_i^0, i = 1, ..., N_\text{e}$. This allows $\text{DA}_{(\text{DL})}$ to transcend the limitations of ES and become more versatile and adaptable. In addition to the choice of predictors, the structure of the DL model also significantly impacts the assimilation result. As the design space of a DL model is infinite, it is impossible to identify the optimal architecture. After comparing popular DL models such as DenseNet (Huang et al., 2017), ResNet (K. He et al., 2016) and U-Net (Ronneberger et al., 2015) in multiple problems that involve both Gaussian and non-Gaussian distributions, it is found that models with a specific encoder-decoder architecture can generally produce satisfying results for both $\text{ES}_{(\text{DL})}$ and $\text{DA}_{(\text{DL})}$. The encoder-decoder architecture consists of two sub-networks: an encoder that compresses the input into a smaller spatial representation with more channels, and a decoder that expands the spatial dimensions while reducing the number of channels.

In highly non-linear DA problems, one single update using $\text{ES}_{(\text{K})}$, $\text{ES}_{(\text{DL})}$, or $\text{DA}_{(\text{DL})}$ may not be sufficient. In this situation, it is suggested to assimilate the observational data multiple times (Emerick & Reynolds, 2013). To guarantee that the updating results are reasonable, random realizations of measurement errors generated in iteration t should be inflated by a factor of β_t, where $t = 1, ..., N_\text{iter}$, N_iter is the number of iterations, and $\sum_{t=1}^{N_\text{iter}} 1/\beta_t^2 = 1$. Then the updating schemes of $\text{ES}_{(\text{K})}$, $\text{ES}_{(\text{DL})}$ and $\text{DA}_{(\text{DL})}$ become

$$\mathbf{m}_i^t = \mathbf{m}_i^{t-1} + \mathbf{C}_\text{MY}^{t-1}\left(\mathbf{C}_\text{YY}^{t-1} + \beta_t^2\mathbf{R}\right)^{-1}\left(\widetilde{\mathbf{y}} + \beta_t\boldsymbol{\epsilon}_i - \mathbf{y}_i^{t-1}\right), \tag{8}$$

$$\mathbf{m}_i^t = \mathbf{m}_i^{t-1} + \mathcal{M}_\text{DL}\left(\widetilde{\mathbf{y}} + \beta_t\boldsymbol{\epsilon}_i - \mathbf{y}_i^{t-1}\right), \tag{9}$$

$$\mathbf{m}_i^t = \mathbf{m}_i^{t-1} + \mathcal{M}_\text{DL}\left(\widetilde{\mathbf{y}} + \beta_t\boldsymbol{\epsilon}_i - \mathbf{y}_i^{t-1}, \mathbf{m}_i^{t-1}, \mathbf{y}_i^{t-1}\right), \tag{10}$$

respectively. Finally, we use $\mathbf{M}^{N_\text{iter}} = \left\{\mathbf{m}_1^{N_\text{iter}}, ..., \mathbf{m}_{N_\text{e}}^{N_\text{iter}}\right\}$ to approximate the posterior distribution of model parameters.

In many situations, evaluating the system model $\mathcal{F}(\cdot)$ can be computationally intensive. As a result, a small ensemble size N_e is often used. Even though the number of training data fed to the DL model, i.e., $N_\text{e}(N_\text{e}-1)/2$, is much larger than N_e, it may still not be enough for training a data-hungry DL model. To address this issue, a simple yet effective data argumentation method is proposed. In the tth iteration, $t = 1, ..., N_\text{iter}$, when using $\left\{\mathbf{m}_i^{t-1}, \mathbf{y}_i^{t-1}\right\}$

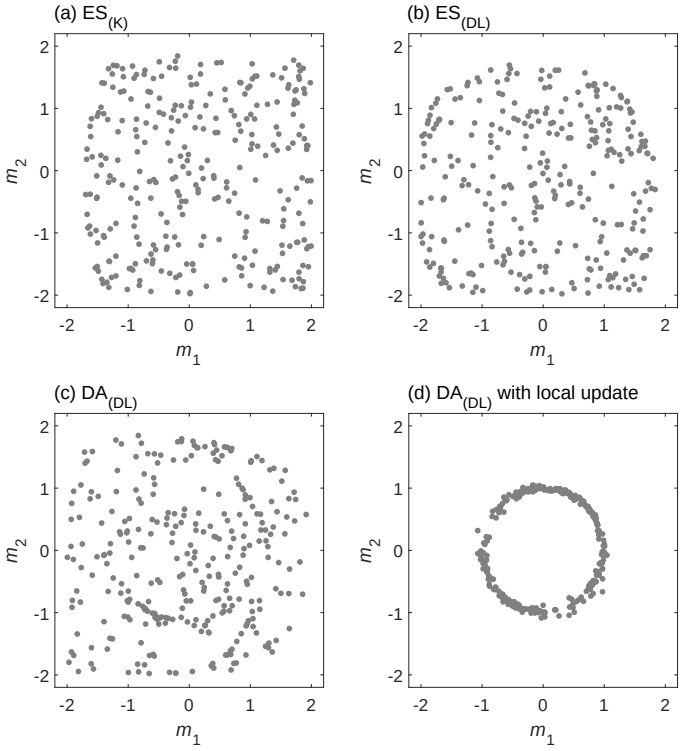

Figure 1. Posterior samples of model parameters obtained by (a) $\mathrm{ES}_{(\mathrm{K})}$, (b) $\mathrm{ES}_{(\mathrm{DL})}$, (c) $\mathrm{DA}_{(\mathrm{DL})}$, and (d) $\mathrm{DA}_{(\mathrm{DL})}$ with local update, respectively.

and $\{\mathbf{m}_j^{t-1}, \mathbf{y}_j^{t-1}\}$ to generate the training datum, $\{\Delta\mathbf{y}_{ij}, \mathbf{m}_j^{t-1}, \mathbf{y}_j^{t-1} \rightarrow \Delta\mathbf{m}_{ij}\}$, we can produce $M \geq 1$ similar copies by adding different random realizations of measurement errors to $\Delta\mathbf{y}_{ij}$, i.e., $\{\Delta\mathbf{y}_{ij,1}, \mathbf{m}_j^{t-1}, \mathbf{y}_j^{t-1} \rightarrow \Delta\mathbf{m}_{ij}\}, ..., \{\Delta\mathbf{y}_{ij,M}, \mathbf{m}_j^{t-1}, \mathbf{y}_j^{t} \rightarrow \Delta\mathbf{m}_{ij}\}$. Finally, a training data set with $M^* N_\mathrm{e}(N\mathrm{e}-1)/2$ samples can be obtained. This way of expanding the training data set can reduce over-fitting and make the DL model more generalized. Performance of this data augmentation method will be demonstrated in Section 3.1.

In hydrological simulations, one major challenge for many DA methods is the equifinality issue, where multiple parameter sets with significantly different values can produce equally good performance. From the Bayesian perspective, it means that the posterior distribution of model parameters is multi-modal. To improve the performance of $\mathrm{DA}_{(\mathrm{DL})}$ in multi-modal DA problems, the local updating approach proposed in our previous work (J. Zhang et al., 2018) can be introduced. The idea behind the local updating approach is straightforward: although globally the distribution is multi-modal, locally it is still approximately single-modal. Based on this idea, the local ensemble of $\mathbf{m}_i^{t-1}(i = 1, ..., N_\mathrm{e}, t = 1, ..., N_\mathrm{iter})$ can be obtained based on the following measure:

$$J(\mathbf{m}) = J_1(\mathbf{m})/J_1^{\max} + J_2(\mathbf{m})/J_2^{\max}, \tag{11}$$

where $J_1(\mathbf{m}) = \left[\mathcal{F}(\mathbf{m}) - \widetilde{\mathbf{y}}\right]^{\mathrm{T}} \mathbf{R}^{-1} \left[\mathcal{F}(\mathbf{m}) - \widetilde{\mathbf{y}}\right]$, $J_2(\mathbf{m}) = \left(\mathbf{m} - \mathbf{m}_i^{t-1}\right)^{\mathrm{T}} \mathbf{C}_{\mathrm{MM}}^{-1} \left(\mathbf{m} - \mathbf{m}_i^{t-1}\right)$, $\mathbf{C}_{\mathrm{MM}}$ is the auto-covariance matrix of model parameters, $J_1^{\max}$ and $J_2^{\max}$ are the maximum values of $J_1(\mathbf{m})$ and $J_2(\mathbf{m})$, respectively. The local ensemble of $\mathbf{m}_i^{t-1}$ is $\mathbf{M}_{i,\mathrm{local}}^{t-1} = \{\mathbf{m}_{i,1}^{t-1}, ..., \mathbf{m}_{i,N_\mathrm{l}}^{t-1}\}$, the samples in $\mathbf{M}^{t-1}$ with the $N_\mathrm{l} = \alpha N_\mathrm{e}(\alpha \in (0, 1])$ smallest J values. Using $\mathrm{DA}_{(\mathrm{DL})}$, we can obtain the updated local ensemble, $\mathbf{M}_{i,\mathrm{local}}^{t}$, from which a random sample, $\mathbf{m}_i^t$, can be drawn as the updated sample of $\mathbf{m}_i^{t-1}$. For more details about the local updating approach and its application to $\mathrm{ES}_{(\mathrm{K})}$, one can refer to (J. Zhang et al., 2018). When implementing $\mathrm{DA}_{(\mathrm{DL})}$ with the local updating approach, the DL model should be

trained N_e times in each iteration. If each time the DL model is trained from scratch, it will require a lot of time and computational resources. To speed up this process, a transfer learning approach can be applied. From the entire prior ensemble, we first train a basic DL model. Then this trained model is updated with new data obtained from each local ensemble. This fine-tuning process is usually very easy and fast.

To demonstrate the effectiveness of the local updating approach, we conduct a simple test on a problem with a multi-modal posterior. The system model considered in this problem is described by $y = m_1^2 + m_2^2$, with both m_1 and m_2 following a uniform prior distribution of $\mathcal{U}(-2, 2)$. To infer the posterior distribution of $\mathbf{m} = \{m_1, m_2\}$, we utilize a single observation of $\widetilde{y} = 1$, with error that follows a Gaussian distribution, $\epsilon \sim \mathcal{N}(0, 0.01^2)$. It is noteworthy that an infinite set of parameter combinations can accurately fit the observation due to the problem's nature, resulting in a posterior distribution that takes on a circular shape. Here, four DA methods with $N_e = 300$ and $N_{\text{iter}} = 2$ are implemented, i.e., $\text{ES}_{(\text{K})}$, $\text{ES}_{(\text{DL})}$, $\text{DA}_{(\text{DL})}$, and $\text{DA}_{(\text{DL})}$ with local update ($\alpha = 0.1$). Both $\text{ES}_{(\text{DL})}$ and $\text{DA}_{(\text{DL})}$ employ a feedforward neural network (FNN) with one hidden layer that has 10 nodes. In $\text{DA}_{(\text{DL})}$, we use $\{\Delta y, y^{t-1}\}$ as inputs to the FNN model instead of $\{\Delta y, \mathbf{m}^{t-1}, y^{t-1}\}$ due to differences in the dimensions of $\mathbf{m}$ and y. DL models are capable of merging predictors with different dimensions, which will be demonstrated in the succeeding section. As shown in Figures 1, introducing the local updating approach to $\text{DA}_{(\text{DL})}$ enables this method to successfully identify the circle-like posterior. The average root-mean-square error (RMSE) between model simulations and $\widetilde{y}$ is also evaluated, with values of 2.34, 2.15, 1.86 and 0.073 for $\text{ES}_{(\text{K})}$, $\text{ES}_{(\text{DL})}$, $\text{DA}_{(\text{DL})}$, and $\text{DA}_{(\text{DL})}$ with local update, respectively.

3 Case Studies

3.1 The Non-Gaussian Condition

In this section, we compare the performance of $\text{DA}_{(\text{DL})}$ with $\text{ES}_{(\text{K})}$ and $\text{ES}_{(\text{DL})}$ in a non-Gaussian setting. For this purpose, we simulate transient groundwater flow in a confined, channelized aquifer. The domain size is 800 (L) × 800 (L), with impervious upper and lower boundaries, as well as two constant-head boundaries at the left (202 L) and right (198 L) sides. The initial hydraulic head is set to 198 (L) everywhere except for the left boundary, where it is prescribed. The system includes an injection well with a rate of 150 (L^3T^{-1}) and a pumping well with a rate of -150 (L^3T^{-1}), which can enhance water flow within the domain. The channelized field comprises two materials with distinct hydraulic conductivities: $\mathcal{K}_1 = 0.5$ (LT^{-1}) and $\mathcal{K}_2 = 2.3$ (LT^{-1}). With the above settings, we can obtain transient hydraulic heads $h(\mathbf{x}, t)$ at different locations and times by solving

$$S_s \frac{\partial h(\mathbf{x}, t)}{\partial t} + \nabla \cdot \mathbf{q}(\mathbf{x}, t) = g(\mathbf{x}, t) \tag{12}$$

with MODFLOW (Harbaugh et al., 2000), where S_s (L^{-1}) represents specific storage, $\mathbf{x}$ (L) denotes location, t (T) is time, $\mathbf{q}(\mathbf{x}, t) = -\mathcal{K}(\mathbf{x})\nabla h(\mathbf{x}, t)$ signifies the water flux, ∇ is the nabla operator, and $g(\mathbf{x}, t)$ (T^{-1}) denotes the source or sink term. For this model, we uniformly divide the domain into 41×41 grids, set the simulation time to be 18 (T), and use $S_s = 0.0001$ (L^{-1}).

In this case, the spatial distribution of $\mathcal{K}$ is unknown and should be inferred from indirect observations. As depicted in Figure 2(b), the reference field of $\mathcal{K}$ is non-Gaussian, making it challenging to be accurately estimated. To accomplish this, hydraulic head measurements are collected from 7×7 wells at $t = \{0.6, 1.2, ..., 6.0\}$ (T) with errors that fit $\epsilon \sim \mathcal{N}(0, 0.01^2)$. For the three DA methods, a same set of prior parameter ensemble with $N_e = 499$ samples are generated with the direct sampling method proposed by Mariethoz et al. (2010). The training image featured in Figure 2(a) serves as the basis for generating these samples. As the level of non-linearity in this problem is relatively low, we only set up one iteration for each of the three DA methods.

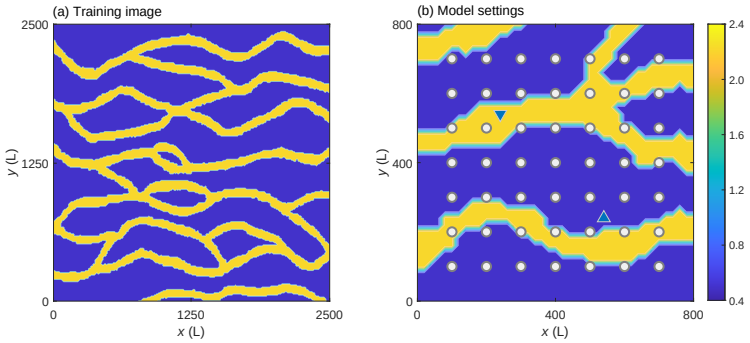

Figure 2. (a) The training image used to generate random realizations of $\mathcal{K}$ field using the direct sampling method; (b) The reference $\mathcal{K}$ field, injection well (the down triangle), pumping well (the up triangle), and measurement locations (the circles).

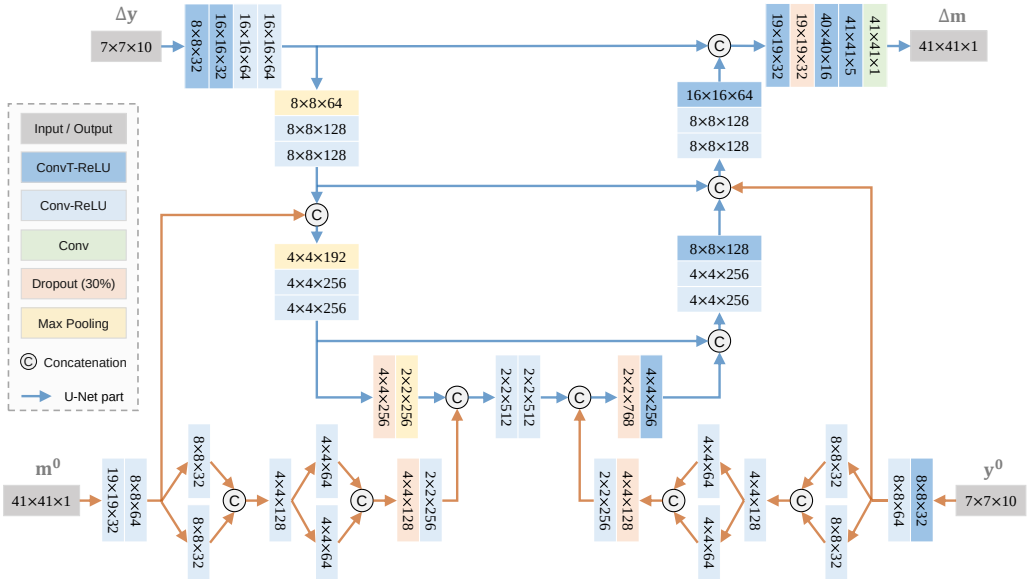

Figure 3. Architecture of the DL model used by $\mathrm{DA}_{(\mathrm{DL})}$ in the non-Gaussian case. Here, the part with blue arrows is the U-Net model used by $\mathrm{ES}_{(\mathrm{DL})}$. Output size of each layer is indicated by height×width×channels. Conv and ConvT mean 2-D convolution layer and transposed 2-D convolution layer, respectively.

Figure 3 depicts the use of a DL model by $\text{DA}_{(\text{DL})}$ with three predictors, namely $\Delta\mathbf{y}$, $\mathbf{m}^0$, and $\mathbf{y}^0$, and one target, namely $\Delta\mathbf{m}$. Dimensions of the input variables $\Delta\mathbf{y}$ and $\mathbf{y}^0$ are both 7×7×10, indicating the presence of 7×7 measurement wells and 10 sampling times. Dimensions of the input variable $\mathbf{m}^0$ and the target variable $\Delta\mathbf{m}$ are both 41×41×1, representing 41×41 model grids. The DL model consists of the U-Net part (blue arrows) that inputs $\Delta\mathbf{y}$ and the extra parts (brown arrows) that input $\mathbf{m}^0$ and $\mathbf{y}^0$. The U-Net part is composed of two pathways, an encoder path on the left and a decoder path on the right. For the input variable $\Delta\mathbf{y}$, we first utilize transposed 2-D convolution (ConvT) and 2-D convolution (Conv) to extend the spatial dimensions from 7×7 to 16×16 and simultaneously increase the channel number from 10 to 64. The non-linear activation function of rectified linear unit (ReLU) is used after ConvT and Conv. As feature maps move through the encoder path, spatial dimensions are progressively reduced while channel numbers increase, utilizing layer types such as Conv, ReLU, Max Pooling and Dropout (with 30% probability). This continues until the feature size reaches 2×2×512. The decoder path, on the other hand, expands the spatial dimensions and reduces the number of channels until the feature reaches a size of 16×16×64. For the purpose of producing finer-grained predictions, skip connections are employed in the U-Net structure to facilitate direct forwarding of feature maps from the encoder to the decoder pathway. To incorporate the information contained in $\mathbf{m}^0$ and $\mathbf{y}^0$, the extra parts with brown arrows transform each of these inputs into two output features with sizes of 8×8×64 and 2×2×256, which are concatenated with features in the encoder and decoder paths that have the same spatial dimensions, further improving the model's performance. Here, $\text{ES}_{(\text{DL})}$ only uses the U-Net part, i.e., the input $\Delta\mathbf{y}$ "flows" through the blue arrows to the target $\Delta\mathbf{m}$. For both $\text{ES}_{(\text{DL})}$ and $\text{DA}_{(\text{DL})}$, we train the DL models using the Adam optimizer with a constant learning rate of 0.001. During the training process, we implement mini-batches containing 512 samples over the course of 50 epochs. Furthermore, we incorporate a gradient threshold method that clips any gradient values that exceed a threshold of 10, preventing potential instability. To mitigate the risk of over-fitting, we include an L_2 regularization factor of 0.0002 for the weights to the loss function.

As shown in Figures 4(a-c), all the three DA methods can capture the non-Gaussian feature in the spatial distribution of $\mathcal{K}$ to varying degrees. However, $\text{ES}_{(\text{K})}$ fails to reproduce the connectivity feature of the reference $\mathcal{K}$ field, as seen in Figure 2b. As the subsurface media consist of only two distinct materials with values of $\mathcal{K}_1 = 0.5$ and $\mathcal{K}_2 = 2.3$, the histogram of an estimated $\mathcal{K}$ field should display bi-modality. Nonetheless, $\text{ES}_{(\text{K})}$ is unable to identify this bi-modality, as illustrated in Figure 4(d). These results confirm that the Kalman-based update method $\text{ES}_{(\text{K})}$ is inadequate in solving non-Gaussian DA problems.

By introducing a non-linear updating scheme with DL, $\text{ES}_{(\text{DL})}$ can better estimate the spatial distribution of $\mathcal{K}$ (Figure 4b) over its Kalman counterpart, and the bi-modality feature can be identified (Figure 4e). Additionally, the standard deviation (std) field as shown in Figure 4(h) also reveals the connectivity feature of $\mathcal{K}$, with larger std values at the interface of the two materials. When we use three predictors $\{\Delta\mathbf{y}, \mathbf{m}^0, \mathbf{y}^0\}$ in $\text{DA}_{(\text{DL})}$ to infer the update vector $\Delta\mathbf{m}$, significant improvement in overall performance can be achieved compared to $\text{ES}_{(\text{DL})}$: the estimated mean field shows a clearer connectivity feature (Figure 4c), the bi-modality distribution is better identified (Figure 4f), and the std field has smaller values (Figure 4i). To conduct more comprehensive comparisons, we perform eight repetitive runs for each of the three DA methods. As shown in Figure 5, $\text{DA}_{(\text{DL})}$ generally provides a more accurate estimation of both the $\mathcal{K}$ field and a better data-match. The average RMSE values between model simulations and measurements calculated from the updated ensembles are 1.44, 1.02, and 0.89 for $\text{ES}_{(\text{K})}$, $\text{ES}_{(\text{DL})}$, and $\text{DA}_{(\text{DL})}$, respectively. The average RMSE values between the estimated and reference $\mathcal{K}$ fields for the three methods are 0.69, 0.66, and 0.51, respectively. These results suggest that $\text{DA}_{(\text{DL})}$ is superior to both $\text{ES}_{(\text{K})}$ and $\text{ES}_{(\text{DL})}$ in solving non-Gaussian DA problems.

When a simulator requires significant computational resources, only a limited number of model simulations can be afforded. In this situation, we propose using the data augmentation

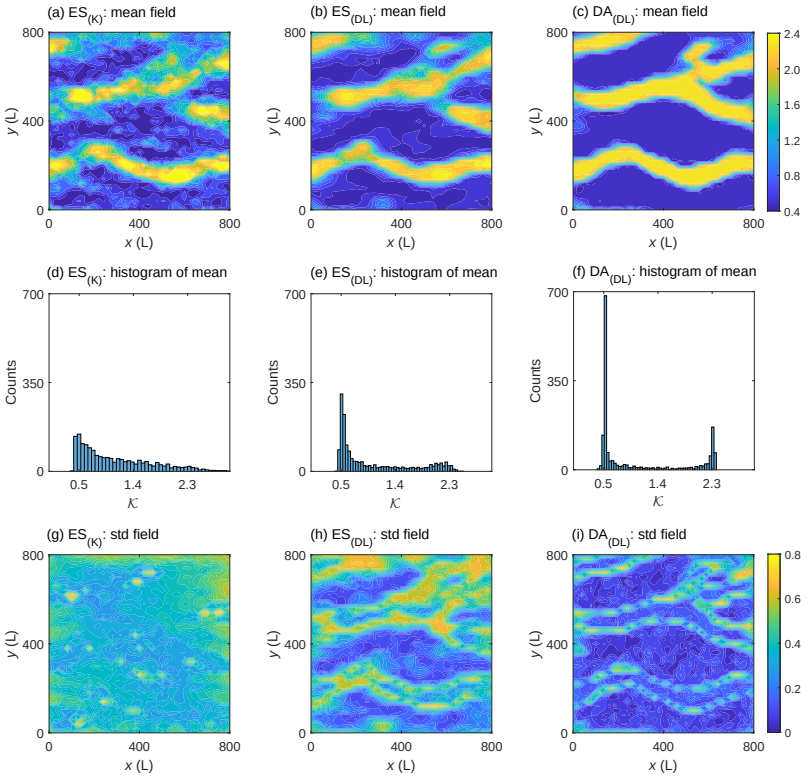

Figure 4. (a-c) Mean fields, (d-f) histograms of the mean fields, and (g-i) standard deviation (std) fields of $\mathcal{K}$ estimated by $\mathrm{ES}_{(\mathrm{K})}$ (left column), $\mathrm{ES}_{(\mathrm{DL})}$ (middle column), and $\mathrm{DA}_{(\mathrm{DL})}$ (right column), respectively.

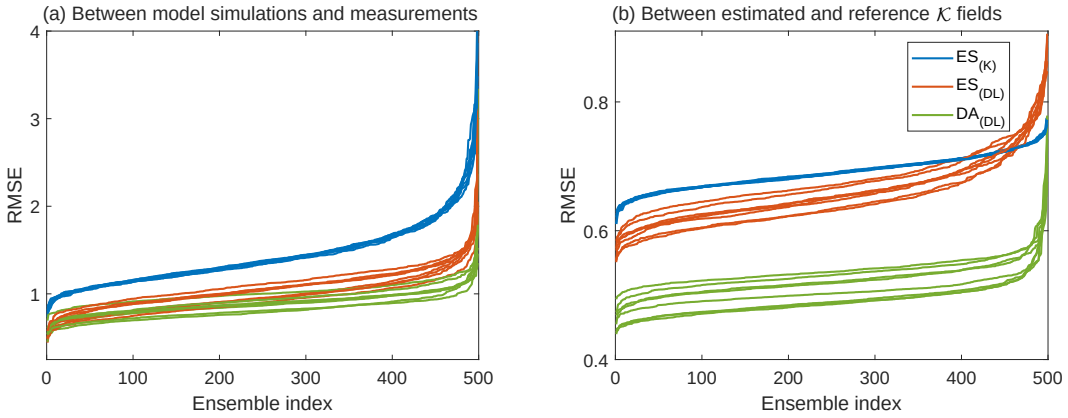

Figure 5. Root-mean-square errors (RMSEs) (a) between model simulations and measurements and (b) between estimated and reference $\mathcal{K}$ fields calculated from the updated ensembles obtained by $\mathrm{ES}_{(\mathrm{K})}$, $\mathrm{ES}_{(\mathrm{DL})}$, and $\mathrm{DA}_{(\mathrm{DL})}$, respectively. The RMSE values are sorted in ascending order. There are eight repetition runs for each method.

method as described in Section 2 in $\mathrm{DA}_{(\mathrm{DL})}$. Specifically, we set $N_\mathrm{e} = 50$ and $M = 20$ and evaluate the performance of $\mathrm{ES}_{(\mathrm{K})}$, $\mathrm{ES}_{(\mathrm{DL})}$, $\mathrm{DA}_{(\mathrm{DL})}$ without data augmentation, and $\mathrm{DA}_{(\mathrm{DL})}$ with data augmentation. Each method is implemented eight times without changing other settings. The four methods yield average RMSE values of 2.31, 3.02, 2.96, and 2.16 for the match of measurements, and average RMSE values of 1.25, 0.97, 0.96, and 0.83 for the match of reference $\mathcal{K}$ field, respectively. It is evident that a significant reduction in ensemble size results in degraded estimation outcomes for all the above DA methods. While $\mathrm{ES}_{(\mathrm{K})}$ yields the worst estimation of $\mathcal{K}$ (RMSE: 1.25), its data-matching result (RMSE: 2.31) is still superior to $\mathrm{ES}_{(\mathrm{DL})}$ (RMSE: 3.02) and $\mathrm{DA}_{(\mathrm{DL})}$ without data augmentation (RMSE: 2.96), which are based on training data sets each with only 1,225 samples. Introducing the data augmentation method to $\mathrm{DA}_{(\mathrm{DL})}$ can increase the number of training data points to 24,500, producing the optimal performance among the four DA methods. In practical applications, it is crucial to strike a balance between the choice of M and the computational resources needed for the simulation of system model and the training of the DL model.

3.2 The Gaussian Condition

In the previous section, we demonstrated that $\mathrm{DA}_{(\mathrm{DL})}$ outperforms two ES methods that rely on the Kalman- and DL-based updates in estimating high-dimensional and non-Gaussian distributed parameters of a groundwater model. To further examine the effectiveness of $\mathrm{DA}_{(\mathrm{DL})}$, we now focus on its performance under the Gaussian condition, where $\mathrm{ES}_{(\mathrm{K})}$ is typically expected to excel. If $\mathrm{DA}_{(\mathrm{DL})}$ can produce similar (or even better) outcomes to $\mathrm{ES}_{(\mathrm{K})}$, it would confirm the versatility and practicality of $\mathrm{DA}_{(\mathrm{DL})}$ as a reliable DA method. Specifically, we explore the joint estimation of contaminant source parameters and heterogeneous conductivity ($\mathcal{K}$) field in the subsurface media, for which the posterior distribution of these parameters is roughly multi-Gaussian.

In this case, we simulate steady-state groundwater flow and contaminant transport in a 2-D confined aquifer. The size of the domain is 20 (L) $\times$ 10 (L), with impervious upper and lower boundaries and constant-head boundaries at the left (12 L) and right (11 L) sides. The domain is discretized into 81$\times$41 grids in the numerical model. The $\mathcal{K}$ field is heterogeneous and its logarithmic transformation ($\mathcal{Y} = \log \mathcal{K}$) is Gaussian distributed. To characterize the spatial correlation of $\mathcal{Y}$ at any two locations $\{x_1, y_1\}$ and $\{x_2, y_2\}$, we adopt the following function:

$$C_\mathcal{Y}(x_1, y_1; x_2, y_2) = \sigma_\mathcal{Y}^2 \exp\left(-\frac{|x_1 - x_2|}{\lambda_x} - \frac{|y_1 - y_2|}{\lambda_y}\right), \tag{13}$$

where $\sigma_\mathcal{Y}^2$ represents variance of $\mathcal{Y}$, λ_x and λ_y denote correlation lengths in the x and y direction, respectively. The steady-state hydraulic heads (h) and water velocity (v_i) within the domain can be determined by utilizing MODFLOW, which involves solving the following equations (Harbaugh et al., 2000):

$$\frac{\partial}{\partial x_i}\left(\mathcal{K}_i \frac{\partial h}{\partial x_i}\right) = 0, \tag{14}$$

and

$$v_i = -\frac{\mathcal{K}_i}{\theta}\frac{\partial h}{\partial x_i}, \tag{15}$$

where θ (-) represents the porosity of the subsurface media, and the subscript i denotes the coordinate axis (1 for the x direction and 2 for the y direction).

The subsurface flow field contains a contaminant source with unknown location and release strengths that change over time. To determine the concentrations (C) at different times in the domain, we use MT3DMS (C. Zheng & Wang, 1999) to solve the following mass-balance equation:

$$\frac{\partial(\theta C)}{\partial t} = \frac{\partial}{\partial x_i}\left(\theta D_{ij}\frac{\partial C}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(\theta v_i C) + q_\mathrm{a} C_\mathrm{s}, \tag{16}$$

Table 1. Prior distributions and reference values of the eight contaminant source parameters.

Parameter	x_s	y_s	S_1	S_2	S_3	S_4	S_5	S_6
Prior distribution	$\mathcal{U}(3,5)$	$\mathcal{U}(4,6)$	$\mathcal{U}(0,8)$	$\mathcal{U}(0,8)$	$\mathcal{U}(0,8)$	$\mathcal{U}(0,8)$	$\mathcal{U}(0,8)$	$\mathcal{U}(0,8)$
Reference value	3.52	4.44	5.69	7.88	6.31	1.49	6.87	5.55

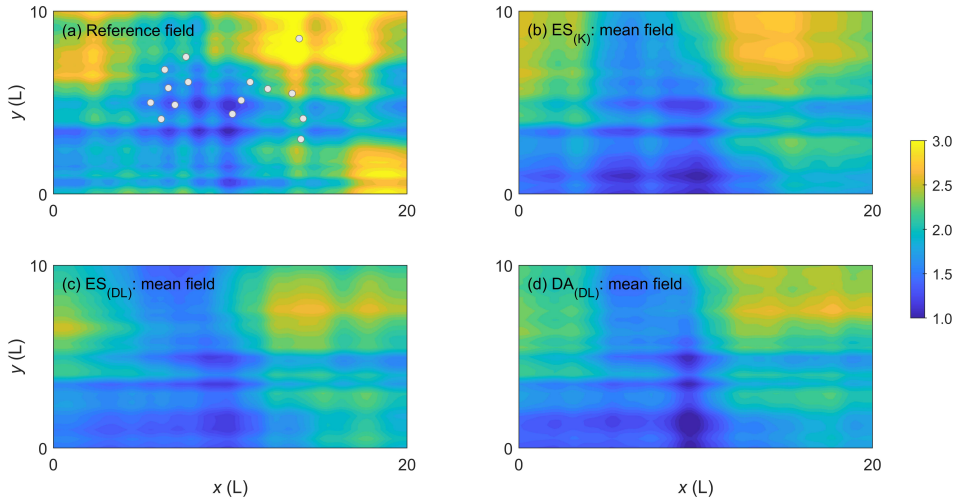

Figure 6. (a) The reference $\mathcal{Y}$ field and measurement well locations (white circles); (b-d) The estimated mean fields of $\mathcal{Y}$ (averaged over five repetition runs) obtained by $\mathrm{ES}_{(\mathrm{K})}$, $\mathrm{ES}_{(\mathrm{DL})}$ and $\mathrm{DA}_{(\mathrm{DL})}$, respectively.

where t (T) represents time, q_a (T^{-1}) is volumetric flow rate per unit volume of the aquifer, C_s (ML^{-1}) represents concentration of the source, and D_{ij} denotes the hydrodynamic dispersion coefficient with the following components:

$$D_{11} = \frac{1}{\sqrt{v_1^2 + v_2^2}} \left(\alpha_{\mathrm{L}} v_1^2 + \alpha_{\mathrm{T}} v_2^2\right), \tag{17}$$

$$D_{22} = \frac{1}{\sqrt{v_1^2 + v_2^2}} \left(\alpha_{\mathrm{L}} v_2^2 + \alpha_{\mathrm{T}} v_1^2\right), \tag{18}$$

$$D_{12} = D_{21} = \frac{1}{\sqrt{v_1^2 + v_2^2}} \left(\alpha_{\mathrm{L}} - \alpha_{\mathrm{T}}\right) v_1 v_2, \tag{19}$$

where α_{L} and α_{T} (L) represents the longitudinal and transverse dispersity, respectively. In this problem, we aim to identify the unknown conductivity field and contaminant source using measurements of hydraulic head and solute concentrations.

To reduce the dimensionality of the problem, we utilize the Karhunen-Loève (KL) expansion (D. Zhang & Lu, 2004) to approximate the random field of $\mathcal{Y}$ as follows:

$$\widetilde{\mathcal{Y}}(\mathbf{x}) = \mu_{\mathcal{Y}} + \sum_{n=1}^{N_{\mathrm{KL}}} \sqrt{\tau_n} s_n(\mathbf{x}) \xi_n. \tag{20}$$

Here, $\widetilde{\mathcal{Y}}(\mathbf{x})$ represents the reconstructed value of $\mathcal{Y}$ at location $\mathbf{x} = \{x, y\}$, where $\mu_{\mathcal{Y}}$ is the mean. The eigenvalues and eigenfunctions of the covariance defined by equation (13) are represented by τ_n and $s_n(\mathbf{x})$, respectively. The uncertainty in the field is represented

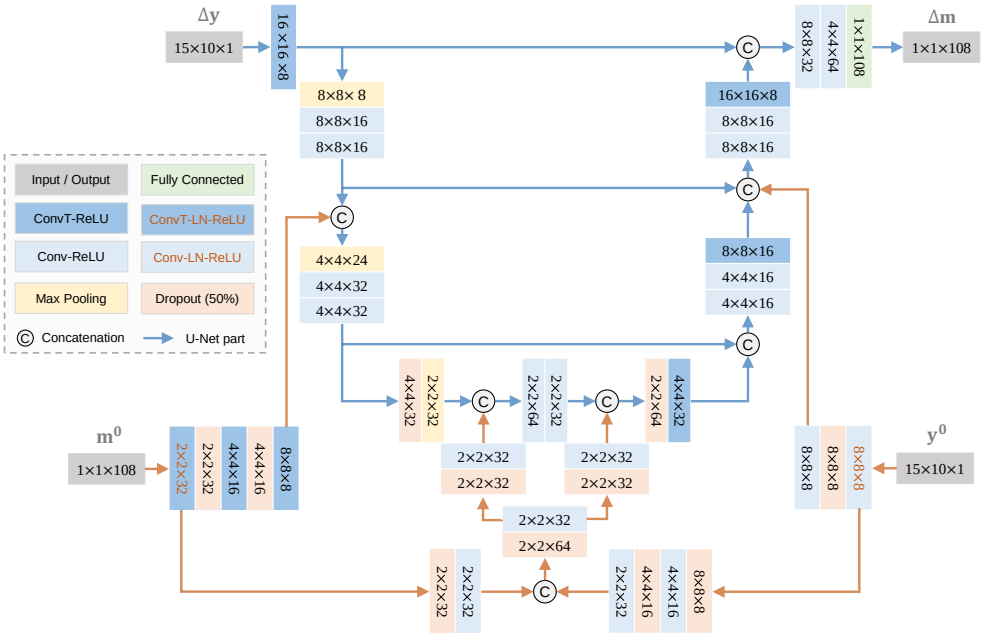

Figure 7. Architecture of the DL model used by $\mathrm{DA}_{(\mathrm{DL})}$ in the Gaussian case. Here, the part with blue arrows represents the U-Net model used by $\mathrm{ES}_{(\mathrm{DL})}$. Size of each layer is indicated by height×width×channels. Conv, ConvT and LN mean 2-D convolution layer, transposed 2-D convolution layer, and layer-normalization layer, respectively.

by independent Gaussian random coefficients $\xi_n \sim \mathcal{N}(0,1)$, where $n = 1, \ldots, N_{\mathrm{KL}}$. To retain 95% variance of the original field, 100 terms ($N_{\mathrm{KL}} = 100$) should be included: $\sum_{n=1}^{100} \tau_n / \sum_{n=1}^{\infty} \tau_n \approx 0.95$. The contaminant source is described by eight parameters: location $\{x_{\mathrm{s}}, y_{\mathrm{s}}\}$ and six mass-loading rates S_k (MT^{-1}) during time interval $[k, k+1]$ (T), where $k = 1, \ldots, 6$. The prior distributions of the eight parameters are all uniform, with ranges listed in Table 1. In summary, there are 108 unknown parameters to be estimated, i.e., $\{\xi_1, \ldots, \xi_{100}, x_{\mathrm{s}}, y_{\mathrm{s}}, S_1, \ldots, S_6\}$. Additional parameters are determined through geological surveys or experiments, including $\mu_{\mathcal{Y}} = 2$, $\sigma^2_{\mathcal{Y}} = 1$, $\lambda_x = 10$ (L), $\lambda_y = 5$ (L), $\alpha_{\mathrm{L}} = 0.3$ (L), $\alpha_{\mathrm{T}} = 0.03$ (L), and $\theta = 0.25$ (-), respectively. To infer the 108 unknown parameters, measurements of steady-state hydraulic head and contaminant concentrations at $t = \{4, 5, ..., 12\}$ (T) are collected from 15 wells in the domain (represented by the circles in Figure 6a). Both types of measurement errors adhere to the Gaussian distribution, with $\epsilon_{\mathrm{h}} \sim \mathcal{N}(0, 0.005^2)$ and $\epsilon_{\mathrm{c}} \sim \mathcal{N}(0, 0.005^2)$.

In this case, we utilize a DL model resembling the one used in the non-Gaussian problem for $\mathrm{DA}_{(\mathrm{DL})}$, as shown in Figure 7. For the $\mathrm{ES}_{(\mathrm{DL})}$ method, we again use the U-Net part indicated by the blue arrows. We incorporate layer-normalization (LN) layers that normalize data across all channels for each sample independently, enabling efficient training and improved performance. The DL model is trained over a total of 600 epochs using the Adam optimizer with an initial learning rate of 0.006, which decreases to its 80% value every 15 epochs. We set the mini-batch size to 3072, and the L_2 regularization factor for the weights to the loss function is 0.0001.

In this case, we once again compare $\mathrm{ES}_{(\mathrm{K})}$, $\mathrm{ES}_{(\mathrm{DL})}$, and $\mathrm{DA}_{(\mathrm{DL})}$, with $N_{\mathrm{iter}} = 5$ and $N_{\mathrm{e}} = 500$. For each DA method, we conduct five repetition runs, and average the outcomes to obtain the estimated mean $\mathcal{Y}$ fields as shown in Figure 6. All three methods are able to identify the major high- and low-value regions of the reference field (Figure 6a). However,

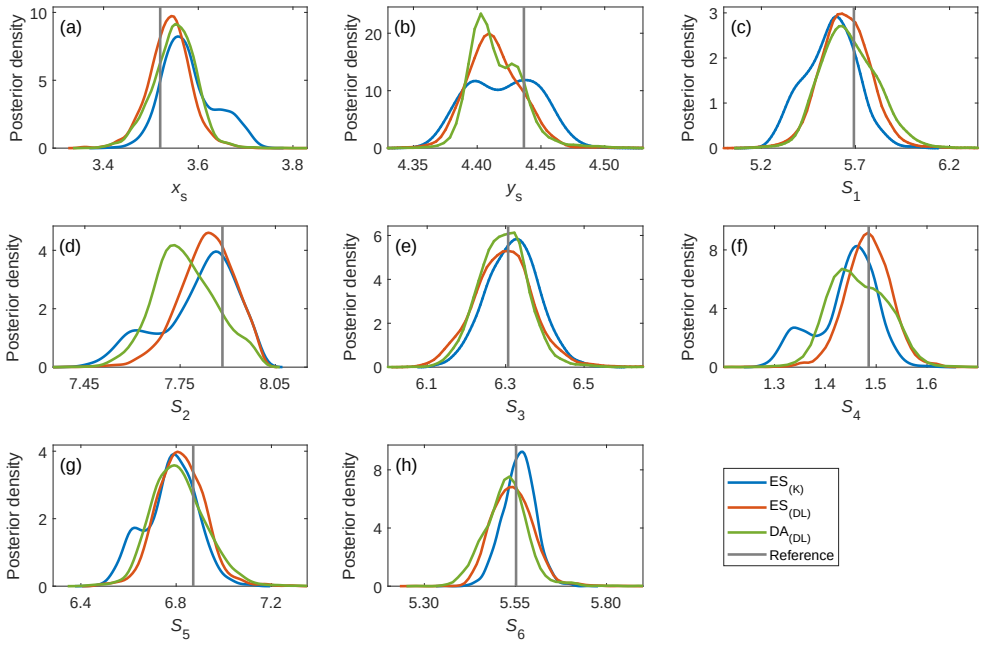

Figure 8. Posterior density curves of the eight contaminant source parameters estimated by $\mathrm{ES}_{(\mathrm{K})}$ (blue lines), $\mathrm{ES}_{(\mathrm{DL})}$ (red lines) and $\mathrm{DA}_{(\mathrm{DL})}$ (green lines), respectively. The reference values are indicated by the black vertical lines. We conduct five repetition runs for each method to obtain the above results.

since the measurements are taken from only 15 wells, which are primarily located in the central region of the flow domain, these DA methods tend to underestimate the high-value regions and overestimate the low-value regions (e.g. the high-value region in the lower bottom corner is not captured). Increasing the number and distribution of measurement wells can improve the precision of the $\mathcal{Y}$ field estimation. The mean RMSE values, calculated between the reference $\mathcal{Y}$ field and the updated ensembles (consisting of 2500 samples from the five repetitions), are 0.4830 ($\mathrm{ES}_{(\mathrm{K})}$), 0.5074 ($\mathrm{ES}_{(\mathrm{DL})}$), and 0.5072 ($\mathrm{DA}_{(\mathrm{DL})}$), with standard deviations of 0.0866, 0.1045 and 0.1167, respectively. In this case, $\mathrm{ES}_{(\mathrm{K})}$ obtains slightly more accurate estimation of the $\mathcal{Y}$ field. Nonetheless, with better designed DL models and training options, $\mathrm{ES}_{(\mathrm{DL})}$ and $\mathrm{DA}_{(\mathrm{DL})}$ have the potential to deliver enhanced performances.

Figure 8 presents a comparison of the posterior density curves for the eight contaminant source parameters, derived from 2500 updated samples of the five repetition runs, using the three DA methods. Although in a single run, each of the three DA methods may produce slightly biased estimates of the contaminant source parameters (results not shown), merging the updated ensembles of the five repetition runs yield consistent results across the three methods. The RMSE values between the measurement data and the updated ensembles for $\mathrm{ES}_{(\mathrm{K})}$, $\mathrm{ES}_{(\mathrm{DL})}$, and $\mathrm{DA}_{(\mathrm{DL})}$ are 0.0574, 0.0574, and 0.0435, respectively. The corresponding standard deviations are 0.0335, 0.0719, and 0.0440, respectively. Although $\mathrm{ES}_{(\mathrm{K})}$ and $\mathrm{ES}_{(\mathrm{DL})}$ yield almost identical mean RMSE values, the results from $\mathrm{ES}_{(\mathrm{DL})}$ exhibit greater variability. Overall, the $\mathrm{DA}_{(\mathrm{DL})}$ method produces the best data-match.

4 Conclusions and Discussion

In this study, a novel DA method, i.e., $\mathrm{DA}_{(\mathrm{DL})}$, is proposed to improve the simulation accuracy of complex hydrological systems involving non-linearity, high-dimensionality,

and non-Gaussianity. Traditional DA methods, such as MCMC, may suffer from low computational efficiency, while others like EnKF and its variants are limited by the Gaussian assumption. $DA_{(DL)}$ takes advantage of DL to recognize complex patterns (including non-Gaussianity) and approximate non-linear relationships automatically from data. By employing ensemble representation of model parameters, states or other related variables, $DA_{(DL)}$ quantifies and reduces the uncertainties inherent in the simulation process from prior knowledge and measurement data. $DA_{(DL)}$ builds non-linear mappings from multiple predictors to the target variable, which is the difference between updated and prior vector of concerned variables (model parameters in the present study). To train the DL model, a large volume of training data are generated from the prior ensemble. When the system model is CPU-demanding, it is preferable to use a small ensemble size, which may not be sufficient for the data-hungry DL model. In this condition, we propose a data argumentation method to enhance the performance of $DA_{(DL)}$. In addition, we address the equifinality issue, which arises when different parameters or forcings can lead to the same outcomes in a complex dynamical system, by introducing a local updating approach proposed in our previous work (J. Zhang et al., 2018) to $DA_{(DL)}$. To evaluate the performance of $DA_{(DL)}$, we conduct numerical experiments involving Gaussian and non-Gaussian distributions and compare $DA_{(DL)}$ with two ES methods, one using the Kalman formula (Emerick & Reynolds, 2013) and the other using a DL-based update (J. Zhang, Zheng, et al., 2020). Our results demonstrate that $DA_{(DL)}$ outperforms its counterparts, especially in the non-Gaussian condition.

Despite the promising results achieved by $DA_{(DL)}$ in this study, there are still several issues that are not well addressed. Firstly, it is difficult or even impossible to design the optimal DL model structure and related training options for a specific problem. During the development of $DA_{(DL)}$, dozens of DL model structures have been tested, and only a few of them can yield satisfactory outcomes. Although the DL models used in this study enable $DA_{(DL)}$ to produce good results, especially in the non-Gaussian case, the current settings are suboptimal. For example, in the Gaussian case, $DA_{(DL)}$ still requires the same number of iterations as $ES_{(K)}$, despite the fact that the mapping defined by $DA_{(DL)}$ is non-linear. It is important to further improve and standardize the implementation of $DA_{(DL)}$. Secondly, as the updating made by $DA_{(DL)}$ relies on statistical learning from data, there is no assurance of physical consistency of the DA outcomes. To address this limitation, incorporating physical constraints into the training of DL model via knowledge-guided machine learning (Karpatne et al., 2022) would be a promising approach. This can help decrease the need of huge training data and enhance the reliability and stability of the inference results. Thirdly, this study only addresses the updating of model parameters form measurement data using $DA_{(DL)}$. However, there is potential to expand the scope of $DA_{(DL)}$ in future research by incorporating other crucial variables such as model states, forcings and error terms. Doing so would provide a more comprehensive understanding of simulation uncertainties, ultimately leading to more informed model enhancements and predictions.

Acknowledgments

Computer codes and data used are available at `https://www.researchgate.net/publication/370927560_MATLAB_codes_of_the_DA_DL_method`

This study is supported by the National Key R&D Program of China (2021YFC3200500), Jiangsu Provincial Innovation and Entrepreneurship Doctor Program (JSSCBS20210260) and National Natural Science Foundation of China (42007004).

References

Anderson, J. L. (2012). Localization and sampling error correction in ensemble Kalman filter data assimilation. *Monthly Weather Review*, *140*(7), 2359–2371. doi: 10.1175/MWR-D-11-00013.1

Bauser, H. H., Berg, D., Klein, O., & Roth, K. (2018). Inflation method for ensemble Kalman filter in soil hydrology. *Hydrology and Earth System Sciences*, *22*(9), 4921–4934. doi:

10.5194/hess-22-4921-2018

Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, *9*(5), e535. doi: 10.1002/wcc.535

Chen, Y., & Oliver, D. S. (2012). Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Mathematical Geosciences*, *44*(1), 1–26. doi: 10.1007/s11004-011-9376-z

Chen, Y., & Zhang, D. (2006). Data assimilation for transient flow in geologic formations via ensemble Kalman filter. *Advances in Water Resources*, *29*(8), 1107–1122. doi: 10.1016/j.advwatres.2005.09.007

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., ... others (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, *51*(4), 2498–2514. doi: 10.1002/2015WR017198

Dausman, A. M., Doherty, J., Langevin, C. D., & Sukop, M. C. (2010). Quantifying data worth toward reducing predictive uncertainty. *Groundwater*, *48*(5), 729–740. doi: 10.1111/j.1745-6584.2010.00679.x

Dechant, C. M., & Moradkhani, H. (2011). Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation. *Hydrology and Earth System Sciences*, *15*(11), 3399–3410. doi: 10.5194/hess-15-3399-2011

Emerick, A. A., & Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, *55*, 3–15. doi: 10.1016/j.cageo.2012.03.011

Etter, S., Strobl, B., Seibert, J., & van Meerveld, H. I. (2020). Value of crowd-based water level class observations for hydrological model calibration. *Water Resources Research*, *56*(2), e2019WR026108. doi: 10.1029/2019WR026108

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

Evensen, G. (2019). Accounting for model errors in iterative ensemble smoothers. *Computational Geosciences*, *23*(4), 761–775. doi: 10.1007/s10596-019-9819-z

Godoy, V. A., Napa-García, G. F., & Gómez-Hernández, J. J. (2022). Ensemble random forest filter: An alternative to the ensemble Kalman filter for inverse modeling. *Journal of Hydrology*, *615*, 128642. doi: 10.1016/j.jhydrol.2022.128642

Gu, Y., & Oliver, D. S. (2007). An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. *SPE Journal*, *12*(04), 438–446. doi: 10.2118/108438-PA

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*. Reston, VA: U. S. Geological Survey. (Retrieved from https://pubs.usgs.gov/of/2000/0092/report.pdf)

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). Las Vegas, NV: IEEE. doi: 10.1109/CVPR.2016.90

He, Q., Barajas-Solano, D., Tartakovsky, G., & Tartakovsky, A. M. (2020). Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport. *Advances in Water Resources*, *141*, 103610. doi: 10.1016/j.advwatres.2020.103610

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700–4708). Honolulu, HI: IEEE. doi: 10.1109/CVPR.2017.243

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440. doi: 10.1038/s42254-021-00314-5

Karpatne, A., Kannan, R., & Kumar, V. (2022). *Knowledge guided machine learning: Accelerating discovery using scientific knowledge and data*. Boca Raton: CRC Press.

Law, K., Stuart, A., & Zygalakis, K. (2015). *Data assimilation: A mathematical introduction*. Switzerland: Springer.

Liu, Y., & Gupta, H. V. (2007). Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water resources research*, *43*(7), W07401. doi: 10.1029/2006WR005756

Lorentzen, R. J., & Naevdal, G. (2011). An iterative ensemble Kalman filter. *IEEE Transactions on Automatic Control*, *56*(8), 1990–1995. doi: 10.1109/TAC.2011.2154430

Man, J., Guo, Y., Jin, J., Zhang, J., Yao, Y., & Zhang, J. (2022). Characterization of vapor intrusion sites with a deep learning-based data assimilation method. *Journal of Hazardous Materials*, *431*, 128600. doi: 10.1016/j.jhazmat.2022.128600

Mariethoz, G., Renard, P., & Straubhaar, J. (2010). The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research*, *46*(11), W11536. doi: 10.1029/2008WR007621

Moradkhani, H., Hsu, K.-L., Gupta, H., & Sorooshian, S. (2005). Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. *Water resources research*, *41*(5), W05012. doi: 10.1029/2004WR003604

Pan, Z., Lu, W., & Bai, Y. (2022). Groundwater contamination source estimation based on a refined particle filter associated with a deep residual neural network surrogate. *Hydrogeology Journal*, *30*(3), 881–897. doi: 10.1007/s10040-022-02454-z

Peel, M. C., & Blöschl, G. (2011). Hydrological modelling in a changing world. *Progress in Physical Geography*, *35*(2), 249–261. doi: 10.1177/0309133311402550

Pulido, M., & van Leeuwen, P. J. (2019). Sequential Monte Carlo with kernel embedded mappings: The mapping particle filter. *Journal of Computational Physics*, *396*, 400–415. doi: 10.1016/j.jcp.2019.06.060

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., ... others (2019). Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, *55*(11), 9173–9190. doi: 10.1029/2019WR024922

Reich, S., & Cotter, C. (2015). *Probabilistic forecasting and Bayesian data assimilation.* United Kingdom: Cambridge University Press.

Reuschen, S., Jobst, F., & Nowak, W. (2021). Efficient Discretization-Independent Bayesian Inversion of High-Dimensional Multi-Gaussian Priors Using a Hybrid MCMC. *Water Resources Research*, *57*(8), e2021WR030051. doi: 10.1029/2021WR030051

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Munich, Germany: Springer. doi: 10.1007/978-3-319-24574-4_28

Shi, P., Yang, T., Yong, B., Xu, C.-Y., Li, Z., Wang, X., ... Zhou, X. (2023). Some statistical inferences of parameter in MCMC approach and the application in uncertainty analysis of hydrological simulation. *Journal of Hydrology*, *617*, 128767. doi: 10.1016/j.jhydrol.2022.128767

Slater, L., & Binley, A. (2021). Advancing hydrological process understanding from long-term resistivity monitoring systems. *Wiley Interdisciplinary Reviews: Water*, *8*(3), e1513. doi: 10.1002/wat2.1513

Smith, P., Beven, K. J., & Tawn, J. A. (2008). Detection of structural inadequacy in process-based hydrological models: A particle-filtering approach. *Water resources research*, *44*(1), W01410. doi: 10.1029/2006WR005205

Spieler, D., Mai, J., Craig, J. R., Tolson, B. A., & Schütze, N. (2020). Automatic model structure identification for conceptual hydrologic models. *Water Resources Research*, *56*(9), e2019WR027009. doi: 10.1029/2019WR027009

Tarakanov, A., & Elsheikh, A. H. (2020). Optimal Bayesian experimental design for subsurface flow problems. *Computer Methods in Applied Mechanics and Engineering*, *370*, 113208. doi: 10.1016/j.cma.2020.113208

Thibaut, R., Compaire, N., Lesparre, N., Ramgraber, M., Laloy, E., & Hermans, T. (2022). Comparing Well and Geophysical Data for Temperature Monitoring within a Bayesian Experimental Design Framework. *Water Resources Research*, *58*(11), e2022WR033045. doi: 10.1029/2022WR033045

van Leeuwen, P. J., & Evensen, G. (1996). Data assimilation and inverse methods in terms

of a probabilistic formulation. *Monthly Weather Review*, *124*(12), 2898–2913. doi: 10.1175/1520-0493(1996)124⟨2898:DAAIMI⟩2.0.CO;2

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling & Software*, *75*, 273–316. doi: 10.1016/j.envsoft.2015.08.013

Vrugt, J. A., Ter Braak, C. J., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, *44*(12), W00B09. doi: 10.1029/2007WR006720

Wang, Y., Shi, L., Zha, Y., Li, X., Zhang, Q., & Ye, M. (2018). Sequential data-worth analysis coupled with ensemble Kalman filter for soil water flow: A real-world case study. *Journal of Hydrology*, *564*, 76–88. doi: 10.1016/j.jhydrol.2018.06.059

Wang, Y., & Yan, B. (2022). On the feasibility of an ensemble multi-fidelity neural network for fast data assimilation for subsurface flow in porous media. *Available at SSRN 4293917*. doi: 10.2139/ssrn.4293917

Xiao, C., Zhang, S., Ma, X., Zhou, T., Hou, T., & Chen, F. (2023). Deep-learning-generalized data-space inversion and uncertainty quantification framework for accelerating geological CO_2 plume migration monitoring. *Geoenergy Science and Engineering*, *224*, 211627. doi: 10.1016/j.geoen.2023.211627

Xue, L., Zhang, D., Guadagnini, A., & Neuman, S. P. (2014). Multimodel Bayesian analysis of groundwater data worth. *Water Resources Research*, *50*(11), 8481–8496. doi: 10.1002/2014WR015503

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*(2), 773–794. doi: 10.1016/j.jcp.2003.09.015

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 Resources Research*, *54*(3), 1716–1733. doi: 10.1002/2017WR020906

Zhang, J., Vrugt, J. A., Shi, X., Lin, G., Wu, L., & Zeng, L. (2020). Improving simulation efficiency of MCMC for inverse modeling of hydrologic systems with a Kalman-inspired proposal distribution. *Water Resources Research*, *56*(3), e2019WR025474. doi: 10.1029/2019WR025474

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. doi: 10.1002/2014WR015740

Zhang, J., Zheng, Q., Wu, L., & Zeng, L. (2020). Using deep learning to improve ensemble smoother: Applications to subsurface characterization. *Water Resources Research*, *56*(12), e2020WR027399. doi: 10.1029/2020WR027399

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*. DTIC Document. (Retrieved from http://www.geology.wisc.edu/courses/g727/mt3dmanual.pdf)

Zheng, F., Tao, R., Maier, H. R., See, L., Savic, D., Zhang, T., ... others (2018). Crowdsourcing methods for data collection in geophysics: State of the art, issues, and future directions. *Reviews of Geophysics*, *56*(4), 698–740. doi: 10.1029/2018RG000616