Identification of contaminant source and hydraulic conductivity field based on an ILUES-SOM surrogate model

Na Zheng
Tongji University
Jinbing Liu
Yangtze Ecology and environment corporation Limited, China
Xuemin Xia
University of Shanghai for Science and Technology
Simin Gu
Tongji University
Xianwen Li
Northwest Agriculture and Forestry University
Simin Jiang (✉ jiangsimin@tongji.edu.cn)
Tongji University

Research Article

Keywords: Groundwater contamination, Contaminant source identification, hydraulic conductivity estimation, Self-organizing map, ensemble smoother, surrogate model

Posted Date: May 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1649935/v1

Identification of contaminant source and hydraulic conductivity field based on an ILUES-SOM surrogate model

Na Zheng[1] · Jinbing Liu[2] · Xuemin Xia[3] · Simin Gu[1] ·Xianwen Li[4] · Simin Jiang[1*]

[1] Department of Hydraulic Engineering, Tongji University, Shanghai, China

[2] Yangtze Ecology and Environment Corporation Limited, Wuhan, China

[3] School of Environment and Architecture, University of Shanghai for Science and Technology, Shanghai,China

[4] Key laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas, Ministry of Education, Northwest A&F University, Yangling, China

Corresponding author: Simin Jiang (jiangsimin@tongji.edu.cn)

Abstract

Contaminant source identification and hydraulic conductivity estimation are of great significance for contaminant transport model in the subsurface media, but their actual values are difficult to obtain and can usually be inversely identified and estimated by sparse observations. In order to reduce computational cost in the process of estimating groundwater model parameters, the surrogate model was often used. This study addresses this challenge by proposing a modified self-organizing map (SOM) based surrogate model, named ILUES-SOM, which combing a modified iterative ensemble smoother method (SGSIM-ILUES) and SOM algorithm, to simultaneously identify contaminant source parameters and hydraulic conductivity field. Considering the characteristics of the proposed method (ILUES-SOM), the comparison of parameter estimation accuracy and computational efficiency is performed with original SOM and SGSIM-ILUES inversion model. Moreover, the robustness of ILUES-SOM model for inversion was illustrated by proposing varying degrees of observation errors and missing early observation data. The results indicated that ILUES-SOM model can successfully retrieve unknown contaminant source simultaneously with heterogeneity hydraulic conductivity field in groundwater system.

Keywords

Groundwater contamination, Contaminant source identification, hydraulic conductivity estimation, Self-organizing map, ensemble smoother, surrogate model

1. Introduction

Identifying the source of groundwater pollution is of great significance to groundwater remediation and management(He et al. 2021). However, the occurrence of groundwater pollution has the characteristics of concealment and discovery lag, and the number of groundwater monitoring wells is small, so it is often difficult to directly obtain pollution sources information and hydrogeological parameters(Prakash and Datta 2013). For this situation, the contaminant source information can be identified by groundwater inverse problem using sparse observations and site prior information, thus to restore the migration and transformation process of pollutants in groundwater (Atmadja and Ba Gtzoglou 2001). In the inversion method for solving the groundwater inverse problem, data assimilation methods can combine dynamic data such as hydraulic head and site pollutant concentration in the groundwater flow and pollutant transport model to reduce the uncertainty of aquifer parameters, and the updated parameters can improve prediction accuracy of groundwater numerical model (Bao et al. 2020).

Among data assimilation algorithms, the ensemble Kalman filter (EnKF) (Evensen 1994) is widely used in the field of hydrogeological research due to its excellent performance (Kang et al. 2021; Li et al. 2012; Schöniger et al. 2012). van Leeuwen and Evensen (van Leeuwen and Evensen 1996) proposed a variant of EnKF: ensemble smoother (ES). It has been shown that ES can obtain similar results to EnKF but with a much lower computational cost (Li et al. 2018)，and widely used in hydrogeology and reservoir research(Bailey and Baù 2010; Bailey et al. 2012; Lima et al. 2020). But when the system is highly nonlinear, iterative application of ES (IES) are needed (Chen and Oliver 2012). Ju et al. (Ju et al. 2018) improved the standard IES and proposed an iterative ensemble smoother algorithm based on Gaussian process, which further improved the computational efficiency；Cao et al. (Cao et al. 2018) coupled IES with multi-point geostatistical method to assimilate dynamic data into non-Gaussian aquifer. To improve the applicability and efficiency of IES for strongly nonlinear problems, Zhang et al. (Zhang et al. 2018) proposed a simple and efficient algorithm, i.e., the iterative local updating ensemble smoother (ILUES), to extend IES to inverse problems with multimodal distributions. Some studies have demonstrated that ILUES can significantly reduce the uncertainty of model parameters (Yang et al. 2020; Liu et al. 2021; Zhang et al. 2020).

For high-dimensional problems, a large ensemble size and iteration number are required to guarantee reliable estimation of unknown parameters in ILUES, leading to a huge computational burden (Zhang et al. 2018). A effective method to improve computational efficiency is to use surrogate models (Asher et al. 2015).

In recent years, with the enhancement of computer performance, the method of constructing surrogate models using machine learning (ML) has become increasingly popular (Chan and Elsheikh 2020; Tang et al. 2020, 2021; Zhong et al. 2019). Among them, Hazrati et al. (Hazrati and Datta 2017a,b) used self-organizing map(SOM) to construct surrogate model and identify the intensity of pollution sources where the location of pollution sources were known and slight to mild heterogeneity of the aquifer was considered；On the basis of Hazrati et al. 's research, Xia et al. (Xia et al. 2019)further explored the effect and robustness of SOM-based surrogate model and identify pollution source parameters (location and release history) in more realistic case, the pollution source parameters were unknown and the heterogeneity was much stronger；Jiang et al. (Jiang et al. 2021) combined the dimensionality reduction idea of pilot points method and applied SOM algorithm to construct surrogate models for simultaneous identification of pollution sources and hydraulic conductivity field.

As a data mining technology, SOM algorithm highlights the nonlinear relationship of data by transforming the original data (Penn 2005). The algorithm converts high-dimensional data into low-dimensional by calculating the main features and correlation between the input data, which effectively improves the data processing ability and further improves the computational efficiency (Simula et al. 1998). The surrogate model of contaminant transport constructed by the SOM algorithm, not merely replaced the complex original numerical model (groundwater flow and solute transport simulation model), but also had the ability to identify unknown model parameters, which meant that other aforementioned inverse solution methods such as data assimilation methods were no longer needed and a large computational cost was subsequently reduced (Jiang et al. 2021).

Among the related researches on the above-mentioned SOM algorithm used for groundwater model parameter inversion, there are few researches on simultaneous inversion of pollution source parameters and hydraulic conductivity field. Furthermore, as a machine learning method, the quality of the training data is one of the important factors to determine the goodness of SOM-based surrogate model. In view of the fact that the data assimilation method can effectively integrate the physical model and observation data, thus generating sample data considering the observed data and complying with the pollutant transport model, does the SOM model based on this posteriori samples have better performance? This study tackled the above challenges via a modified SOM, constructed using posterior samples from ILUES algorithm.

The remainder of this paper is organized as follows. Section 2 presents the detailed description of groundwater flow and the contaminant transport model. Section 3 outlines the framework of the proposed methodology. In Section 4, results obtained from numerical experiments are discussed. Section 5 discusses different scenarios and analyzes the results. Some conclusions are given in Section 6.

2. Problem Formulation

The transport of contaminant in saturated aquifer may involve diffusion, advection, dispersion, absorption. In this study, advection and dispersion are dominated processes in a two-dimensional contaminant transport system under steady-state groundwater flow conditions.

The governing equation for the steady-state groundwater flow can be written as follows:

$$\frac{\partial}{\partial \mathrm{x}}\left(K\frac{\partial \mathrm{h}}{\partial \mathrm{x}}\right)=0 \tag{1}$$

and the flow velocity v [LT^{-1}] can be obtained by Darcy's law:

$$v=-\frac{K}{\theta}\frac{\partial h}{\partial x} \tag{2}$$

where h [L] is the hydraulic head; K [LT^{-1}] represents the hydraulic conductivity; θ is effective porosity (dimensionless); The flow governing equation is solved by numerical simulator MODFLOW (Harbaugh et al. 2000). Then, the resulting velocity v is used as input for the advection-dispersion equation to calculate the contaminant concentration by MT3DMS (Zheng and Wang 1999). The advection-dispersion equation for a 2D saturated aquifer is:

$$\frac{\partial(bC)}{\partial t}=\frac{\partial}{\partial x}\left(bD\frac{\partial C}{\partial x}\right)-\frac{\partial}{\partial x}(bvC)+\frac{C_sW}{\theta} \tag{3}$$

where t [T] is time; b [L] is the saturated thickness of aquifer; C [ML^{-3}] is the concentration of the dissolved chemical species; Cs[ML^{-3}] is the concentration of source or sink; W[LT^{-1}] is the volumetric flux per unit area; D[L^{2}T^{-1}] is the hydrodynamic dispersion coefficient, determined by v, and longitudinal and transverse dispersivities (α_{L} and α_{T}) [L].

3. Methodology

The self-organizing maps (SOM) is a clustering algorithm proposed by Kohonen (Kohonen 1982), which consists of an input layer and an output layer representing the grid topology. Each neuron in the input layer is connected to the neuron in the output layer by a weight vector. The principle of SOM algorithm is as follows(Chaudhary et al. 2015)：

Firstly, initialize the weight vector of each neuron in the output layer in a random or linear manner, and set the learning rate and the type of the neighborhood function, the neighborhood function is used for quantitatively describe the relationship between any neuron and surrounding neurons. Secondly, Euclidean distance (in Equation (4)) between the input training data and each weight vector is calculated. Neuron with the smallest Euclidean distance (the winner neuron) is activated together with neurons in the topological neighborhood (in Equation (5)), and their corresponding weight vector are adjusted to make them move to the training data. Repeat the above steps until the learning rate decays to zero. Finally, the network (the output layer with modified weight vector) is obtained to represent the topological relationship between training data, in which each neuron represents a cluster.

$$d_j(x) = \sum_{i=1}^{D} (x_i - \omega_{ji})^2 \tag{4}$$

where D is dims of input; x_i is input data; ω_{ji} is weight vector between neuron j and input data.

$$T_{j,I(x)}(\mathrm{t}) = \exp\left(-\frac{S_{j,I(x)}^2}{2\sigma(t)^2}\right) \tag{5}$$

I(x) is the winner neuron; $S_{j,I(x)}$ is distance between the winner neuron and neuron j; σ is a coefficient of decay with time.

$$\Delta\omega_{ji} = \eta(t) \cdot T_{j,I(x)}(t) \cdot (x_i - \omega_{ji}) \tag{6}$$

$\eta(\mathrm{t})$ is learning rate and decrease with time.

For the trained surrogate model, the information contained in the neurons is called the map codebook. When using it for forward prediction or inverse source identification, the surrogate model determines the winner neuron by calculating the distance between the neuron and the input vector, and the data vector in the map codebook corresponding to the winner neuron is output. The flow chart of this study using the SOM surrogate model for prediction is shown in Fig. 1. A detailed explanation of the original SOM surrogate model can be found in the previous research work (Jiang et al. 2021; Xia et al. 2019).

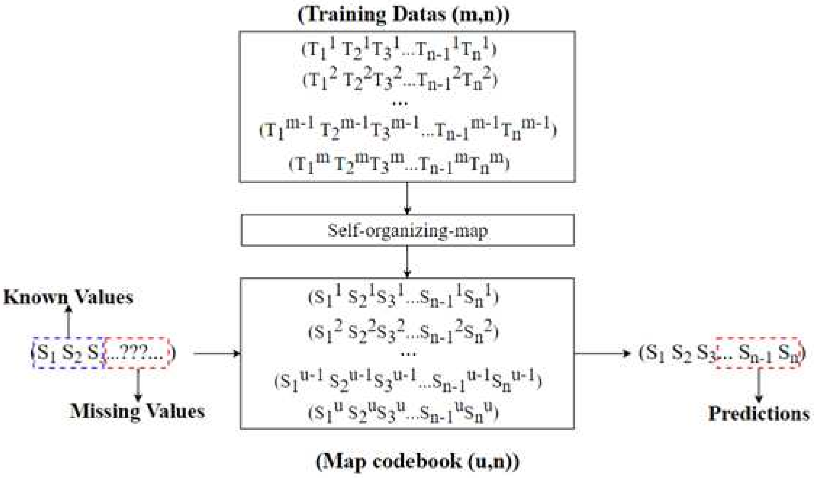
(Training Datas (m,n))
$(T_1^1\ T_2^1 T_3^1 ... T_{n-1}^1 T_n^1)$
$(T_1^2\ T_2^2 T_3^2 ... T_{n-1}^2 T_n^2)$
...
$(T_1^{m-1}\ T_2^{m-1} T_3^{m-1} ... T_{n-1}^{m-1} T_n^{m-1})$
$(T_1^m\ T_2^m T_3^m ... T_{n-1}^m T_n^m)$
Self-organizing-map
Known Values
$(S_1\ S_2\ S_3 ...???...)$
Missing Values
$(S_1^1\ S_2^1 S_3^1 ... S_{n-1}^1 S_n^1)$
$(S_1^2\ S_2^2 S_3^2 ... S_{n-1}^2 S_n^2)$
...
$(S_1^{u-1}\ S_2^{u-1} S_3^{u-1} ... S_{n-1}^{u-1} S_n^{u-1})$
$(S_1^u\ S_2^u S_3^u ... S_{n-1}^u S_n^u)$
$(S_1\ S_2\ S_3 ... S_{n-1}\ S_n)$
Predictions
(Map codebook (u,n))

Fig. 1 The schematic diagram for application of the constructed surrogate model

The previous research work (Xia et al. 2019) has indicated that Imp SOM algorithm performs better than batch SOM in the construction of surrogate models for groundwater pollute transport. Besides the training algorithm, the hyperparameter such as map units was the most important parameter and determined by trial and error method (Jiang et al. 2021). Furthermore, the quantity and quality training data were also important factor to determine the quality of SOM-based surrogate model. With regard to the quantity of training data, almost all previous research had taken this into account, but as far as we know, there was almost no research on the quality of training data in SOM-based surrogate model.

Considering that the ensemble-based data assimilation methods was the most widely used method for groundwater inverse problems and because of the similarity between the a priori/posterior set and the training data, the ensemble-based data assimilation method was adopted to improve the training data, then the surrogate model was constructed on the basis of posterior samples.

The main procedures for constructing the modified SOM based surrogate model (ILUES-SOM) for the solute transport model are as follows.

Step 1: Generaton of training and validation data.

A large amount of training data is needed to obtain an accurate surrogate model. In this study, the training data for the SOM based surrogate model consists of the inputs and outputs of the original groundwater numerical model. The input is the pollution source parameter and the hydraulic conductivities at pilot points, where the pollution source parameter includes the location of the pollution source and the release concentration of pollutants in each stress period. The output is the pollutant concentration at the observation points.

A modified iterative ensemble smoother (SGSIM-ILUES) proposed by Jiang et al. (Jiang et al. 2022) was adopted as the inversion framework to improve the training data. The SGSIM-ILUES method was based on the coupling of ILUES and sequential gaussian simulation (SGSIM) in geostatistics. Specially, the inversion of hydraulic conductivities was converted to the estimation of hydraulic conductivity at pilot points. The posterior samples from ILUES algorithm (1 iteration) was used as training data. In order to evaluate the accuracy of the surrogate model obtained, the same steps are used to generate the validation data, and the validation data is fixed to 500 groups in this study. A detailed explanation of the SGSIM-ILUES method can be found in the study (Jiang et al. 2022).

Step 2: Training of the SOM based surrogate model.

As mentioned, the codebook size (units) and the training data size (TDS) have significant impact on the performance of the SOM based surrogate model. Therefore, the number of units is set to 100, 500, 1000, 1500, 2000, 2500, 3000, respectively, and the training data size is set to 500, 1000, 2000, respectively. After training multiple surrogate models, use the validation data to evaluate the accuracy of each candidate surrogate model, and select the optimal surrogate model for subsequent groundwater model parameter inversion.

Step 3: Using the surrogate model to identify unknown values.

Finally, the constructed ILUES-SOM based surrogate model can be used to estimate the missing components. Using the known observed true values, find the best matching unit (BMU) in the optimal ILUES-SOM based surrogate model, from which the estimated values of the groundwater model parameters can be retrieved, and then use the geostatistical method (i.e. SGSIM) to obtain the estimated hydraulic conductivity field, so as to complete the inversion of the pollution source parameters and the hydraulic conductivity field.

4. Illustrative Example

4.1 A hypothetical aquifer site

In this study, advection and dispersion were dominated processes in a two-dimensional contaminant transport system under steady-state groundwater flow conditions. The hypothetical aquifer was saturated and confined aquifer with a 2D steady-state groundwater flow. Specifically, the aquifer size was 40 m in the x-direction and 20 m in the y-direction. The top and bottom boundary was no-flux. The constant hydraulic heads of 12 m and 10 m were the west and east boundary, and the aquifer thickness was 1 m. Values of model parameters are listed in Table 1.

Table 1 Hydrogeological characteristics of the hypothetical aquifer

Parameter	Units	Value
Grid size	m	0.5 × 0.5
Aquifer thickness	m	1.0
Effective porosity	-	0.30
Longitudinal dispersivity	m	2.0
Transverse dispersivity	m	0.6
Simulation time	day	40

Considering the spatial heterogeneity, the reference hydraulic conductivity field (Fig. 2(a)) was lognormally distributed with mean =4.0 and variation = 0.5, and the correlation length along x-direction and y-direction were 8.0 m and 4.0 m, respectively. The reference hydraulic conductivity field was generated based on known hydraulic conductivity of hard data using SGSIM method.

In this hypothetical aquifer, some amount of contaminant was released from a point source (Fig. 2(a) asterisk). A contaminant source is placed in this hypothetical aquifer. The contaminant source was characterized by ten parameters, i.e., S_x, S_y, $SP_i(MT^{-1})$ during the i_{th} stress periods, for $i = 1,...,8$ as listed in Table 2. It is assumed that the possible location range of pollution sources (Fig. 2(a) Red dotted area) and prior range of pollution source (in Table 2) can be determined in the investigation of groundwater pollution sites.

The simulation mode for MODFLOW and MT3DMS was steady-state and transient, respectively. The whole simulation time was 40 days, which was equally divided into 8 stress periods. There were twenty observation wells with their locations shown in Fig. 2(a) to gather observation head and concentration every four days, i.e., $t = 4,8,12...,40$. The number of pilot points in this study was fixed at 80, and the distribution is shown in Fig. 2(b). Consequently, 10 unknown contaminant source parameters and hydraulic conductivities at 80 pilot points need to be estimated.

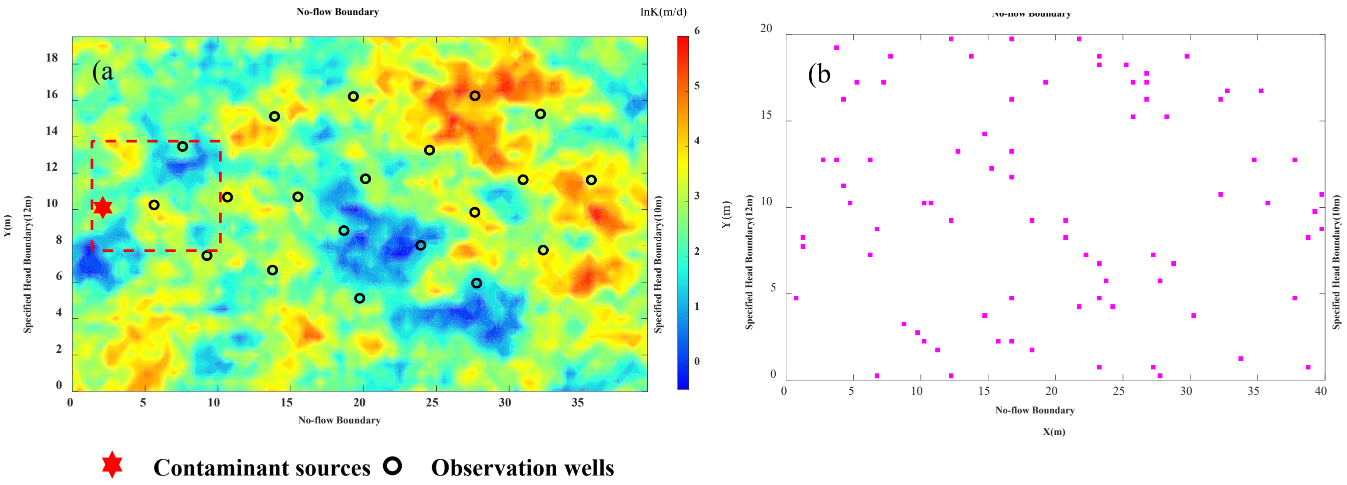

Fig. 2 (a)The reference hydraulic conductivity with location of contaminant sources and observation wells; (b)location of pilot points 1 – 80

Table 2 Actual values of the contaminant source flux and prior range

Contaminant Parameter	Prior range	Actual value
S_x [L]	[1.25 10.25]	2.25
S_y [L]	[7.75 13.75]	10.25
SP1(g/s)	[35 75]	50
SP2(g/s)	[35 75]	48
SP3(g/s)	[30 70]	45
SP4(g/s)	[30 60]	40
SP5(g/s)	[25 55]	36
SP6(g/s)	[22 45]	30
SP7(g/s)	[15 30]	20
SP8(g/s)	[7 15]	10

4.2 Assessment criteria

In this study, the pollution source parameter (SS), hydraulic conductivities at pilot points (KPP) and observation values (OBS) in the validation data were regarded as missing parts, and these missing values were estimated by SOM based surrogate model. The normalized absolute error of estimation (NAEE) and root mean square error (RMSE) were used to evaluate the inversion results. The NAEE can quantitatively characterize the deviation degree between the estimated value and the actual value, which is defined as follows:

$$\mathrm{NAEE}(\%) = \frac{\sum_{i=1}^{N} |(d_{est})_i - (d_{act})_i|}{\sum_{i=1}^{N} (d_{act})_i} \times 100 \tag{7}$$

RMSE can measure the degree of match between the estimated value and the actual value, and is defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} [(d_{est}) - (d_{act})]^2}{N}} \tag{8}$$

where d_{est} is estimated value, d_{act} is actual value, N is the number of values.

In this study, NAEE_{SS}, NAEE_{KPP} and NAEE_{OBS} represent NAEE calculation results of pollution source parameters, hydraulic conductivities at pilot points and observation values, respectively. The surrogate model with the highest estimation accuracy (NAEE) of both pollution source parameters and hydraulic conductivities at pilot points was selected as the optimal surrogate model.

5. Results and discussion

Considering the characteristics of the proposed methodology (in Section 3), the construction of SOM surrogate models, identification of contaminant source and hydraulic conductivity field based on SOM surrogate model were carried out and organized into three subsections.

(1) Firstly, the original SOM based surrogate model (S1) and the ILUES-SOM based surrogate model (S2) were constructed using randomly generated training samples and posterior samples from SGSIM-ILUES algorithm (1 iteration), respectively, and the optimal sizes of the codebook and training sample for SOM surrogate models were obtained.

(2) Then, the unknown contaminant source and hydraulic conductivities at pilot points were identified based on surrogate model S1 and S2, and the better surrogate model was found by comparison.

(3) Finally, the identification performance of the better surrogate model was evaluated in the presence of observation error and in the absence of observational data, respectively.

5.1 SOM based surrogate model

The first stage was to obtain the optimal sizes of the codebook and training sample for these two SOM based surrogate model (original SOM and ILUES-SOM), because their sizes greatly affected the accuracy and efficiency of the SOM based surrogate models. Specifically, the training data size was 500, 1000, 2000 groups, respectively, and the codebook size of the SOM was set to 100, 500, 1000, 1500, 2000, 2500, 3000, respectively. Their corresponding original SOM and ILUES-SOM based surrogate models were constructed.

5.1.1 Original SOM based surrogate model

The verification results ($\mathrm{NAEE}_{\mathrm{OBS}}$, $\mathrm{NAEE}_{\mathrm{SS}}$, $\mathrm{NAEE}_{\mathrm{KPP}}$) of the SOM surrogate models with different parameter (training data size, TDS; codebook size, Units) combinations were compared in Fig. 3. The NAEE of estimating OBS, SS and KPP were obviously decreased with Units (precision increased with increased Units, in Fig. 3(a, b, c)) except the estimation of OBS with TDS=2000. There were significant increasing trends of $\mathrm{NAEE}_{\mathrm{OBS}}$, $\mathrm{NAEE}_{\mathrm{SS}}$, $\mathrm{NAEE}_{\mathrm{KPP}}$ with TDS (precision decreased with increased TDS). Note that the required CPU time is a crucial factor in addition to the model's accuracy. The CPU time increased exponentially with codebook size (Fig. 3d), and the computational time for SOM based surrogate model (in Fig. 3d) was much smaller than the computational time for the physically based model (i.e. training data generation).

After consideration of above-mentioned accuracy and efficiency, the optimal surrogate model is the codebook trained by the combination of TDS=500 with Units =3000.

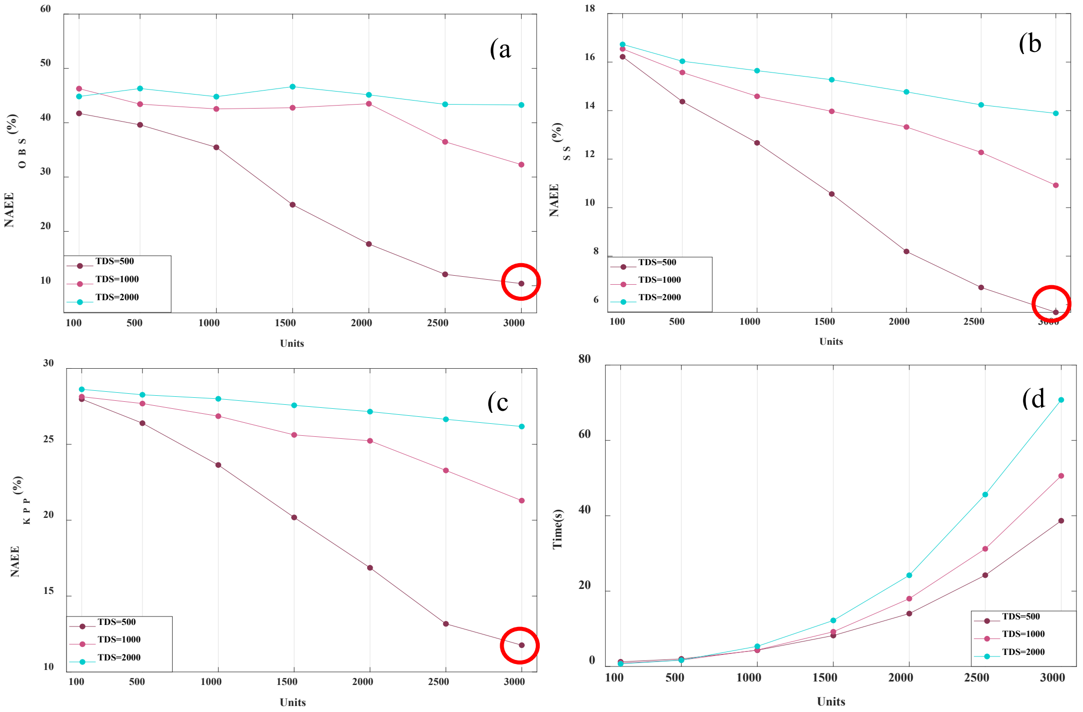

Fig. 3 Validation results and Training time of surrogate model S1: NAEE of (a) OBS, (b) SS, (c) KPP, (d) Training time

5.1.2 ILUES-SOM based surrogate model

Considering the improvement of training data can help improve the quality of surrogate model, posterior samples from data assimilation algorithm was adopted as training data for SOM based surrogate model. Considering that the ensemble-based data assimilation methods such as ILUES algorithm usually meant higher computational burden, which increased nearly linearly with the iteration number, only one iterative SGIM-ILUES operation was performed in the proposed ILUES-SOM surrogate model.

The verification results ($NAEE_{OBS}$, $NAEE_{SS}$, $NAEE_{KPP}$) of the trained ILUES-SOM based surrogate models were shown in Fig. 4. In comparison with Fig. 3, $NAEE_{OBS}$, $NAEE_{SS}$, $NAEE_{KPP}$ have been significantly improved, which proved that ILUES-SOM surrogate model was superior to original SOM model. It can be seen that there were no significant variations of $NAEE_{OBS}$, $NAEE_{SS}$, $NAEE_{KPP}$ with Units. As the validation results depicted in Fig. 4, when training data size was 500 groups and codebook size was 3000, the surrogate model had the highest inversion accuracy for the hydraulic conductivities at pilot points, and the inversion accuracy of pollution source parameters was also high. As selection principles suggested (in sect 4.2), the optimal surrogate model is the codebook trained by the combination of TDS=500 with Units =3000.

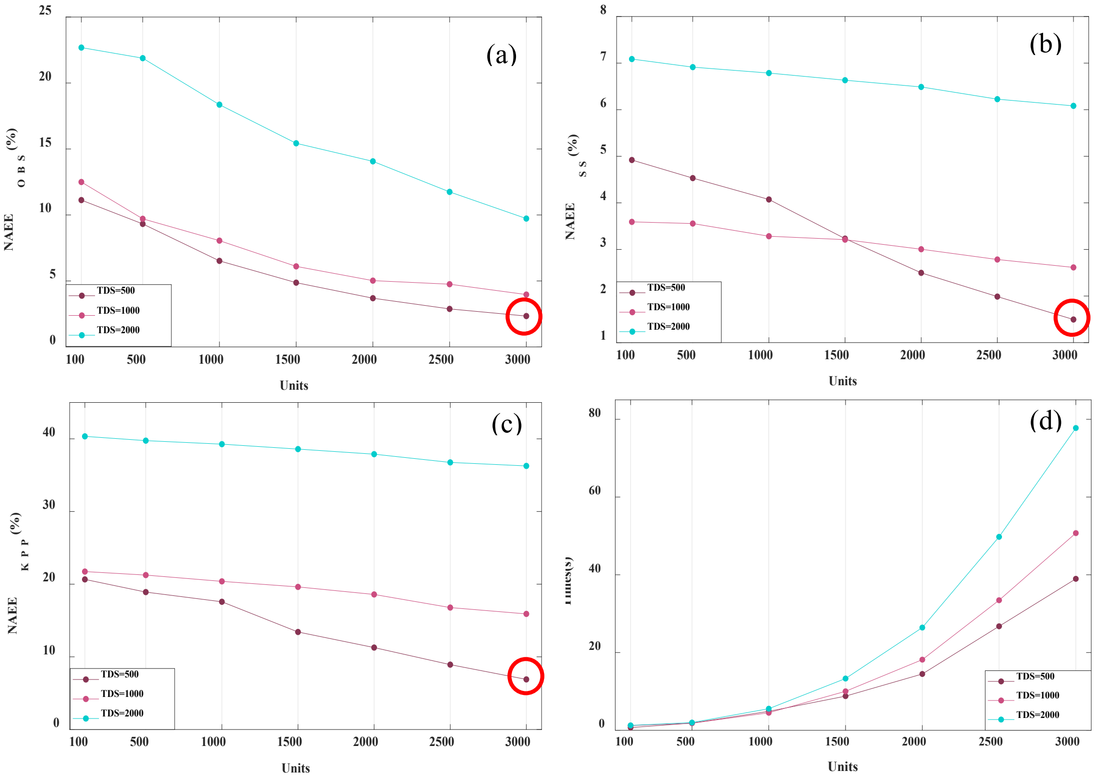

Fig. 4 Validation results of surrogate model S2: NAEE of (a) OBS, (b) SS, (c) KPP

5.2 Application for constructed SOM based surrogate model

The unknown contaminant source and hydraulic conductivities at pilot points were identified based on surrogate model S1 and S2. For further comparison, the contaminant source and hydraulic conductivity field were identified by SGSIM-ILUES inversion model with N_e =2000 and N_{iter} =8 (model O).

The inversion results of the hydraulic conductivity field of the above-mentioned three models (O, S1, S2) were shown in Fig. 5. Compared with the reference log-transformed conductivity field, the SGSIM-ILUES inversion model (model O) has the best inversion result, and the two SOM based surrogate models have a relatively large error in characterizing low conductivity areas. After further comparison, the morphology of K-field depicted in model S2 was slightly better than that in model S1.

To test the accuracy of SOM based surrogate model, the identified contaminant source information from different inversion models were compared with the actual values. It can be seen in Fig. 6 that the estimation precision of ILUES-SOM (S2) was fairly high for both contaminant source location and source fluxes, and was close to that of SGSIM-ILUES (O). Further comparison of the inversion accuracy of model S1 and S2, model S2 was more accurate except the source flux at SP1. Fig. 7 showed the estimated values and estimated deviations of model S1 and S2. It was clear that the deviation bar of model S1 was much bigger than model S2, indicating that ILUES-SOM (S2) was the better model for estimating unknown contaminant source.

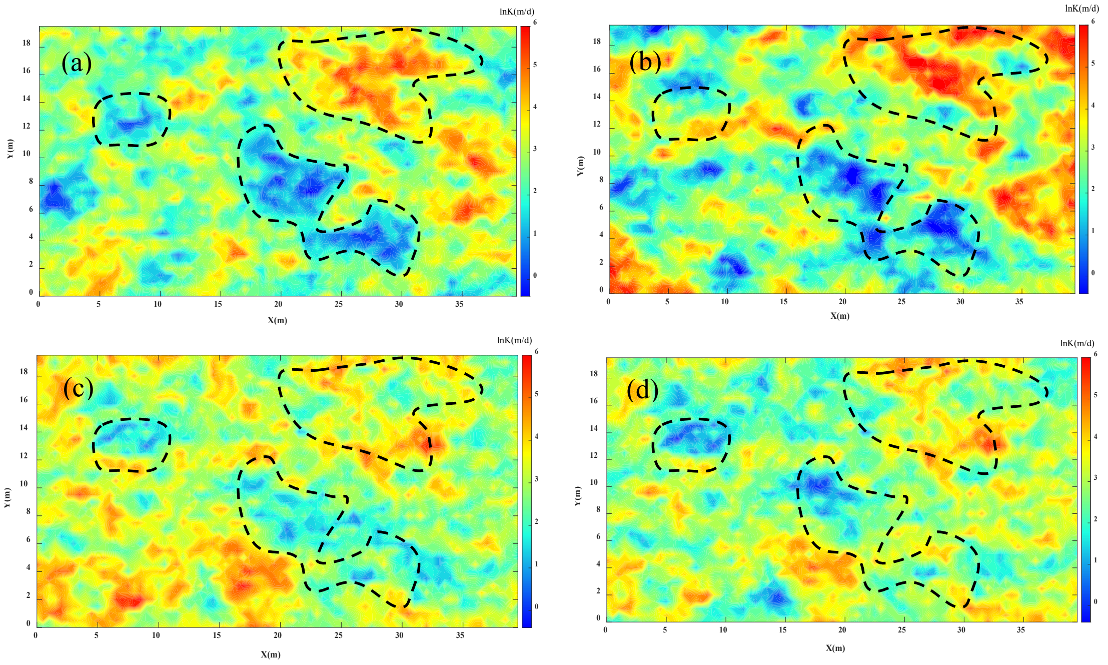

Fig. 5 (a)The reference K-field; (b)-(d) The corresponding interpolated K-field based on estimated KPP of ILUES based on original model (O) and optimal surrogate models of S1, S2

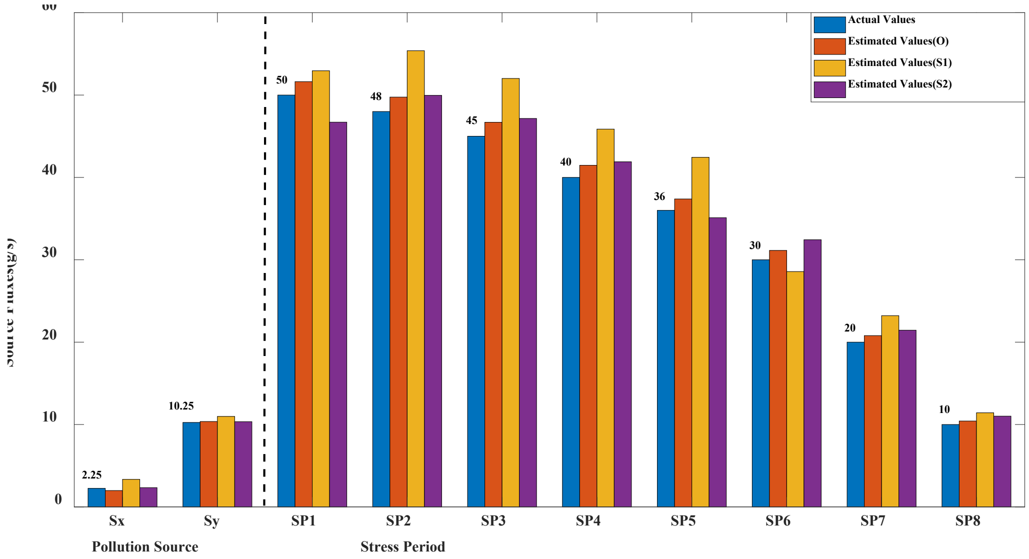

Fig. 6 Comparison of estimated SS by ILUES based on original model (O) and optimal surrogate models of S1, S2

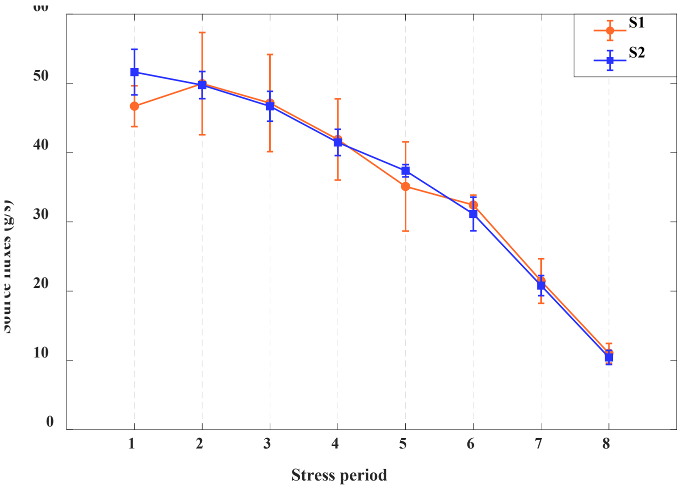

Fig. 7 Error comparison of estimated source fluxes by optimal surrogate models of S1 and S2

After a comparative analysis of the inversion performance of the two SOM surrogate models in terms of K-field and unknown contaminant source (SS), respectively, the performance of the two SOM models (S1, S2) and the ILUES inversion model (O) are synthesized by RMSE criteria (Table 3). Specially, for estimating SS and KPP, the inversion accuracy of model S2 was closer to that of ILUES inversion model, and clearly superior to that of model S1.

Table 3 Comparison of RMSE and time cost of estimated SS and KPP by ILUES inversion model (O) and optimal surrogate models of S1, S2

	RMSE		Time consuming(s)
	SS	KPP	
O	1.22	1.20	157856 (43.85h)
S1	4.52	1.87	3251.5 (0.90h)
S2	1.81	1.33	21932.8 (6.09h)

Table 3 showed the computation time for ILUES inverse model (O) and two SOM based surrogate model (S1, S2), where the computation time of ILUES inverse model (N_e =2000 and N_{iter} =8) was 43.85 hours. The computational time of the SOM based surrogate model mainly included the time to generate training data and the training time of the SOM model, where the former was the main computational burden. It should be noted that model S2 were constructed using posterior samples from SGSIM-ILUES algorithm (1 iteration), thus its time to generate training data included the time for initial sample generation and the time for 1 iteration of SGSIM-ILUES operation.

As can be seen from Table 3, model S1 constructed based on the original data has the largest improvement in computational efficiency compared to model O by 97% (i.e., the computational time is reduced from 43.85h to 0.90h), but there were significant deviations in the inversion accuracy of the parameters (SS and KPP). Meanwhile, the computational efficiency improvement of model S2 was slightly lower than that of model S1, but also reached 86% (from 43.85 h to 6.09 h), and the inversion accuracy of the unknown model parameters was closer to that of the ILUES inversion model (SS and KPP).

Considering the calculation accuracy and computational efficiency of the two SOM-based surrogate model, model S2 (ILUES-SOM model) was a better choice, which could not only ensure the parameter inversion accuracy, but also significantly improve the computational efficiency.

5.3 Further discussion

As can be seen from the results of 5.1 and 5.2, model S2 not only showed better accuracy in the validation stage of surrogate model, but also had better performance in the parameter inversion stage. Considering that groundwater system was a complex system affected by many factors, there were various uncertainties, and in practical problems, the uncertainty of site information was mainly caused by incomplete observation data. In order to fully consider the actual situation, two scenarios were designed to further analyze the performance of the optimal ILUES-SOM surrogate. In scenario 1, a varying degree of observation error was introduced to test the robustness of the proposed ILUES-SOM surrogate model; and early observation data was missing in scenario 2 to complicate the identification process.

Scenario1

In this scenario, a varying degree of observation errors were introduced to test the robustness of the proposed ILUES-SOM model. These observation errors were generated by adding different degree of random errors to the numerically simulated concentrations (C) at observation wells.

It is assumed that the random errors at the observation wells follow a normal distribution, where the arithmetic mean is zero and the standard deviation is 1.

$$C' = C + \varepsilon \times a \times C \qquad (9)$$

where C' is the perturbed value; ε is a normally distributed random value; a is the error level of observation, and three error level with the values of 5%, 10% and 15% were chosen to estimate the effects of errors on parameter inversion.

The inversion results of hydraulic conductivity field with varying degrees of noise in the contaminant concentrations were shown in Fig. 8. It can be seen that the estimated K-fields under different error level (error-free, 5%, 10%, 15%) were stable and slightly affected by error level (up to 15%).

The identification results of unknown contaminant source by the optimal ILUES-SOM model were shown in Table 4, and the inversion results were not significantly affected when the error level was below 15%. Specially, for estimating SS and KPP, RMSE(SS) and RMSE(KPP) for the cases with different observation noise were almost stable and only slightly larger than that of the error-free case (Table 4).

Therefore, the proposed ILUES-SOM based surrogate model was able to handle varying degrees of observation errors for identifying unknown contaminant source and estimating K-field. The identification error was stable when the observation errors range from 5% to 15%.

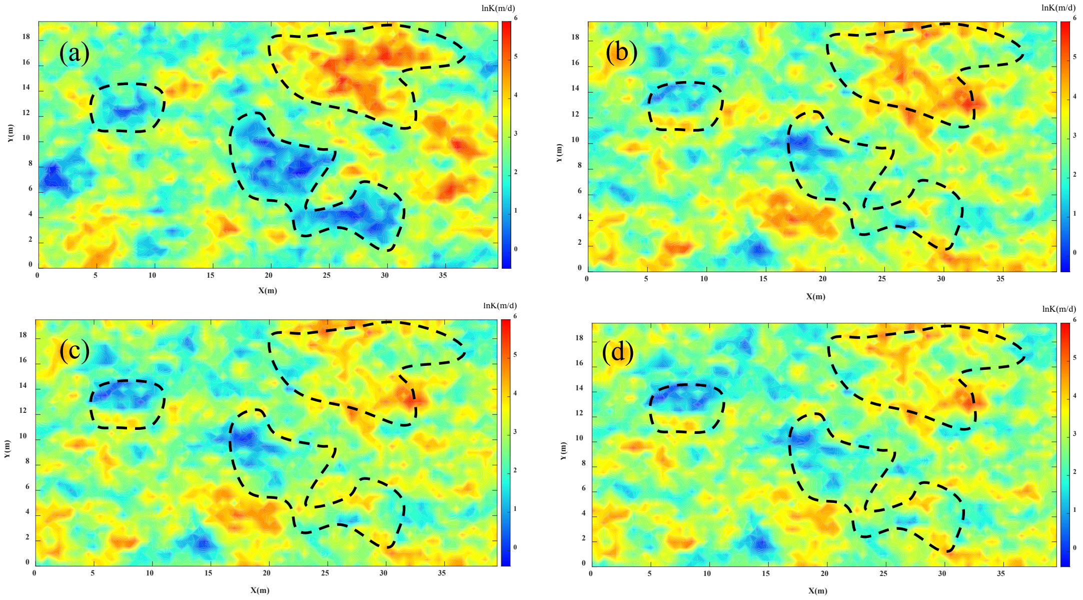

Fig. 8 (a) The reference K-field; (b)-(d) The corresponding interpolated K-field based on estimated KPP of data with 5%,10% and 15%error

Table 4 Comparison of actual and estimated SS using optimal surrogate models of S2. The last two lines are the RMSE for SS and KPP

Stress period	Source location (m) & Actual flux (g/s)	Estimated source flux (g/s)				
		Data with error				Missing data
		Error-free	5% error	10% error	15% error	
S_x	2.25	2.34	2.74	2.34	2.98	2.28
S_y	10.25	10.35	10.77	10.35	10.42	10.78
SP1	50	46.71	47.49	46.71	46.09	46.94
SP2	48	49.96	48.25	49.96	49.84	49.63
SP3	45	47.16	46.94	47.16	42.76	46.12
SP4	40	41.90	41.29	41.90	38.20	41.92
SP5	36	35.11	38.14	35.11	33.84	38.27
SP6	30	32.44	32.73	32.44	29.43	32.61
SP7	20	21.45	22.05	21.45	19.23	21.72
SP8	10	11.02	11.52	11.02	11.71	11.16
RMSE(SS)	-	1.81	1.76	1.81	1.90	1.83
RMSE(KPP)	-	1.33	1.35	1.33	1.33	1.38

Scenario 2

In this section, the missing observation data for the first three observation time $(t = 4,8,12)$ were considered, and the observation error level was set 5%. The inversion result of hydraulic conductivity field under incomplete observation data was shown in Fig. 9. In comparison to the reference K-field (Fig. 9a), the major low and high conductivity zones were effectively captured by ILUES-SOM model (Fig. 9b).

The identification results of unknown contaminant source under incomplete observation data were shown in last column of Table 4 and Fig. 10, it can be seen that the estimated values of contaminant source did not change significantly except for source flux in the 5_{th} stress period. Specially, for estimating SS and KPP, RMSE(SS) and RMSE(KPP) were increased by 0.02 (from 1.81 to 1.83) and 0.05 (from 1.33 to 1.38), respectively.

Therefore, the proposed ILUES-SOM based surrogate model showed satisfactory performance when the early observation data were missing. The identified contaminant source with and without missing data were similar to the actual values, and the estimated K-field under incomplete observation data could depict the morphological characteristics of reference conductivity field.

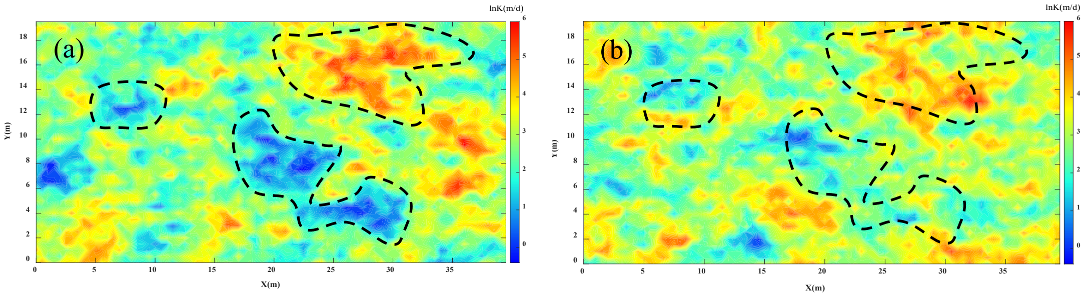

Fig. 9 (a) The reference K-field; (b)The corresponding interpolated K-field based on estimated KPP of missing data

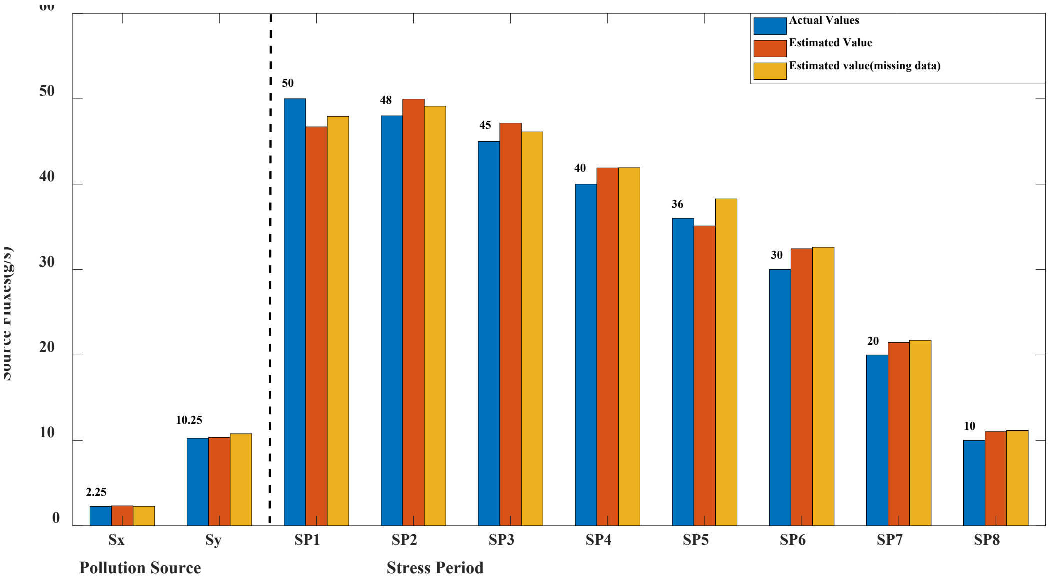

Fig. 10 Comparison of estimated SS by optimal surrogate models of S2

6. Conclusions

1. In this study, the proposed ILUES-SOM surrogate model was constructed for simultaneous inversion of hydraulic conductivity field and contaminant source parameters by combining the SGSIM-ILUES and SOM. Specifically, the inversion of hydraulic conductivity field was converted to the estimation of hydraulic conductivity at pilot points.

2. Considering the estimation accuracy and computational efficiency of the two SOM-based surrogate model, ILUES-SOM model was a better choice, which could not only ensure the parameter inversion

accuracy, but also significantly improve the computational efficiency. This indicated that the quality of training data can efficiently improve the performance of SOM-based surrogate model.

3. In terms of parameter inversion accuracy, ILUES-SOM model (1 iteration) was close to ILUES inverse model, but with significantly lower time cost. In other words, ILUES-SOM model has the qualities of fast inversion of SOM based surrogate model, while being able to achieve the inversion accuracy that can be achieved only by multiple iterations of ILUES inversion.

4. The proposed ILUES-SOM surrogate model for contaminant transport showed remarkable robustness. Varying degrees of observation errors were added to the limited observation data, and the estimation performance was still well and stable when the error level was under 15%. Even though early observation data were missing, the estimation precision of contaminant source was almost the same as that without missing, and the estimated K-field also could depict the morphological characteristics of reference conductivity field.

Author contribution:

Data statistics: Na Zheng: conceptualization, methodology, programming, writing—original draft; Jinbing Liu: methodology, programming tune; Xuemin Xia: case designing, data collecting; Simin Gu: data collection, validation; Xianwen Li: methodology, case designing; Simin Jiang: writing—review and editing, supervision, fund acquisition.

Funding This research was supported by the National Key R&D Program of China (2019YFE0114900), and the Belt and Road Special Foundation of the State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering (2019nkzd01).

Data availability Data and Matlab codes of this study are available upon request to the corresponding author.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no competing interests.

References

Asher MJ, Croke BFW, Jakeman AJ, Peeters LJM (2015) A review of surrogate models and their application to groundwater modeling. Water Resources Research, 51(8), 5957–5973. https://doi.org/10.1002/2015WR016967

Atmadja J, Ba Gtzoglou AC (2001) State of the Art Report on Mathematical Methods for Groundwater Pollution Source Identification. Environmental Forensics, 2(3), 205–214.

Bailey R, Baù D (2010) Ensemble smoother assimilation of hydraulic head and return flow data to estimate hydraulic conductivity distribution. Water Resources Research, 46(12). W12543, doi:10.1029/2010WR009147.

Bailey RT, Baù DA, Gates TK (2012) Estimating spatially-variable rate constants of denitrification in irrigated agricultural groundwater systems using an Ensemble Smoother. Journal of Hydrology, 468–469, 188–202. https://doi.org/10.1016/j.jhydrol.2012.08.033

Bao J, Li L, Redoloza F (2020) Coupling ensemble smoother and deep learning with generative adversarial networks to deal with non-Gaussianity in flow and transport data assimilation. Journal of Hydrology, 590, 125443. https://doi.org/10.1016/j.jhydrol.2020.125443

Cao Z, Li L, Chen K (2018) Bridging iterative Ensemble Smoother and multiple-point geostatistics for better flow and transport modeling. Journal of Hydrology, 565, 411–421.

https://doi.org/10.1016/j.jhydrol.2018.08.023

Chan S, Elsheikh AH (2020) Parametrization of Stochastic Inputs Using Generative Adversarial Networks With Application in Geology. Frontiers in Water, 2, 5. https://doi.org/10.3389/frwa.2020.00005

Chaudhary V, Bhatia RS, Ahlawat AK (2015) Community SOM (CSOM): An Improved Self-Organizing Map Learning Technique. International Journal of Fuzzy Systems, 17(2), 129–132. https://doi.org/10.1007/s40815-015-0022-7

Chen Y, Oliver DS (2012) Ensemble Randomized Maximum Likelihood Method as an Iterative Ensemble Smoother. Mathematical Geosciences, 44(1), 1–26. https://doi.org/10.1007/s11004-011-9376-z

Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research, 99(C5), 10143. https://doi.org/10.1029/94JC00572

Harbaugh AW, Banta ER, Hill MC, McDonald MG (2000) Modflow-2000, the u. s. geological survey modular ground-water model-user guide to modularization concepts and the ground-water flow process. Open-file report. U. S. Geological Survey (Open-File Report) (p. 134).

Hazrati YS, Datta B (2017) Self-organizing map based surrogate models for contaminant source identification under parameter uncertainty. International Journal of GEOMATE, 13(36). https://doi.org/10.21660/2017.36.2750

Hazrati YS, Datta B (2017) Adaptive Surrogate Model Based Optimization (ASMBO) for Unknown Groundwater Contaminant Source Characterizations Using Self-Organizing Maps. Journal of Water Resource and Protection, 09(02), 193–214. https://doi.org/10.4236/jwarp.2017.92014

He X, Li P, Wu J, Wei M, Ren X, Wang D (2021) Poor groundwater quality and high potential health risks in the Datong Basin, northern China: research from published data. Environmental Geochemistry and Health, 43(2), 791–812. https://doi.org/10.1007/s10653-020-00520-7

Jiang SM, Liu JB, Xia XM, Wang ZY, Cheng L, Li XW (2021) Simultaneous identification of contaminant sources and hydraulic conductivity field by combining geostatistics method with self-organizing maps algorithm. Journal of Contaminant Hydrology, 241, 103815. https://doi.org/10.1016/j.jconhyd.2021.103815

Jiang SM, Zhang RC, Liu JB, Xia XM, Li XW, Zheng MH (2022) Simultaneous Estimation of a Contaminant Source and Hydraulic Conductivity Field by Combining an Iterative Ensemble Smoother and Sequential Gaussian Simulation. Water, 14(5). https://doi.org/10.3390/w14050757

Liu JB, Jiang SM, Zhou NQ, Cai Y, Cheng L, Wang ZY (2021) Groundwater contaminant source identification based on QS-ILUES, 9(10):73-82. DOI: 10.19637/j.cnki.2305-7068.2021.01.007

Ju L, Zhang JJ, Meng L, Wu LS, Zeng LZ (2018) An adaptive Gaussian process-based iterative ensemble smoother for data assimilation. Advances in Water Resources, 115, 125–135. https://doi.org/10.1016/j.advwatres.2018.03.010

Kang XY, Kokkinaki A, Power C, Kitanidis PK, Shi XQ, Duan LM, Liu TX, Wu JC (2021) Integrating deep learning-based data assimilation and hydrogeophysical data for improved monitoring of DNAPL source zones during remediation. Journal of Hydrology, 601, 126655. https://doi.org/10.1016/j.jhydrol.2021.126655

Kohonen T (1982) Analysis of a simple self-organizing process. Biological Cybernetics, 44(2), 135–140. https://doi.org/10.1007/BF00317973

Li LP, Puzel R, Davis A (2018) Data assimilation in groundwater modelling: ensemble Kalman filter versus ensemble smoothers. Hydrological Processes, 32(13), 2020–2029. https://doi.org/10.1002/hyp.13127

Li LP, Zhou H, Gómez-Hernández JJ, Hendricks Franssen HJ (2012) Jointly mapping hydraulic conductivity and porosity by assimilating concentration data via ensemble Kalman filter. Journal of Hydrology, 428–429, 152–169. https://doi.org/10.1016/j.jhydrol.2012.01.037

Lima MM, Emerick AA, Ortiz CEP (2020) Data-space inversion with ensemble smoother. Computational Geosciences, 24(3), 1179–1200. https://doi.org/10.1007/s10596-020-09933-w

Penn BS (2005) Using self-organizing maps to visualize high-dimensional data. Computers & Geosciences, 31(5), 531–544. https://doi.org/10.1016/j.cageo.2004.10.009

Prakash O, Datta B (2013) Sequential optimal monitoring network design and iterative spatial estimation of pollutant concentration for identification of unknown groundwater pollution source locations. Environmental Monitoring and Assessment, 185(7), 5611–5626. https://doi.org/10.1007/s10661-012-2971-8

Schöniger A, Nowak W, Hendricks Franssen HJ (2012) Parameter estimation by ensemble Kalman filters with transformed data: Approach and application to hydraulic tomography: PARAMETER ESTIMATION BY tEnKFs. Water Resources Research, 48(4). https://doi.org/10.1029/2011WR010462

Simula O, Vesanto J, Alhoniemi E, Hollmn J (1998) Analysis and Modeling of Complex Systems Using the Self-Organizing Map. Kasabov N. & Kozma R.physica Verlag.

Tang M, Liu Y, Durlofsky LJ (2020) A deep-learning-based surrogate model for data assimilation in dynamic subsurface flow problems. Journal of Computational Physics, 413, 109456. https://doi.org/10.1016/j.jcp.2020.109456

Tang M, Liu Y, Durlofsky LJ (2021) Deep-learning-based surrogate flow modeling and geological parameterization for data assimilation in 3D subsurface flow. Computer Methods in Applied Mechanics and Engineering, 376, 113636. https://doi.org/10.1016/j.cma.2020.113636

Van Leeuwen PJ, Evensen G (1996) Data Assimilation and Inverse Methods in Terms of a Probabilistic Formulation. Monthly Weather Review, 124(12), 2898–2913. https://doi.org/10.1175/1520-0493(1996)124<2898:DAAIMI>2.0.CO;2

Xia XM, Zhou NQ, Wang L, Li XW, Jiang SM (2019) Identification of transient contaminant sources in aquifers through a surrogate model based on a modified self-organizing-maps algorithm. Hydrogeology Journal, 27(7), 2535–2550. https://doi.org/10.1007/s10040-019-02003-1

Yang AL, Jiang SM, Liu JB, Jing QY, Zhou T, Zhang W (2020) Groundwater contaminant source identification based on iterative local update ensemble smoother, 28(1), 3–11.

Zhang JJ, Lin G, Li WX, Wu LS, Zeng LZ (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. https://doi.org/10.1002/2017WR020906

Zhang RC, Zhou NQ, Xia XM, Zhao GX, Jiang SM (2020) Joint Estimation of Hydraulic and Biochemical Parameters for Reactive Transport Modelling with a Modified ILUES Algorithm. Water, 12(8), 2161. https://doi.org/10.3390/w12082161

Zheng CM, Wang PP (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. Ajr American Journal of Roentgenology, 169(4), 1196–7.

Zhong Z, Sun AY, Jeong H (2019) Predicting CO_2 Plume Migration in Heterogeneous Formations Using Conditional Deep Convolutional Generative Adversarial Network. Water Resources Research, 55(7), 5830–5851. https://doi.org/10.1029/2018WR024592