water

MDPI

Article

Groundwater Contamination Source Recognition Based on a Two-Stage Inversion Framework with a Deep Learning Surrogate

Zibo Wang [1,2,3] and Wenxi Lu [1,2,3,*]

[1] Key Laboratory of Groundwater Resources and Environment, Ministry of Education, Jilin University, Changchun 130021, China
[2] Jilin Provincial Key Laboratory of Water Resources and Water Environment, Jilin University, Changchun 130021, China
[3] College of New Energy and Environment, Jilin University, Changchun 130021, China
* Correspondence: luwenxi9966@163.com

Abstract: Groundwater contamination source recognition is an important prerequisite for subsequent remediation efforts. To overcome the limitations of single inversion methods, this study proposed a two-stage inversion framework by integrating two primary inversion approaches—simulation-optimization and simulation-data assimilation—thereby enhancing inversion accuracy. In the first stage, the ensemble smoother with multiple data assimilation method (a type of simulation-data assimilation) conducted a global broad search to provide better initial values and ranges for the second stage. In the subsequent stage, a collective decision optimization algorithm (a type of simulation-optimization) was used for a refined deep search, further enhancing the final inversion accuracy. Additionally, a deep learning method, the multilayer perceptron, was utilized to establish a surrogate of the simulation model, reducing computational costs. These theories and methods were applied and validated in a hypothetical scenario for the synchronous identification of the contamination source and boundary conditions. The results demonstrated that the proposed two-stage inversion framework significantly improved search accuracy compared to single inversion methods, with a mean relative error and mean absolute error of just 4.95% and 0.1756, respectively. Moreover, the multilayer perceptron surrogate model offered greater approximation accuracy to the simulation model than the traditional shallow learning surrogate model. Specifically, the coefficient of determination, mean relative error, mean absolute error, and root mean square error were 0.9860, 9.72%, 0.1727, and 0.47, respectively, highlighting its significant advantages. The findings of this study can provide more reliable technical support for practical case applications and improve subsequent remediation efficiency.

Keywords: groundwater contamination; source recognition; multilayer perceptron; ensemble smoother with multiple data assimilation; collective decision optimization algorithm

Citation: Wang, Z.; Lu, W. Groundwater Contamination Source Recognition Based on a Two-Stage Inversion Framework with a Deep Learning Surrogate. *Water* **2024**, *16*, 1907. https://doi.org/10.3390/w16131907

Academic Editors: Simin Jiang and Zhenbo Chang

Received: 11 June 2024
Revised: 28 June 2024
Accepted: 2 July 2024
Published: 3 July 2024

1. Introduction

Groundwater contamination, unlike surface water contamination, is often difficult to detect and, if untreated, poses significant risks to the environment and human health [1–3]. Therefore, designing an effective and efficient remediation strategy is critical. Groundwater contamination source recognition (GCSR) is essential for obtaining reliable contamination source information, which is a prerequisite for remediation efforts.

GCSR involves the inversion and solution of the groundwater contamination numerical simulation model based on actual water level and contaminant concentration monitoring data, along with auxiliary information such as site investigations and professional expertise. This process makes groundwater contamination visible by determining relevant contamination source information. Currently, the inversion approaches for GCSR are categorized into three types [4]: simulation-Bayesian inference (S-BI), simulation-data

assimilation (S-DA), and simulation-optimization (S-O). The S-BI is a probabilistic method known for its ability to consider the uncertainty of unknown variables. It not only provides point estimates but also offers comprehensive interval estimates. However, the search efficiency of S-BI methods may decrease when dealing with high-dimensional complex problems. S-DA methods, such as the ensemble Kalman filter, ensemble smoother (ES), and their variants, have recently gained popularity in GCSR due to their fast calculation speed and efficiency [5]. Among them, the ensemble smoother with multiple data assimilation (ES-MDA) is an iterative variant of the ES [6,7]. Its core idea is to use the covariance matrix representing unknown variables and system outputs to perform multiple data assimilation with monitoring data, thereby realizing iterative updates of unknown variables. Compared to the ES, the ES-MDA improves the approximation accuracy of nonlinear systems by performing multiple iterations. Studies have demonstrated the advantages of ES-MDA in parameter estimation [8–11]. However, this method has certain limitations, as it is based on linear estimation theory and the Gaussian assumption. Consequently, as system nonlinearity increases, the refined deep search capability of S-DA may diminish, and the inversion accuracy still needs to be improved. In contrast, the S-O method has strong search capabilities and is the most widely used in GCSR [4,12]. This method minimizes the objective function value by establishing and solving an optimization model until a set of unknown variable values is found [5]. The core of solving the optimization model lies in the optimization algorithms, which are divided into heuristic and non-heuristic types. Heuristic optimization algorithms, in particular, can avoid local optima and improve search efficiency to some extent, making them widely applied in GCSR [13–18]. These research results sufficiently demonstrated the potential of heuristic optimization algorithms. The collective decision optimization algorithm (CDOA) [19] is a novel heuristic optimization algorithm that simulates human collective decision-making behavior. The global optimal solution is achieved through information transmission and cooperation among individuals. The advantages of CDOA have been demonstrated in our previous study. Therefore, this work adopted the CDOA to solve the optimization model of GCSR. However, the S-O method has limitations, such as reliance on initial value selection. If the initial value is far from the reference value, it may decrease search efficiency and fall into the local optima, thereby affecting final accuracy [12,20].

Each of the aforementioned methods has its own advantages and disadvantages. Previous studies [21–24] have usually only used one method separately, making it easy to fall into the limitations of the methods themselves. To better capitalize on the strengths of each method, this work proposed a two-stage inversion framework that combined the two primary inversion techniques (S-O and S-DA), thereby enhancing recognition accuracy. In the first stage, the ES-MDA method took advantage of its fast speed and was employed for a global broad search to quickly obtain point and ensemble estimation results, thus providing better initial values and intervals of each unknown variable for the subsequent stage, effectively avoiding the limitation of the CDOA relying on initial value selection. Based on these results, in the second stage, the CDOA was applied for GCSR based on the optimization model, performing a refined deep search to achieve the final inversion results, to effectively compensate for the shortcomings of the weak refined search ability of the ES-MDA in dealing with complex nonlinear problems. To the best of our knowledge, this research is the first to combine ES-MDA and CDOA for GCSR. This approach breaks the constraints of previous research that relied on a single inversion method, offering a new perspective on inversion techniques, enriching the theoretical foundation of GCSR, and providing more reliable technical support for practical applications, thereby significantly improving subsequent remediation efficiency.

Additionally, whether employing the S-DA or S-O method for GCSR, a substantial number of iterations of the numerical simulation model are required, incurring high costs. To mitigate this, many scholars favor using surrogate models with lower costs to replace numerical simulation models for calculations [25–28]. Surrogate models are black box models that learn input–output mapping through large datasets. With the

rise of artificial intelligence, the modeling methods of surrogate models in GCSR have shifted from traditional shallow learning (SL) methods [18,29–31] to deep learning (DL) methods [32–35]. DL has demonstrated significant advantages in fitting complex nonlinear relationships. Among them, the multilayer perceptron (MLP) has gained a strong reputation in groundwater studies [36,37]. Therefore, this work utilized MLP to establish a surrogate of the simulation model, and the results were compared with those of the traditional SL surrogate model.

In summary, the contamination source information and boundary conditions were treated as unknown variables and were identified using the theories and methods mentioned above in this work. Their effectiveness was tested through a hypothetical case. The main contributions of this work can be summarized as follows: (1) For the first time, combining the ES-MDA and CDOA for GCSR, overcoming the limitations of using a single method and further improving the inversion accuracy. (2) Use the deep learning model in the two-stage inversion framework, effectively improving the approximation accuracy to the simulation model.

The following sections are structured as follows: Section 2 describes the principles of the methods used in this work. Section 3 presents the application and validation of these methods in a hypothetical scenario. Section 4 provides an in-depth discussion on this work. Finally, Section 5 summarizes the related conclusions.

2. Methodology

2.1. Simulation Model

The establishment of a reliable simulation model is fundamental for GCSR, encompassing both groundwater flow and solute transport simulation models. In a 2D steady-state-flow confined aquifer, the partial differential equations governing these models can be expressed as follows:

$$\frac{\partial}{\partial(x_i)}\left(T_{ij}\frac{\partial h}{\partial x_j}\right) = W \quad i,j = 1,2; \tag{1}$$

$$\frac{\partial(\theta C)}{\partial t} = \frac{\partial}{\partial x_i}\left(\theta D_{ij}\frac{\partial C}{\partial x_j}\right) - \frac{\partial}{\partial x_j}(\theta C u_i) + \frac{C_s W}{e} \quad i,j = 1,2. \tag{2}$$

Equations (1) and (2) are linked by Equation (3):

$$u_i = -\frac{K_{ij}}{\theta}\frac{\partial h}{\partial x_j} \quad i,j = 1,2 \tag{3}$$

where x_i and x_j are the coordinates; u_i is the actual average flow velocity; t is the time; h is the hydraulic head; e is the aquifer thickness; θ is the porosity; W is the volumetric flux per unit area (positive sign for inflow and negative sign for outflow); C is the contaminant concentration; C_s is the contaminant concentration in the sources or sinks; and K_{ij}, T_{ij}, and D_{ij} are the hydraulic conductivity tensor, transmissivity tensor, and hydrodynamic dispersion tensor, respectively.

The above partial differential equations, along with specific initial and boundary conditions (Equation (22)), constituted the groundwater flow and solute transport simulation models.

2.2. Low-Cost Surrogate Models

2.2.1. The KELM

The extreme learning machine (ELM), proposed by Huang et al. [38], is a single-hidden-layer feedforward neural network. When the input sample x_j is given, the ELM output can be calculated as follows:

$$\sum_{i=1}^{L} \beta_i \cdot p(\omega_i x_j + b_i) = f(x_j) = y_j \quad j = 1, 2, \ldots, m \tag{4}$$

where L represents the number of hidden layer neurons, β_i represents the connection weight between the i-th hidden layer neuron and the output layer neuron, $p(\cdot)$ is the activation function, $p(\omega_i x_j + b_i)$ represents the i-th hidden layer neuron output, ω_i represents the connection weight between the input layer neuron and the i-th hidden layer neuron, x_j represents the j-th input sample, and b_i represents the bias of the i-th hidden layer neuron. Equation (4) can also be expressed as follows:

$$Y = Q\beta; \tag{5}$$

$$\beta = Q^{+}T \tag{6}$$

where $\beta = [\beta_1, \beta_2, \ldots, \beta_L]^T$, $Y = [y_1, y_2, \ldots, y_m]^T$, Q is the output matrix of the hidden layer, Q^+ is the Moore–Penrose generalized inverse of Q, and $T = [t_1, t_2, \ldots, t_m]^T$ is the target data when the ELM can perform unbiased learning of training samples.

However, the ELM suffers from stability issues. Consequently, Huang [39] introduced a kernel function into the training process, replacing traditional random maps with kernel maps, and, thus, proposed the KELM. Compared to the ELM, the KELM produces more stable output results, as expressed below:

$$f(x) = h(x)Q^T\left(QQ^T + \frac{I}{Re}\right)^{-1}T = \begin{bmatrix} Ke(x, x_1) \\ Ke(x, x_2) \\ \vdots \\ Ke(x, x_m) \end{bmatrix}^T \left(\Omega_{ELM} + \frac{I}{Re}\right)^{-1}T, \tag{7}$$

$$\Omega_{ELM} = QQ^T, \tag{8}$$

$$\Omega_{ELM(i,j)} = h(x_i) \cdot h(x_j) = Ke(x_i, x_j) \tag{9}$$

where Re is the regularization coefficient, I is the identity matrix, $h(x)$ is the feature mapping function, Ω_{ELM} is the kernel matrix, and $Ke(u,v)$ is the kernel function.

In this study, the Gaussian RBF kernel function was selected, and the number of hidden layer neurons was set to 200.

2.2.2. The MLP

The MLP is a type of feedforward artificial neural network. Its structure (Figure 1) consists of an input layer, multiple hidden layers, and an output layer, providing strong nonlinear fitting capabilities [37]. The training process of the MLP involves two main stages: forward propagation and backward propagation. During forward propagation, data are inputted into the input layer and processed sequentially through each hidden layer. The prediction result is finally obtained at the output layer. The output of the m-th neuron in the l-th layer y_m^l can be calculated as follows:

$$\begin{cases} x_m^l = b_m^l + \sum_{i=1}^{k} \omega_{im}^{l-1} y_i^{l-1} \\ y_m^l = f(x_m^l) \end{cases} \tag{10}$$

where x_m^l is the corresponding input, b_m^l is the bias, ω_{im}^{l-1} is the connection weight between the m-th neuron in l-th layer and the i-th neuron in $(l-1)$-th layer, and $f(\cdot)$ is the activation function.

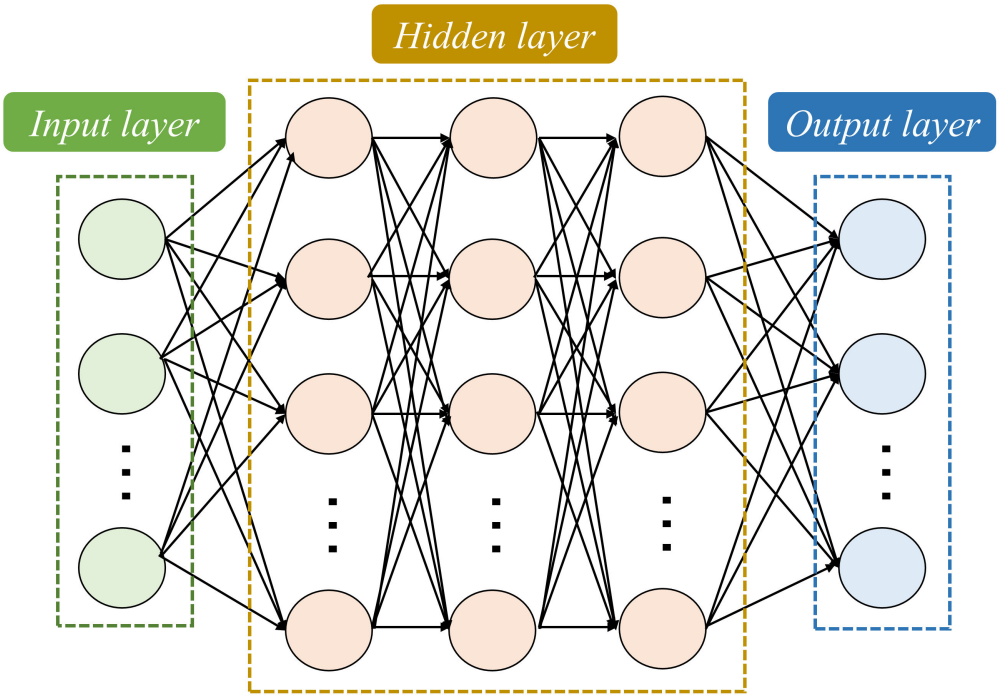

Figure 1. The general structure of the MLP.

Then, in the backward propagation stage, the error between the network's prediction result and the target result is calculated. The network parameters (ω and b) are continuously adjusted until the error is minimized. The update equations for the network parameters are shown as follows:

$$\begin{cases} \widetilde{\omega}_{im}^{l-1} = \omega_{im}^{l-1} - \eta \frac{\partial E}{\partial \omega_{im}^{l-1}} \\ \widetilde{b}_m^l = b_m^l - \eta \frac{\partial E}{\partial b_m^l} \end{cases} \tag{11}$$

where η is the learning rate and E is the loss function.

In this study, the MLP model was structured with one input layer, three hidden layers, and one output layer. The hidden layers contained 16, 64, and 128 neurons, respectively, and all used the Logistic Sigmoid activation function. The output layer employed the Purelin activation function. During the training process, the learning rate was set to 0.0025.

2.3. Inversion Framework

2.3.1. The ES-MDA

To enhance robustness in addressing nonlinear problems, the ES-MDA method was proposed. Unlike the ES method, the ES-MDA method performs multiple iterations. The core update equations are as follows:

$$\widetilde{X}_i^j = X_i^j + C_i^{XY}(C_i^{YY} + \alpha_i M)^{-1}(Z_{obs} + \sqrt{\alpha_i}\varepsilon^j - Y_i^j), \tag{12}$$

$$C_i^{XY} = \frac{1}{Ne-1}\sum_{j=1}^{Ne}(X_i^j - \overline{X}_i)(Y_i^j - \overline{Y}_i)^T, \tag{13}$$

$$C_i^{YY} = \frac{1}{Ne-1}\sum_{j=1}^{Ne}(Y_i^j - \overline{Y}_i)(Y_i^j - \overline{Y}_i)^T \tag{14}$$

where X_i^j, $\widetilde{X}_i^j$, Y_i^j are the j-th realization at the i-th iteration in the prior ensemble, posterior ensemble, and predicted ensemble, respectively; ε^j is the monitoring error, following the Gaussian distribution (the mean value is 0 and covariance matrix is M); Z_{obs} is the actual monitoring data; C_i^{XY} and C_i^{YY} are the cross-covariance matrix and auto-covariance matrix, respectively; $\overline{X}_i$, $\overline{Y}_i$ are the mean values of X_i, Y_i, respectively; Ne is the ensemble capacity; and α_i is the inflation factor. More extensive descriptions are provided by Emerick and Reynolds [40], Evensen [41], and Wang et al. [42].

In this work, when we used the ES-MDA for GCSR, the Ne and the number of iterations were set to 800 and 8, respectively.

2.3.2. The CDOA

The CDOA [19] is a population-based intelligent heuristic optimization algorithm, with its specific principles detailed in our previous work. In CDOA, for each individual, multiple candidate positions are generated through information transmission and cooperation within the population (Equations (15)–(21)). These candidate positions are then compared with the corresponding initial positions to select the best positions in the current iteration and update the corresponding population. This process is repeated until the iteration termination condition is satisfied, i.e., the maximum number of iterations T is reached. Finally, the iteration process is terminated, and the optimal solution is output.

1. Experience-based phase

$$\begin{cases} newX_{j0} = X_j(i) + a_0 \cdot step(i) \cdot d_0 \\ d_0 = \varphi_p - X_j(i) \end{cases} \tag{15}$$

2. Others-based phase

$$\begin{cases} newX_{j1} = newX_{j0} + a_1 \cdot step(i) \cdot d_1 \\ d_1 = r_1 \cdot d_0 + r_2 \cdot (X_o(i) - X_j(i)) \end{cases} \tag{16}$$

3. Group-thinking-based phase

$$\begin{cases} newX_{j2} = newX_{j1} + a_2 \cdot step(i) \cdot d_2 \\ d_2 = r_3 \cdot d_1 + r_4 \cdot (X_G - X_j(i)) \\ X_G = \frac{1}{N} \sum_{k=1}^{N} X_k(i) \end{cases} \tag{17}$$

4. Leader-based phase

$$\begin{cases} newX_{j3} = newX_{j2} + a_3 \cdot step(i) \cdot d_3 \\ d_3 = r_5 \cdot d_2 + r_6 \cdot (X_L - X_j(i)) \end{cases} \tag{18}$$

Additionally, a random walk strategy was used to change the position of X_L (Equation (19)), and the new X_L can be obtained by Equation (20).

$$newX_q = X_L + W_q \; q = 1, 2, 3, 4, 5 \tag{19}$$

$$X_L = newX_k \; k = \min\_\text{ObjFun}(newX_q) \tag{20}$$

5. Innovation-based phase

$$\begin{cases} newX_{j4} = newX_{j3} \\ r_7 \le MF \; x_{j4}^p = LB^p + r_8 \cdot (UB^p - LB^p) \end{cases} \tag{21}$$

where $X_j(i) = [x_j^1(i), x_j^2(i), \ldots, x_j^d(i)]$ $(j = 1, \ldots, N)$ is the position of the j-th individual in the population N at the i-th iteration; d is the dimension; a_0, a_1, a_2, a_3, and W_q are the random vectors in (0, 1), respectively; $step(i) = 2 - 1.7 \cdot (\frac{i-1}{T-1})$ is the step size at the i-th iteration; d_0,

d_1, d_2, and d_3 are the directions of movement, respectively; φ_p, $X_o(i)$, X_G, and X_L are the best position of the individual, the position of a randomly selected superior individual, geometric center position of all individuals, and the position of the best individual in the population, respectively; r_1, r_3, and r_5 are the random numbers in (−1, 1), respectively; r_2, r_4, and r_6 are the random numbers in (0, 2), respectively; r_7 and r_8 are the random numbers in (0, 1), respectively; MF is the innovation factor, p is randomly generated in the range [1, d]; and UB and LB are the upper and lower bounds, respectively. In CDOA, the objective function was used to evaluate the quality of individuals, and min_ObjFun is an indicator of the minimum objective function value.

When the CDOA was used for GCSR, an individual represented a set of the unknown variables in the inverse problem, and the position of the individual can be seen as the solution of the unknown variable. The process of finding the optimal position was the process of obtaining the optimal solution, i.e., the process of solving the inverse problem. In GCSR, the sum of squared errors between the surrogate model output data and the actual monitoring data was adopted as the objective function (Equation (29)). The smaller the objective function value, the closer the solution is to the reference value. In this work, the key parameters N and T were set to 75 and 900, respectively.

2.3.3. Two-Stage Inversion Framework

In this study, the ES-MDA and CDOA were combined to construct a two-stage inversion framework, improving inversion accuracy. The main steps are as follows:

Step 1. The first-stage inversion process: based on the specific principles provided in Section 2.3.1, the ES-MDA algorithm was employed for a global broad search to quickly obtain point estimation and ensemble estimation results, which served as the initial values and intervals for the next stage.

Step 2. The second-stage inversion process: based on the results of the first-stage inversion process, the S-O method was then applied for GCSR using the optimization model and CDOA, conducting a refined deep search to obtain the final results.

The main parameter settings of the two-stage inversion framework proposed in this study are shown in Sections 2.3.1 and 2.3.2, and the flowchart is depicted in Figure 2.

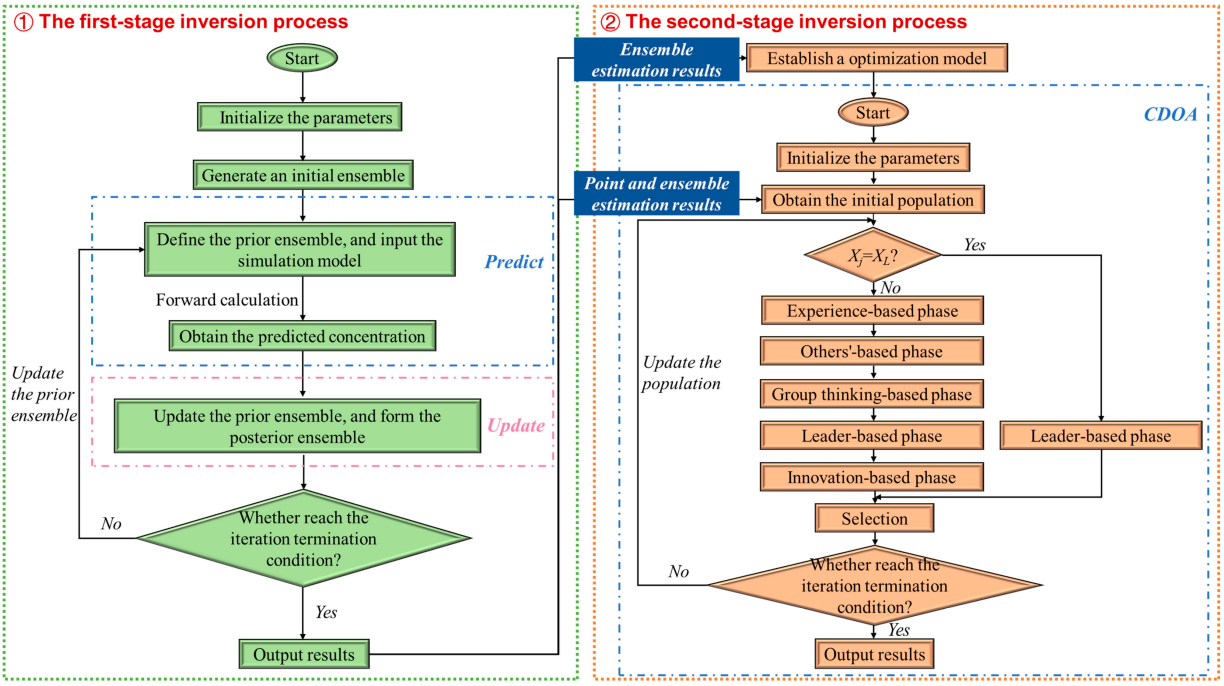

Figure 2. The two-stage inversion framework.

3. Site Overview

3.1. Case Description

In this study, a hypothetical numerical case was utilized to verify the effectiveness of the proposed methods and theories. This case was modified from Zhang et al. [43]. As depicted in Figure 3, the simulation area was a 2D confined aquifer with a size of 20 × 10 [L], featuring stable water flow from left to right. In the solute transport model, the left and right boundaries were seen as specified concentration boundaries, with the concentration at the left boundary being unknown. The expressions of initial and boundary conditions are shown in Equation (22). Additionally, to better characterize the uncertainty of hydrogeological parameters, the hydraulic conductivity field K was considered heterogeneous. The logarithmic form $Y = \ln K$ was used to represent the field K. The covariance function for any two points (x_a, y_a) and (x_b, y_b) is shown in Equation (23). In this work, we adopted the Karhunen–Loève (KL) expansion [44] to describe the reference Y field (Figure 4), as shown in Equation (24).

$$\begin{cases} C(x,y)|_{t=0} = C_0(x,y) & (x,y) \in \Omega \\ h(x,y)|_{\Gamma_2,\Gamma_4} = \phi(x,y,t) & (x,y) \in \Gamma_2,\Gamma_4,\ t \geq 0 \\ \frac{\partial h}{\partial \vec{n}}\Big|_{\Gamma_1,\Gamma_3} = 0 & (x,y) \in \Gamma_1,\Gamma_3,\ t \geq 0 \\ C(x,y)|_{\Gamma_2,\Gamma_4} = \varphi(x,y,t) & (x,y) \in \Gamma_2,\Gamma_4,\ t \geq 0 \\ \left(C\vec{u} - DgradC\right)\cdot\vec{n}\Big|_{\Gamma_1,\Gamma_3} = 0 & (x,y) \in \Gamma_1,\Gamma_3,\ t \geq 0 \end{cases} \tag{22}$$

where Ω is the simulation area, $\phi(x,y,t)$, $\varphi(x,y,t)$ are the known functions, $\vec{n}$ is the normal of the outer boundary, and $C_0(x,y)$ is the known function of the initial concentration.

$$C_Y(x_a,y_a;x_b,y_b) = \sigma_Y^2 \exp\left(-\frac{|x_a - x_b|}{l_x} - \frac{|y_a - y_b|}{l_y}\right), \tag{23}$$

$$Y(x,y) \approx \overline{Y}(x,y) + \sum_{i=1}^{N_{KL}} \sqrt{\tau_i} s_i(x,y)\zeta_i \tag{24}$$

where N_{KL} is the number of KL items, $\sigma_Y^2, \overline{Y}(x,y)$ are the variance and mean function, respectively, l_x, l_y are the correlation lengths in different directions, $\tau_i, s_i(x,y)$ are the eigenvalues and eigenfunctions, respectively, and ζ_i is a random variable that follows a Gaussian distribution with a mean of 0 and a standard deviation of 1. In this work, N_{KL}, σ_Y^2, $\overline{Y}(x,y)$ were set to 20, 0.4, and 2, respectively, and l_x and l_y were set to 10 and 5, respectively.

In the simulation area, one unknown contamination source and 15 known monitoring wells were present, with the total simulation duration set to 16 [T]. The contamination source continuously released a conservative contaminant throughout the simulation period. The release intensity, which varied over time ($t = i$: $i + 1$ [T], for $i = 1,\ldots,6$), was unknown. There were nine unknown variables in total, including the horizontal coordinate Sx and longitudinal coordinate Sy of the contamination source location, the release intensity RI_1~RI_6 for six time periods, and the left boundary concentration (BC). Their reference values are presented in Table 1. These reference values were input into the simulation model for forward calculation, yielding monitoring data for water levels and contaminant concentrations at the 15 monitoring wells. These data were considered the actual monitoring data in the hypothetical case, comprising a total of 90 dimensions: 15 dimensions for water level monitoring and 75 dimensions for concentration monitoring at five different times (t = 6, 8, 10, 12, and 14 [T]).

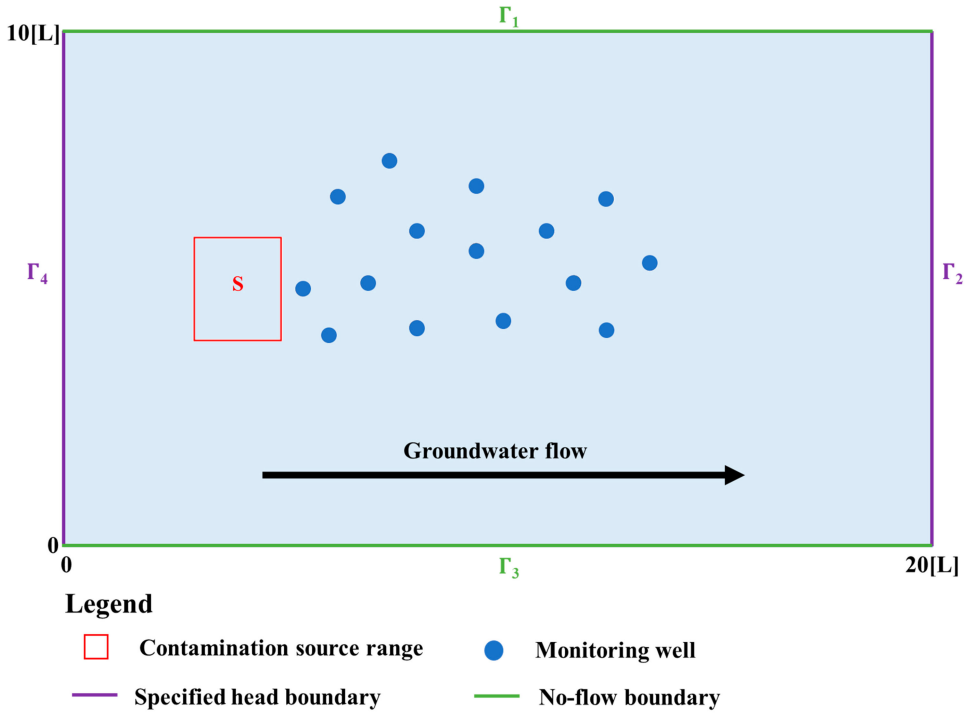

Figure 3. The schematic diagram of the simulation area.

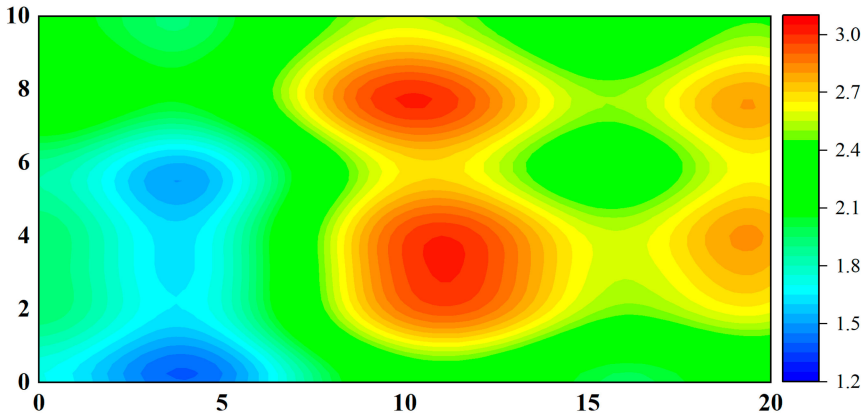

Figure 4. The reference *Y* field.

Table 1. Reference details of unknown variables.

Unknown Variable	Reference Value	Initial Range	Only the First-Stage Inversion Process Result	Only the Second-Stage Inversion Process Result	The Two-Stage Inversion Framework Result
Sx [L]	3.99	*U* [3, 5]	3.95	3.95	3.96
Sy [L]	4.71	*U* [4, 6]	4.72	4.72	4.71
RI_1 [MT^{-1}]	4.90	*U* [0, 8]	4.75	4.72	4.98
RI_2 [MT^{-1}]	3.65	*U* [0, 8]	3.72	3.99	3.46
RI_3 [MT^{-1}]	2.36	*U* [0, 8]	3.16	2.66	2.67
RI_4 [MT^{-1}]	6.70	*U* [0, 8]	5.00	5.46	6.38
RI_5 [MT^{-1}]	2.78	*U* [0, 8]	4.73	4.38	3.15
RI_6 [MT^{-1}]	7.76	*U* [0, 8]	6.41	6.68	7.50
BC [ML^{-3}]	0.84	*U* [0.6, 1.2]	0.88	0.90	0.86

3.2. Application of the Low-Cost Surrogate Model

In this section, we adopted the DL and SL methods, respectively, to establish the low-cost surrogate models of the simulation model. The accuracy of these two surrogate models was then compared to evaluate their respective advantages and applicability.

Step 1. Based on the initial distribution of each unknown variable (Table 1), sampling was conducted within the feasible domain of each variable, resulting in 550 sets of input samples. These samples were inputted into the simulation model for forward calculations, yielding 550 sets of output samples.

Step 2. The first 500 sets of input–output samples were designated as training samples, used to train the KELM and MLP surrogate models, respectively. The parameter settings are shown in Sections 2.2.1 and 2.2.2.

Step 3. The remaining 50 sets of input–output samples were used as testing samples to evaluate the approximation accuracy of the two surrogate models. Four accuracy indicators were selected for verification in this work: the coefficient of determination (R^2), mean relative error (*MRE*), mean absolute error (*MAE*), and root mean square error (*RMSE*):

$$R^2 = 1 - \frac{\sum_{k=1}^{l}\sum_{i=1}^{s}(y_{k,i} - \hat{y}_{k,i})^2}{\sum_{k=1}^{l}\sum_{i=1}^{s}(y_{k,i} - \overline{y})^2}, \tag{25}$$

$$MRE = \frac{\sum_{k=1}^{l}\sum_{i=1}^{s}\left|y_{k,i} - \hat{y}_{k,i}\right|/y_{k,i}}{l \cdot s}, \tag{26}$$

$$MAE = \frac{\sum_{k=1}^{l}\sum_{i=1}^{s}\left|y_{k,i} - \hat{y}_{k,i}\right|}{l \cdot s}, \tag{27}$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{l}\sum_{i=1}^{s}(y_{k,i} - \hat{y}_{k,i})^2}{l \cdot s}} \tag{28}$$

where l is the output dimension, s is the number of samples, $y_{k,i}$, $\hat{y}_{k,i}$ are the k-th dimension in the i-th sample of the simulation model and the low-cost surrogate model, respectively, and $\overline{y}$ is the average value of the simulation model output.

The final results of the two surrogate models are presented in Table 2. To illustrate the effectiveness more intuitively, fitting diagrams of the different surrogate models to the simulation model and relative error box plots were drawn, as shown in Figures 5 and 6. In Figure 5, the data points obtained by the MLP surrogate model were closer to the $y = x$ line, indicating a more concentrated and better fit to the simulation model. In Figure 6, data points within the 1.5 IQR range were considered normal, with 25% and 75% representing the lower and upper quartiles, respectively. Compared to the relative error results of the KELM surrogate model, the upper whisker, upper quartile, median line, and mean line of the MLP surrogate model were lower, demonstrating that the MLP surrogate model had higher prediction accuracy. Additionally, as shown in Table 2, each indicator result of the MLP surrogate model was superior to that of the KELM surrogate model. Thus, in this case, the DL surrogate model established by the MLP method exhibited advantages and overall better prediction performance. Therefore, the MLP surrogate model was used for GCSR.

Table 2. The performance of the different surrogate models.

Indicator	KELM Surrogate Model	MLP Surrogate Model
R^2	0.9577	0.9860
MRE (%)	13.07	9.72
MAE [ML^{-3}]	0.2379	0.1727
RMSE [ML^{-3}]	0.82	0.47

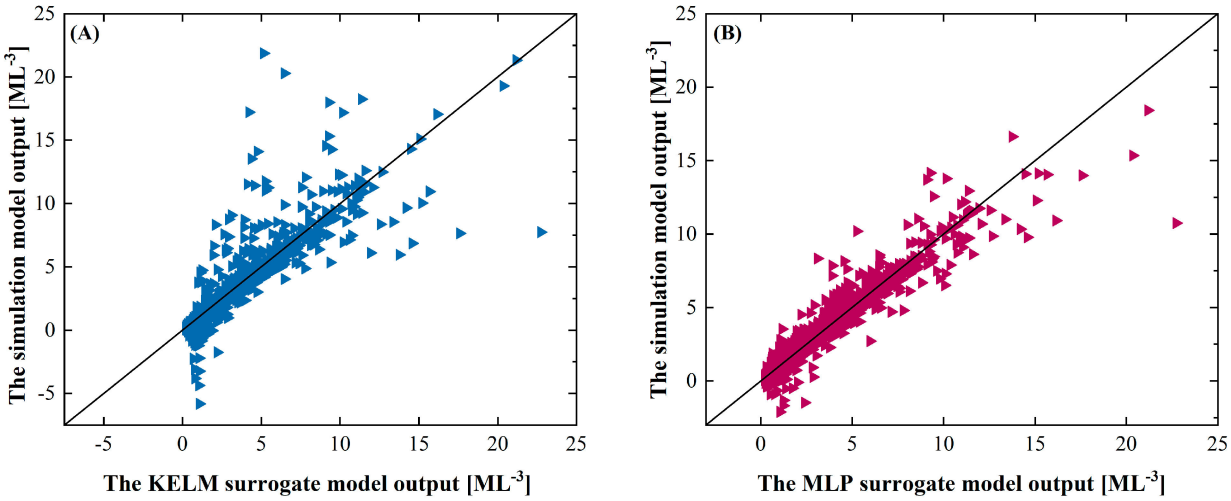

Figure 5. The fitting diagrams of the different surrogate models. (**A**) The KELM surrogate model and (**B**) The MLP surrogate model.

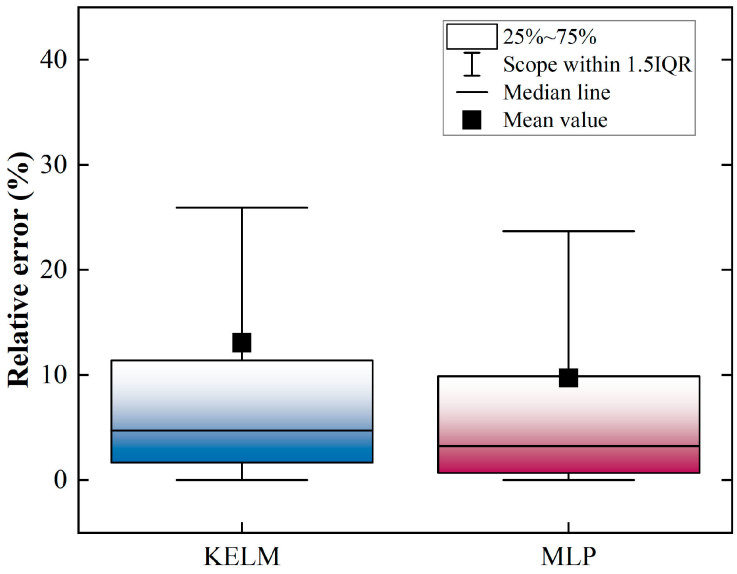

Figure 6. The relative error box plots of the different surrogate models.

3.3. Application of the Two-Stage Inversion Framework

Based on the MLP surrogate model, a two-stage inversion framework was employed. In the first stage, the fast ES-MDA method was utilized for a global broad search. The point estimation results from the first-stage inversion process are presented in Table 1, while the ensemble estimation results are depicted in Figure 7. It is evident that the variable distribution obtained by the ES-MDA method was generally concentrated and nearly encompassed the reference values. However, a few variables (e.g., RI_4~RI_6) exhibited a maximum posterior probability distribution that significantly deviated from the corresponding reference

values, indicating a need for improved inversion accuracy. Furthermore, based on the obtained ensemble estimation results, the upper and lower limits of each unknown variable were redefined to form a new search interval (i.e., optimized interval) for the second-stage inversion process. As shown in Figure 7, compared to the initial interval in Table 1, the optimized interval obtained from the first-stage inversion process was narrower, potentially providing strong support for the subsequent inversion stage.

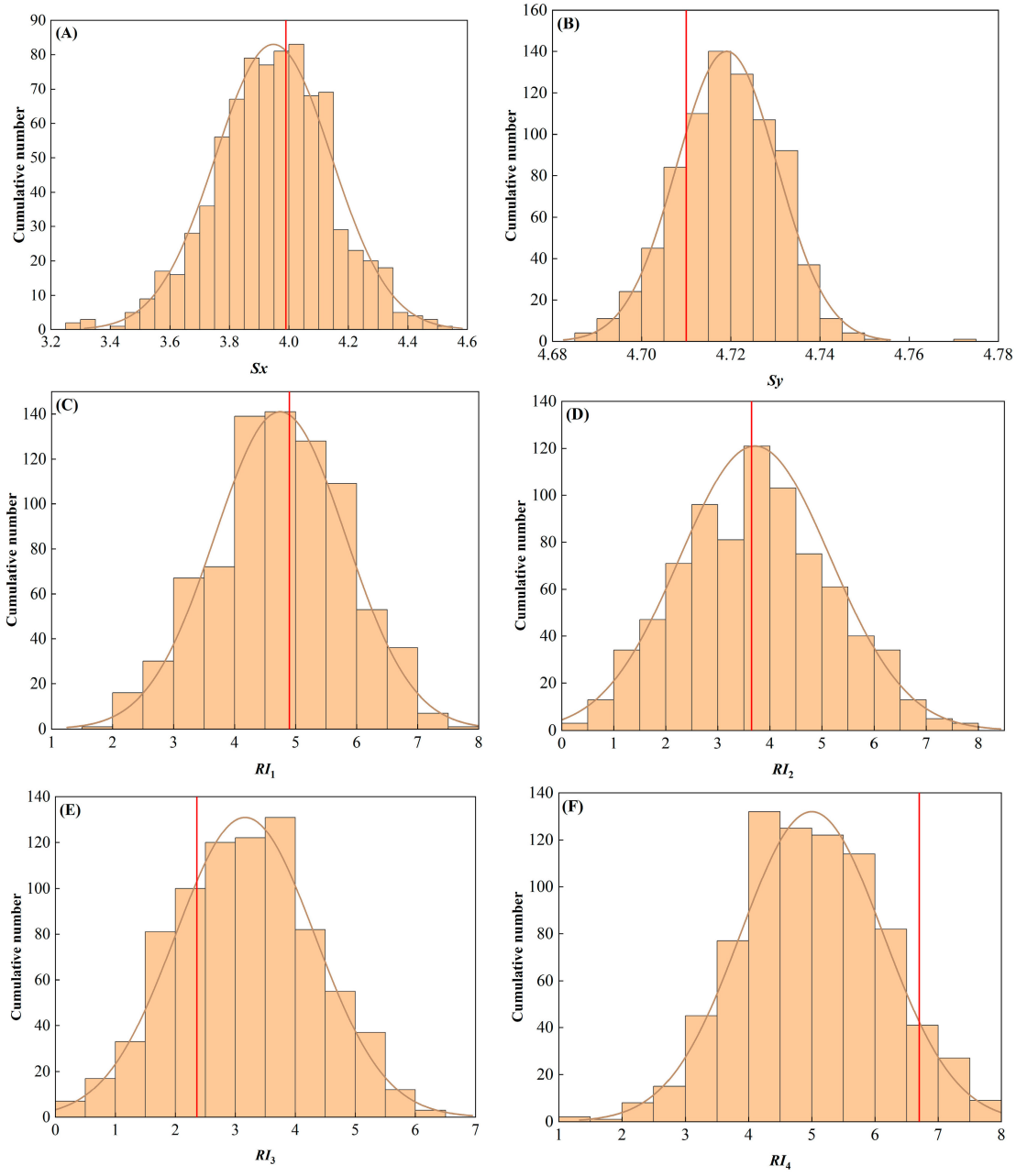

Figure 7. *Cont.*

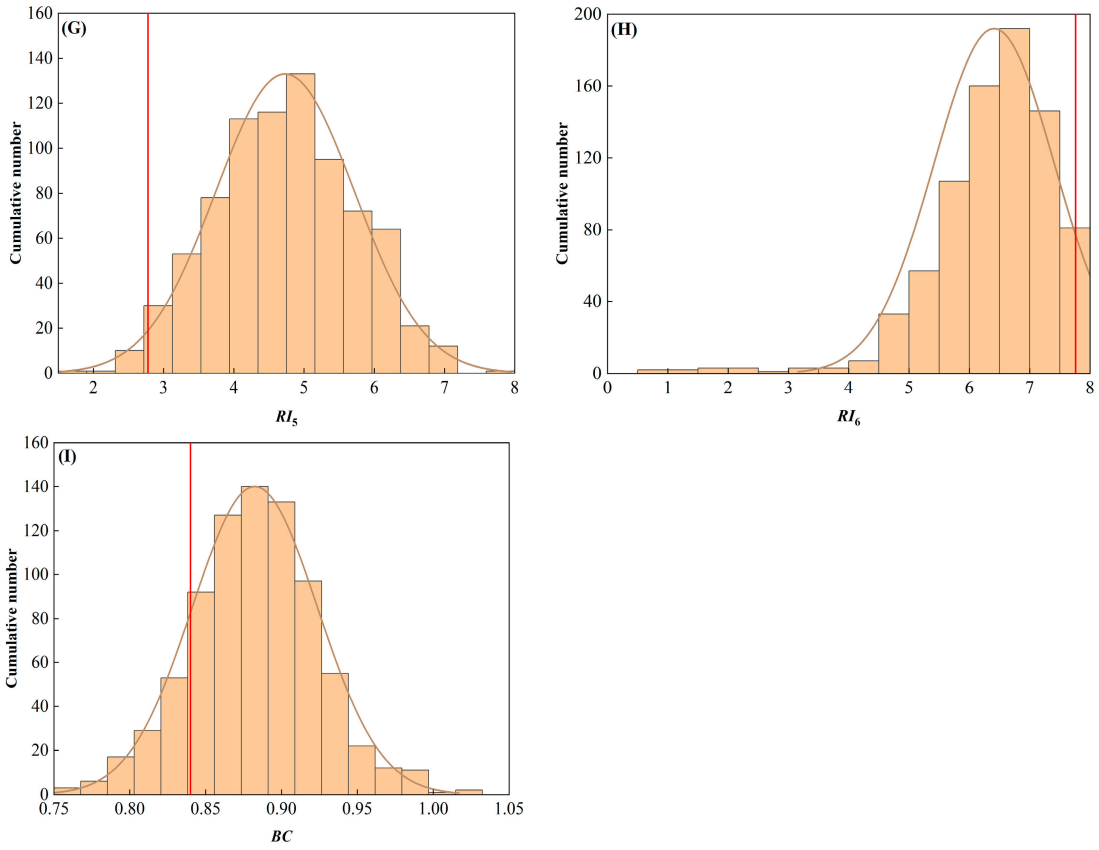

Figure 7. The ensemble distribution obtained in the first-stage inversion process. (**A**) Sx, (**B**) Sy, (**C**) RI_1, (**D**) RI_2, (**E**) RI_3, (**F**) RI_4, (**G**) RI_5, (**H**) RI_6, and (**I**) BC. The red line represents the corresponding reference value of each variable.

In the second-stage inversion process, the S-O method based on the CDOA was employed for a refined deep search. During this process, the inversion information from the first stage was fully utilized. Specifically, the values of each unknown variable were re-constrained within the optimized interval as inequality constraints, while the MLP surrogate model served as an equality constraint. The sum of squared errors between the surrogate model output data and the actual monitoring data was used as the objective function, establishing the optimization model (Equation (29)). Subsequently, we adopted the CDOA to solve the optimization model. Notably, the initial population was generated based on the optimized interval obtained from the first-stage inversion process, and the worst individual in the initial population was replaced by the point estimation results from the first stage. This provided better initial values and intervals, thereby accelerating the search efficiency and improving the inversion accuracy.

$$\min\left[\sum_{t=1}^{5}\sum_{k=1}^{15}\left(C_{k,t}-\widetilde{C}_{k,t}\right)^2+\sum_{k=1}^{15}\left(h_k-\widetilde{h}_k\right)^2\right]$$
$$s.t.\begin{cases}(h_k,C_{k,t})=g(Sx,Sy,RI_1,RI_2,RI_3,RI_4,RI_5,RI_6,BC)\\ Sx^{LB}\le Sx\le Sx^{UB}\\ Sy^{LB}\le Sy\le Sy^{UB}\\ RI_1^{LB}\le RI_1\le RI_1^{UB}\\ RI_2^{LB}\le RI_2\le RI_2^{UB}\\ RI_3^{LB}\le RI_3\le RI_3^{UB}\\ RI_4^{LB}\le RI_4\le RI_4^{UB}\\ RI_5^{LB}\le RI_5\le RI_5^{UB}\\ RI_6^{LB}\le RI_6\le RI_6^{UB}\\ BC^{LB}\le BC\le BC^{UB}\end{cases} \quad (29)$$

where h_k represents the water level of the k-th monitoring well obtained by the MLP surrogate model, $C_{k,t}$ represents the contaminant concentration of the k-th monitoring well at the t-th time obtained by the MLP surrogate model, and $\widetilde{h}_k$, $\widetilde{C}_{k,t}$ represent the actual water level and contaminant concentration, respectively.

Finally, the identification results obtained through the two-stage inversion framework are presented in Table 1. We employed the *MRE* and *MAE* to evaluate the performance. Compared to the first-stage inversion results, the identification results from the two-stage inversion framework were generally closer to the reference values. Further calculations indicated that the *MRE* and *MAE* of the inversion results from the first stage were 17.53% and 0.6789, respectively. However, upon completion of the two-stage inversion framework, the *MRE* and *MAE* decreased to 4.95% and 0.1756, respectively. This demonstrated that the introduction of the second stage for a refined deep search significantly enhanced the inversion accuracy. Additionally, a comparison was made between the final estimated results and those obtained using the second-stage inversion process alone. The convergence curves of the CDOA search process under these two scenarios are plotted in Figure 8, where a logarithmic scale was used for the vertical axis for better visualization. It is evident that the initial value of the CDOA based on the two-stage inversion framework was significantly better than that using CDOA alone, attributable to the first-stage inversion process. Moreover, the CDOA based on the two-stage inversion framework had a significantly smaller final objective function value compared to CDOA alone. This indicated that the two-stage inversion framework improved the search efficiency of the optimization algorithm and enhanced its ability to avoid local optima to some extent. For inversion accuracy, the degree of approximation of the identification results to the reference values under different inversion scenarios is plotted in Figure 9. From Table 1 and Figure 9, it can be seen that the introduction of the first-stage inversion process reduced the *MRE* from 13.78% to 4.95%, and the *MAE* was also reduced from 0.5389 to 0.1756. Compared to using the second-stage inversion process alone, the preliminary global broad search of the first stage proved to be beneficial and effective. The first-stage inversion process provided better initial values and optimized intervals for each unknown variable, thereby enhancing the accuracy of the subsequent inversion stage.

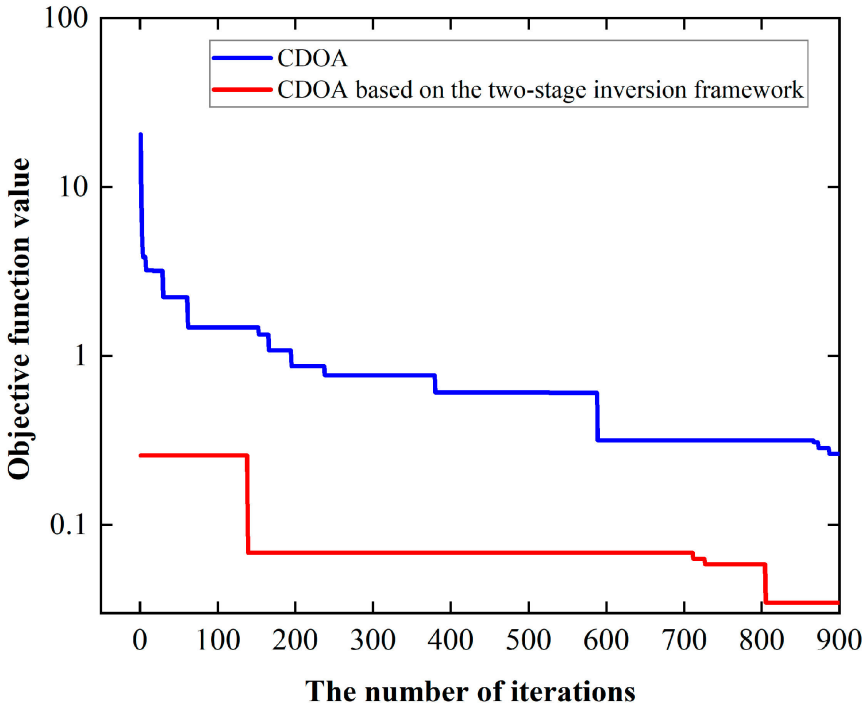

Figure 8. The convergence curves of the different inversion situations.

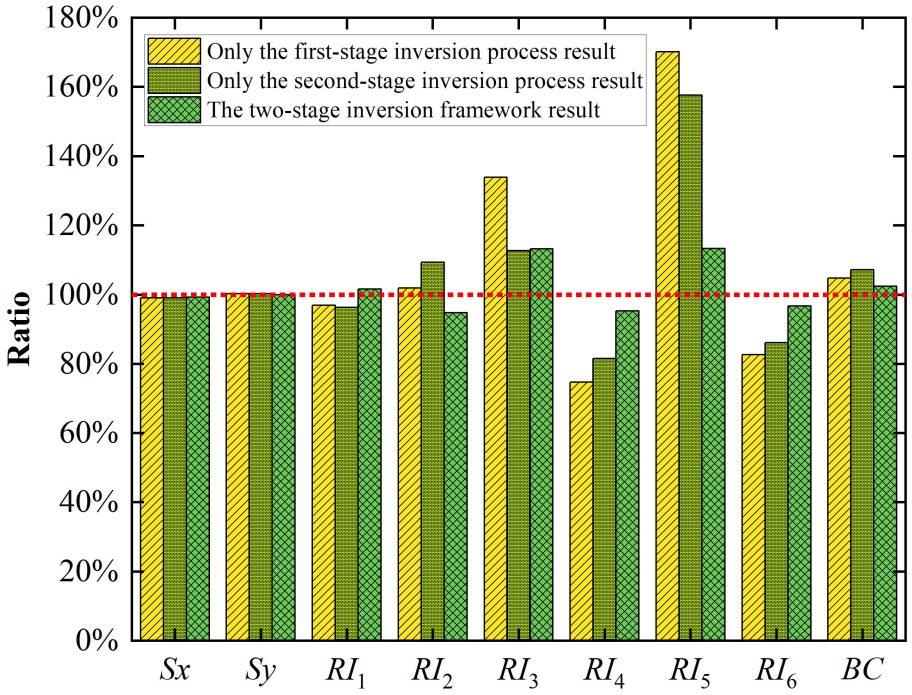

Figure 9. The performance of the different inversion situations. The red dotted line indicates that the identification value is 100% close to the reference value.

In summary, the two-stage inversion framework proposed in this study significantly improved inversion accuracy compared to the single-stage inversion process, proving to be both effective and feasible.

4. Discussion

Several new insights were gained during the research process. First, it was demonstrated that, compared to the traditional SL surrogate model, the DL surrogate model significantly enhances the approximation accuracy of the simulation model. The result was consistent with that of previous research [45,46]. However, deep neural networks often require the setting of more hyperparameters (such as neurons and network layers), which greatly influence prediction accuracy. In this study, these hyperparameters were manually adjusted, a process that is both time-consuming and labor-intensive. Some existing studies [4,32] have shown that using the hyperparameter optimization strategy can effectively replace previous manual tuning and improve efficiency. While this was not the focus of the current study, it represents a valuable research avenue for future exploration. Furthermore, although the DL method has advantages over the SL method, it entails longer and more complex training processes. If the model input dimensions are not excessively high and the mapping relationship between input and output is not overly complex, the SL surrogate model may suffice. Therefore, an optimal choice between DL and SL surrogate models should be made based on the specific research needs and actual conditions, to better balance training time and approximation accuracy.

For the inversion method, we adopted a two-stage inversion framework for GCSR. The research results indicated that this combination strategy achieved higher inversion accuracy compared to a single method. Therefore, the combination inversion strategy has certain development potential. However, it is important to note that combining different methods may require more time than using a single method. In this study, for example, under equivalent parameter settings on a 2.60 GHz dual-core E5 CPU and 96 GB RAM PC platform, the two-stage inversion framework required approximately 8740s, whereas the CDOA alone required about 8681 s. Although the former takes slightly more time, the difference is not significant. We believe that sacrificing a small amount of time for improved inversion accuracy is acceptable. In practical groundwater contamination scenarios, if

the two-stage inversion framework proposed in this work can enhance accuracy, more precise contamination source information can be obtained, allowing for swift remediation and treatment measures. Consequently, the improvement in inversion accuracy plays a crucial role in subsequent groundwater contamination remediation plans, risk assessment, and the accurate determination of contamination responsibilities, thereby significantly enhancing the efficiency of subsequent efforts. This is of great significance for real-world environmental management and remediation efforts.

Additionally, several limitations and uncertainties are present in this study. First, a hypothetical case modified from previous research [43] was adopted. The model assumed the actual monitoring data quality was perfect. However, in real-world scenarios, data quality may be imperfect and contain some errors. Although our previous study [4] demonstrated that the ES-MDA in the two-stage inversion framework has strong noise resistance, it is hypothesized that the two-stage inversion framework may inherit this advantage and have a certain degree of noise resistance. However, this hypothesis was not tested in this study, as robustness testing with monitoring data of varying noise levels was not conducted. The impact of noise on the accuracy of the proposed inversion framework remains uncertain and should be addressed in future research. Secondly, in our case, the K field was known, while the contamination source information and boundary conditions were treated as unknown variables. In practical applications, many variables (such as the K field) cannot be reliably estimated in advance, potentially requiring more variables to be considered unknown. Moreover, actual case simulation models are more complex, posing greater challenges for inversion techniques. Therefore, future work will focus on applying this framework to high-dimensional and complex practical cases to further test its efficacy in addressing real-world problems.

5. Conclusions

In this study, a two-stage inversion framework was proposed to synchronously identify groundwater contamination source information and boundary conditions. The results were compared with those obtained from single methods. To reduce computational costs, an MLP surrogate model was constructed to replace the simulation model in completing the above inversion task. The effectiveness of the MLP surrogate model was further compared and analyzed against that of the KELM surrogate model. These theories and methods were tested using a hypothetical site, leading to the following conclusions:

(1) In the two-stage inversion framework, the global broad search in the first-stage inversion process effectively provided better initial values and optimized intervals for the subsequent stage, accelerating search efficiency and improving identification accuracy. The introduction of the second-stage inversion process for a refined deep search further enhanced the inversion effect. Therefore, compared to using the first and second stages separately, the proposed two-stage inversion framework significantly improved inversion accuracy and effectiveness.

(2) Compared to the KELM surrogate model, the R^2, *MRE*, *MAE*, and *RMSE* metrics of the MLP surrogate model were significantly better, demonstrating the improved approximation accuracy of the simulation model and proving its advantages.

Author Contributions: Conceptualization, Z.W. and W.L.; Formal analysis, Z.W.; Methodology, Z.W.; Project administration, W.L.; Resources, W.L.; Software, Z.W.; Supervision, W.L.; Validation, Z.W.; Writing—original draft, Z.W.; Writing—review and editing, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 42272283).

Data Availability Statement: Data are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shi, C.; Zhao, Y.; Lu, W. Application of stochastic programming in groundwater pollution source identification. *Environ. Forensics* **2021**, *23*, 170–178. [CrossRef]
2. Covaciu, D.C.; Balint, A.C.; Neamtu, C.V.; Mosneag, S.C.; Bordea, D.; Dirjan, S.; Odagiu, A.C.M. Assessment of Groundwater Quality in Relation to Organic versus Mineral Fertilization. *Water* **2023**, *15*, 2895. [CrossRef]
3. Yang, C.; Dong, J.; Ren, L.; Fan, Y.; Li, B.; Hu, W. Influencing factors on the stabilization of colloid biliquid aphrons and its effectiveness used for density modification of DNAPLs in subsurface environment. *Colloid Surf. A-Physicochem. Eng. Asp.* **2018**, *553*, 439–445. [CrossRef]
4. Chang, Z.; Guo, Z.; Chen, K.; Wang, Z.; Zhan, Y.; Lu, W.; Zheng, C. A comparison of inversion methods for surrogate-based groundwater contamination source identification with varying degrees of model complexity. *Water Resour. Res.* **2024**, *60*, e2023WR036051. [CrossRef]
5. Chen, Z.; Zong, L.; Gómez-Hernández, J.J.; Xu, T.; Jiang, Y.; Zhou, Q.; Yang, H.; Jia, Z.; Mei, S. Contaminant source and aquifer characterization: An application of ES-MDA demonstrating the assimilation of geophysical data. *Adv. Water Resour.* **2023**, *181*, 104555. [CrossRef]
6. Kim, S.; Jung, H.; Choe, J. Enhanced History Matching of Gas Reservoirs with an Aquifer Using the Combination of Discrete Cosine Transform and Level Set Method in ES-MDA. *J. Energy Resour. Technol.-Trans. ASME* **2019**, *141*, 072906. [CrossRef]
7. Todaro, V.; D'Oria, M.; Tanda, M.G.; Gómez-Hernández, J.J. Ensemble smoother with multiple data assimilation to simultaneously estimate the source location and the release history of a contaminant spill in an aquifer. *J. Hydrol.* **2021**, *598*, 126215. [CrossRef]
8. Silva, T.M.D.; Bela, R.V.; Pesco, S.; Barreto, A. ES-MDA applied to estimate skin zone properties from injectivity tests data in multilayer reservoirs. *Comput. Geosci.* **2021**, *146*, 104635. [CrossRef]
9. Todaro, V.; D'Oria, M.; Zanini, A.; Gómez-Hernández, J.J.; Tanda, M.G. Experimental sandbox tracer tests to characterize a two-facies aquifer via an ensemble smoother. *Hydrogeol. J.* **2023**, *31*, 1665–1678. [CrossRef]
10. Cui, F.; Bao, J.C.; Cao, Z.D.; Li, L.P.; Zheng, Q. Soil hydraulic parameters estimation using ground penetrating radar data via ensemble smoother with multiple data assimilation. *J. Hydrol.* **2020**, *583*, 124552. [CrossRef]
11. Xu, T.; Zhang, W.; Gómez-Hernández, J.J.; Xie, Y.; Yang, J.; Chen, Z.; Lu, C. Non-point contaminant source identification in an aquifer using the ensemble smoother with multiple data assimilation. *J. Hydrol.* **2022**, *606*, 127405. [CrossRef]
12. Wang, Z.; Lu, W.; Chang, Z.; Wang, H. Simultaneous identification of groundwater contaminant source and simulation model parameters based on an ensemble Kalman filter—Adaptive step length ant colony optimization algorithm. *J. Hydrol.* **2022**, *605*, 127352. [CrossRef]
13. Wang, Z.; Lu, W.; Chang, Z.; Luo, J. A combined search method based on a deep learning combined surrogate model for groundwater DNAPL contamination source identification. *J. Hydrol.* **2023**, *616*, 128854. [CrossRef]
14. Hou, Z.; Lao, W.; Wang, Y.; Lu, W. Hybrid homotopy-PSO global searching approach with multi-kernel extreme learning machine for efficient source identification of DNAPL-polluted aquifer. *Comput. Geosci.* **2021**, *155*, 104837. [CrossRef]
15. Jiang, S.; Zhang, Y.; Wang, P.; Zheng, M. An almost-parameter-free harmony search algorithm for groundwater pollution source identification. *Water Sci. Technol.* **2013**, *68*, 2359–2366. [CrossRef]
16. Aral, M.M.; Guan, J.; Maslia, M.L. Identification of contaminant source location and release history in aquifers. *J. Hydrol. Eng.* **2001**, *6*, 225–234. [CrossRef]
17. Chakraborty, A.; Prakash, O. Identification of clandestine groundwater pollution sources using heuristics optimization algorithms: A comparison between simulated annealing and particle swarm optimization. *Environ. Monit. Assess.* **2020**, *192*, 791. [CrossRef]
18. Zhao, Y.; Qu, R.; Xing, Z.; Lu, W. Identifying groundwater contaminant sources based on a KELM surrogate model together with four heuristic optimization algorithms. *Adv. Water Resour.* **2020**, *138*, 103540. [CrossRef]
19. Zhang, Q.; Wang, R.; Yang, J.; Ding, K.; Li, Y.; Hu, J. Collective decision optimization algorithm: A new heuristic optimization method. *Neurocomputing* **2017**, *221*, 123–137. [CrossRef]
20. Li, J.; Lu, W.; Fan, Y. Groundwater pollution sources identification based on hybrid homotopy-genetic algorithm and simulation optimization. *Environ. Eng. Sci.* **2021**, *38*, 777–788. [CrossRef]
21. Ayvaz, M.T. A linked simulation-optimization model for solving the unknown groundwater pollution source identification problems. *J. Contam. Hydrol.* **2010**, *117*, 46–59. [CrossRef] [PubMed]
22. Xu, T.; Gomez-Hernandez, J.J. Joint identification of contaminant source location, initial release time, and initial solute concentration in an aquifer via ensemble Kalman filtering. *Water Resour. Res.* **2016**, *52*, 6587–6595. [CrossRef]
23. Han, K.; Zuo, R.; Ni, P.; Xue, Z.; Xu, D.; Wang, J.; Zhang, D. Application of a genetic algorithm to groundwater pollution source identification. *J. Hydrol.* **2020**, *589*, 125343. [CrossRef]
24. Bai, Y.; Lu, W.; Li, J.; Chang, Z.; Wang, H. Groundwater contamination source identification using improved differential evolution Markov chain algorithm. *Environ. Sci. Pollut. Res.* **2022**, *29*, 19679–19692. [CrossRef] [PubMed]
25. Wang, H.; Lu, W.; Chang, Z.; Li, J. Heuristic search strategy based on probabilistic and geostatistical simulation approach for simultaneous identification of groundwater contaminant source and simulation model parameters. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 891–907. [CrossRef]
26. Anshuman, A.; Eldho, T.I. A parallel workflow framework using encoder-decoder LSTMs for uncertainty quantification in contaminant source identification in groundwater. *J. Hydrol.* **2023**, *619*, 129296. [CrossRef]

27. Xing, Z.; Qu, R.; Zhao, Y.; Fu, Q.; Ji, Y.; Lu, W. Identifying the release history of a groundwater contaminant source based on an ensemble surrogate model. *J. Hydrol.* **2019**, *572*, 501–516. [CrossRef]
28. Luo, J.; Liu, Y.; Li, X.; Xin, X.; Lu, W. Inversion of groundwater contamination source based on a two-stage adaptive surrogate model-assisted trust region genetic algorithm framework. *Appl. Math. Model.* **2022**, *112*, 262–281. [CrossRef]
29. Yan, X.; Dong, W.; An, Y.; Lu, W. A Bayesian-based integrated approach for identifying groundwater contamination sources. *J. Hydrol.* **2019**, *579*, 124160. [CrossRef]
30. Chang, Z.; Lu, W.; Wang, H.; Li, J.; Luo, J. Simultaneous identification of groundwater contaminant sources and simulation of model parameters based on an improved single-component adaptive Metropolis algorithm. *Hydrogeol. J.* **2021**, *29*, 859–873. [CrossRef]
31. Li, J.; Lu, W.; Wang, H.; Bai, Y.; Fan, Y. Groundwater contamination sources identification based on kernel extreme learning machine and its effect due to wavelet denoising technique. *Environ. Sci. Pollut. Res.* **2020**, *27*, 34107–34120. [CrossRef] [PubMed]
32. Chang, Z.; Lu, W.; Wang, Z. Study on source identification and source-sink relationship of LNAPLs pollution in groundwater by the adaptive cyclic improved iterative process and Monte Carlo stochastic simulation. *J. Hydrol.* **2022**, *612*, 128109. [CrossRef]
33. Pan, Z.; Lu, W.; Bai, Y. Groundwater contamination source estimation based on a refined particle filter associated with a deep residual neural network surrogate. *Hydrogeol. J.* **2022**, *30*, 881–897. [CrossRef]
34. Xia, X.; Jiang, S.; Zhou, N.; Cui, J.; Li, X. Groundwater contamination source identification and high-dimensional parameter inversion using residual dense convolutional neural network. *J. Hydrol.* **2023**, *617*, 129013. [CrossRef]
35. Mo, S.; Zabaras, N.; Shi, X.; Wu, J. Deep autoregressive neural networks for high-dimensional inverse problems in groundwater contaminant source identification. *Water Resour. Res.* **2019**, *55*, 3856–3881. [CrossRef]
36. Müller, J.; Park, J.; Sahu, R.; Varadharajan, C.; Arora, B.; Faybishenko, B.; Agarwal, D. Surrogate optimization of deep neural networks for groundwater predictions. *J. Glob. Optim.* **2021**, *81*, 203–231. [CrossRef]
37. Li, Y.; Lu, W.; Pan, Z.; Wang, Z.; Dong, G. Simultaneous identification of groundwater contaminant source and hydraulic parameters based on multilayer perceptron and flying foxes optimization. *Environ. Sci. Pollut. Res.* **2023**, *30*, 78933–78947. [CrossRef]
38. Huang, G.; Zhu, Q.; Siew, C.K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]
39. Huang, G. An insight into extreme learning machines: Random neurons, random features and kernels. *Cogn. Comput.* **2014**, *6*, 376–390. [CrossRef]
40. Emerick, A.A.; Reynolds, A.C. Ensemble smoother with multiple data assimilation. *Comput. Geosci.* **2013**, *55*, 3–15. [CrossRef]
41. Evensen, G. Analysis of iterative ensemble smoothers for solving inverse problems. *Comput. Geosci.* **2018**, *22*, 885–908. [CrossRef]
42. Wang, Z.; Lu, W.; Chang, Z.; Zhang, T. Joint identification of groundwater pollution source information, model parameters, and boundary conditions based on a novel ES-MDA with a wheel battle strategy. *J. Hydrol.* **2024**, *636*, 131320. [CrossRef]
43. Zhang, J.; Zheng, Q.; Chen, D.; Wu, L.; Zeng, L. Surrogate-Based Bayesian Inverse Modeling of the Hydrological System: An Adaptive Approach Considering Surrogate Approximation Error. *Water Resour. Res.* **2020**, *56*, e2019WR025721. [CrossRef]
44. Zhang, D.; Lu, Z. An efficient, high-order perturbation approach for flow in random porous media via Karhunen-Loeve and polynomial expansions. *J. Comput. Phys.* **2004**, *194*, 773–794. [CrossRef]
45. Li, J.; Lu, W.; Luo, J. Groundwater contamination sources identification based on the Long-Short Term Memory network. *J. Hydrol.* **2021**, *601*, 126670. [CrossRef]
46. Wang, H.; Lu, W. Groundwater contamination source-sink analysis based on random statistical method for a practical case. *Stoch. Environ. Res. Risk Assess.* **2022**, *36*, 4157–4174. [CrossRef]