water

MDPI

Article

An Integrated Bayesian and Machine Learning Approach Application to Identification of Groundwater Contamination Source Parameters

Yongkai An [1,2], Yanxiang Zhang [3] and Xueman Yan [4,*]

[1] Key Laboratory of Subsurface Hydrology and Ecological Effects in Arid Region of the Ministry of Education, Chang'an University, No. 126 Yanta Road, Xi'an 710054, China
[2] School of Water and Environment, Chang'an University, No. 126 Yanta Road, Xi'an 710054, China
[3] Power China Northwest Engineering Corporation Limited, Xi'an 710065, China
[4] College of Urban and Environmental Sciences, Northwest University, Xi'an 710027, China
* Correspondence: yanxm666@126.com

Abstract: The identification of groundwater contamination source parameters is an important prerequisite for the control and risk assessment of groundwater contamination. This study developed an innovative approach for the optimal design of observation well locations and the high-precision identification of groundwater contamination source parameters. The approach involves Bayesian theory and integrates Markov Chain Monte Carlo, Bayesian design, information entropy, machine learning, and surrogate modeling. The optimal observation well locations are determined by information entropy, which is adopted to mine valuable information about unknown groundwater contamination source parameters from measurements of contaminant concentration according to Bayesian design. After determining the optimal observation well locations, the identification of groundwater contamination source parameters is implemented through a Bayesian-based Differential Evolution Adaptive Metropolis with Discrete Sampling–Markov Chain Monte Carlo approach. However, the processes of both determination and identification are time-consuming because the original simulation model (that is, the contaminant transport model) needs to be invoked multiple times. To overcome this challenge, a machine learning approach, that is, Multi-layer Perceptron, is used to build a surrogate model for the original simulation model, which can greatly accelerate the determination and identification processes. Finally, two hypothetical numerical case studies involving homogeneous and heterogeneous cases are used to verify the performance of the proposed approach. The results show that the optimal design of observation well locations and high-precision identification of groundwater contamination source parameters can be implemented accurately and effectively by using the proposed approach. In summary, this study highlights that the integrated Bayesian and machine learning approach provides a promising solution for high-precision identification of groundwater contamination source parameters.

Keywords: groundwater contamination source identification; Bayesian; Markov Chain Monte Carlo; surrogate model; multi-layer perceptron

Citation: An, Y.; Zhang, Y.; Yan, X. An Integrated Bayesian and Machine Learning Approach Application to Identification of Groundwater Contamination Source Parameters. *Water* **2022**, *14*, 2447. https://doi.org/10.3390/w14152447

Academic Editor: Dimitrios E. Alexakis

Received: 8 July 2022
Accepted: 5 August 2022
Published: 7 August 2022

1. Introduction

Groundwater contamination has become a global issue due to industrial wastewater discharge, agricultural fertilizer utilization, landfill leakage, and so on, which damages the groundwater environment and drinking water safety [1–4]. The spatiotemporal distribution characteristics of groundwater contamination sources are an important basis for formulating remediation plans and risk assessments of groundwater contamination [5–7]. However, groundwater contamination is characterized by concealment and hysteresis, so it is difficult to obtain groundwater contamination source parameters (GCSPs) directly, such as the location and release history of the groundwater contamination source [8–10].

Consequently, it is urgent to explore an effective approach for high-precision identification of GCSPs.

Since it is almost impossible to measure GCSPs directly, many existing identification approaches infer GCSPs from the easily obtained measurements of contamination concentration and hydraulic head, which results in an inverse problem. At present, the inverse problem can be solved through the following approaches [8]: (i) the simulation-optimization approach [11–13]; (ii) the probabilistic and geostatistical simulation approach [14–16]; (iii) the analytical solution and regression approach [17–19]; and (iv) the direct approach [8,18,20]. The simulation-optimization approach has been applied extensively in previous studies, but it is limited by an inadequate description of the uncertainty of solutions [21]. Fortunately, a promising approach, that is, Bayesian inversion, can effectively solve this uncertainty problem [22–24]. The generalized likelihood uncertainty estimation (GLUE) [25] is a Bayesian approach first applied to the hydrology field, and it has been applied to the identification of GCSPs [26,27]. However, the GLUE approach is only suitable for low-dimensional systems due to its inefficient sampling technique [28,29]. Markov Chain Monte Carlo (MCMC) [30,31] uses a more efficient sampling technique than that of GLUE. To date, MCMC has been successfully used for the identification of GCSPs in hydrology and groundwater fields [7,16,23,24,32].

To date, the MCMC approach has developed many sampling algorithms, including the Metropolis–Hastings (MH) algorithm [30], the adaptive Metropolis (AM) algorithm [33,34], the delayed rejection adaptive Metropolis (DRAM) algorithm [35], the Differential Evolution Adaptive Metropolis (DREAM) algorithm [36], and Differential Evolution Adaptive Metropolis with Discrete Sampling ($DREAM_{(D)}$) [37]. For GCSP identification, the unknown parameters are both continuous (such as contamination source intensity) and discrete (such as contamination source location). However, many studies assume the contamination source location as a continuous variable [23,24,38,39]. To identify GCSPs more accurately and effectively, the $DREAM_{(D)}$-MCMC approach, which can consider both discrete and continuous variables, is used for GCSP identification in this study.

Moreover, the observation well locations (OWLs) largely affect the identification accuracy of GCSPs [40,41]. However, it has usually been assumed that OWLs are known prior to Bayesian inversion of GCSPs in previous studies [7,39,42]. Generally, the identification of GCSPs will be more effective and accurate when the measurements of OWLs have more valuable information related to unknown GCSPs [16,23]. Moreover, the Bayesian design is applicable to high-dimensional nonlinear systems such as the groundwater contaminant transport model [19,43]. Therefore, information entropy (IE), which is a quantitative measurement index of information, is employed to scientifically and effectively determine optimal OWLs through Bayesian design prior to the high-precision identification of GCSPs in this study.

Furthermore, the processes of the optimal design of OWLs and high-precision identification of GCSPs need to invoke the original simulation model thousands of times, which not only requires an enormous amount of time but also results in a huge computational load. Establishing a surrogate model for the original simulation model has been regarded as a promising approach to overcome this challenge [32,44–46]. Multi-layer Perceptron (MLP) introduces the hidden layer on the basis of a single-layer neural network, which can be effectively applied to the regression for high-dimensional nonlinear systems [47]. Hence, a surrogate model is constructed to describe the relationship between the input and output of the high-dimensional nonlinear simulation model by using the MLP approach in this study. To significantly reduce the computational load and time requirement of optimal design of OWLs and high-precision identification of GCSPs, the surrogate model is directly invoked in the whole subsequent iteration process after constructing the surrogate model.

The main objectives of this study are the optimal design of OWLs and the high-precision identification of GCSPs. An innovative approach that integrates the Bayesian design, IE, MCMC, and surrogate modeling was developed for the above objectives. The overall research process is shown in Figure 1. Section 2 mainly explains the proposed

theoretical framework of this study, including the theory and research approaches. Section 3 mainly examines the performance of the proposed approaches using two hypothetical numerical case studies. Section 4 shows the results and discusses them. Section 5 presents the main conclusions of this study.

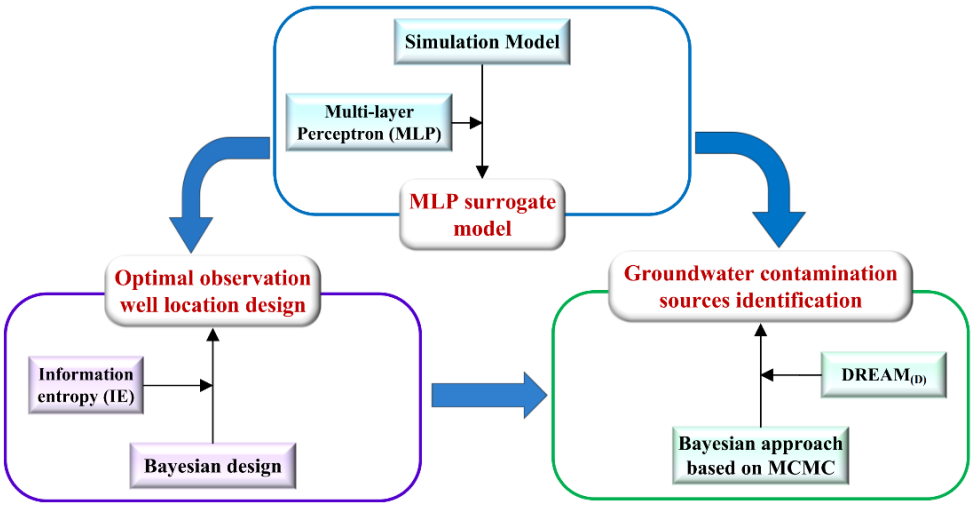

Figure 1. Flow diagram of research process.

2. Theoretical Framework

2.1. Simulation Model

A numerical model of groundwater contamination can describe the real processes of advection, dispersion, and biochemical reactions of groundwater contaminants so as to research the excitation response relationship between the input and output of an actual groundwater system. This study assumes that there is only one conservative contaminant, so only advection and dispersion in groundwater need to be considered. A two-dimensional homogeneous (heterogeneous) steady-flow partial differential equation is described as follows:

$$\frac{\partial}{\partial x_i}\left(K_i \frac{\partial h}{\partial x_i}\right) = 0 \tag{1}$$

A two-dimensional advection and dispersion groundwater contaminant transport equation is described as follows:

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial c}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(cu_i) + \frac{qc_s}{nb} \tag{2}$$

where x_i and x_j represent the Cartesian coordinates for $i, j = 1, 2$ (L); K_i symbolizes a principal component of the hydraulic conductivity tensor (LT^{-1}); h represents the hydraulic head (L); c is the concentration of a contaminant dissolved in groundwater (ML^{-3}); t is time (T); q is the volumetric flow rate per unit area of the aquifer representing fluid sources (positive) (LT^{-1}); c_s symbolizes the concentration of the source or sink (ML^{-3}); n denotes the porosity of the porous medium; b symbolizes the aquifer thickness (L); D_{ij} is the hydrodynamic dispersion tensor (L^2T^{-1}); and u_i represents the actual flow velocity (LT^{-1}). D_{ij} and u_i can be written as:

$$\begin{cases} D_{xx} = \alpha_L \frac{u_x^2}{|u|} + \alpha_T \frac{u_y^2}{|u|} \\ D_{yy} = \alpha_L \frac{u_y^2}{|u|} + \alpha_T \frac{u_y^2}{|u|} \\ D_{xy} = D_{yx} = (\alpha_L - \alpha_T)\frac{u_x u_y}{|u|} \end{cases} \tag{3}$$

$$u_i = -\frac{K_i}{n}\frac{\partial h}{\partial x_i} \tag{4}$$

where α_L and α_T represent the longitudinal and transversal dispersivities (L), respectively; u_x and u_y are the components of the actual flow velocity (LT^{-1}); and $|u|$ denotes the modulus of u, such that $|u| = \sqrt{u_x^2 + u_y^2}$.

The partial differential equation ((Equations (1)–(4)) is combined with initial conditions and boundary conditions to form the groundwater flow model and the contaminant transport model. In this study, MODFLOW [48] and MT3DMS [49] are employed to solve the groundwater flow and contaminant transport equations, respectively.

2.2. Optimal Observation Well Location Design

According to Bayes' theorem, the unknown GCSPs ϕ can be estimated by concentration measurements m under an observation condition ϖ, and the expression [43] is as follows:

$$p(\phi|m,\ \varpi) = \frac{p(\phi|\varpi)p(m|\phi,\ \varpi)}{p(m|\varpi)} \tag{5}$$

where $p(\phi|m,\ \varpi)$ is the posterior distribution,$p(\phi|\varpi)$ symbolizes the prior distribution, $p(m|\phi,\ \varpi)$ is the likelihood, $p(m|\varpi)$ is usually regarded as a normalization constant, and $p(m|\varpi) = \int p(m|\phi,\ \varpi)p(\phi|\varpi)\mathrm{d}\phi$. In this study, ϖ symbolizes OWLs corresponding to the concentration measurements m. The prior distribution $p(\phi|\varpi)$ represents the knowledge of the unknown parameters prior to obtaining the measurements, which is independent of the OWLs, so $p(\phi|\varpi) = p(\phi)$.

The unknown GCSPs ϕ are identified based on the concentration measurements m. The IE of the posterior distribution $p(\phi|m,\ \varpi)$ can be expressed as follows [50]:

$$H(\varpi) = -\int p(\phi|m,\ \varpi)\ln p(\phi|m,\ \varpi)d\phi \tag{6}$$

In this study, the expected IE is used as the utility function. The expected IE of an observation condition ϖ can be expressed as follows [50]:

$$\begin{aligned} E(\varpi) &= -\int[\int p(\phi|m,\ \varpi)\ln p(\phi|m,\ \varpi)d\theta]p(\phi|m,\ \varpi)dm \\ &= -\int\int p(\phi|m,\ \varpi)p(m|\varpi)\ln p(\phi|m,\ \varpi)d\phi dm \end{aligned} \tag{7}$$

where $E(\varpi)$ denotes the expected IE.

The analytic solutions of expected IE in Equation (7) are nonexistent, and expected IE can be approximately numerically solved using the approach [43] by Huan and Marzouk. Then, optimal OWLs can be obtained by solving the expected IE.

2.3. Parameter Identification

2.3.1. Bayesian Inversion

After determining the optimal OWLs, the corresponding measurements can be obtained. Subsequently, Bayesian inversion is used for the identification of GCSPs in this study. The posterior distribution $p(\phi|m)$ can be calculated as:

$$p(\phi|m) = \frac{p(\phi)p(m|\phi)}{p(m)} \tag{8}$$

where $p(m|\phi)$ denotes the likelihood equation, and $p(m)$ is usually regarded as a normalization constant.

This study assumes that the measurement errors conform to the normal distribution with a mean of 0 and a covariance of R, so the likelihood equation can be expressed as follows:

$$p(m|\phi) = \frac{1}{(2\pi)^{(n/2)}|R|^{1/2}} \exp\left\{-\frac{1}{2}[m - F(\phi)]^T R^{-1}[m - F(\phi)]\right\} \quad (9)$$

where n represents the number of measurements, $|R|$ means the determinant of the covariance matrix R, R^{-1} symbolizes the inverse of R, $F(\cdot)$ denotes the simulation model, and T denotes transposition.

The analytical form of the posterior parameter distribution is usually difficult to obtain directly for a groundwater simulation model because it is a highly nonlinear system. MCMC is used to calculate and analyze the statistical characteristics of the posterior parameter distribution $p(\phi|m)$ in this study.

2.3.2. MCMC

Markov Chain Monte Carlo (MCMC) can randomly search and sample the posterior parameter space. In the sampling process, MCMC constructs a suitable Markov chain, which can reach a stable distribution $\pi(\phi)$, that is, the posterior probability distribution of the parameters, after running for a long enough time. Then, a sampling approach is used to draw samples from the posterior probability distribution $p(\phi|m)$, and these samples are used to analyze the statistical characteristics of $p(\phi|m)$.

In this study, Differential Evolution Adaptive Metropolis with Discrete Sampling ($DREAM_{(D)}$), which can consider both discrete and continuous variables, is used to identify GCSPs. The $DREAM_{(D)}$ approach is not described in detail here; interested readers are referred to Vrugt and Ter Braak [37].

2.4. Multi-Layer Perceptron

Multi-layer Perceptron (MLP) is a supervised machine learning approach that learns a function $F(\cdot)$ by training on a dataset. Given a set of features $\Theta = [\phi_1, \phi_2, \cdots, \phi_k]$ and a target y, it can be used to learn a nonlinear function approximator for regression, where k is the number of dimensions for input. MLP contains one or more nonlinear hidden layers between the input and the output layer, which is different from logistic regression [47], as shown in Figure 2.

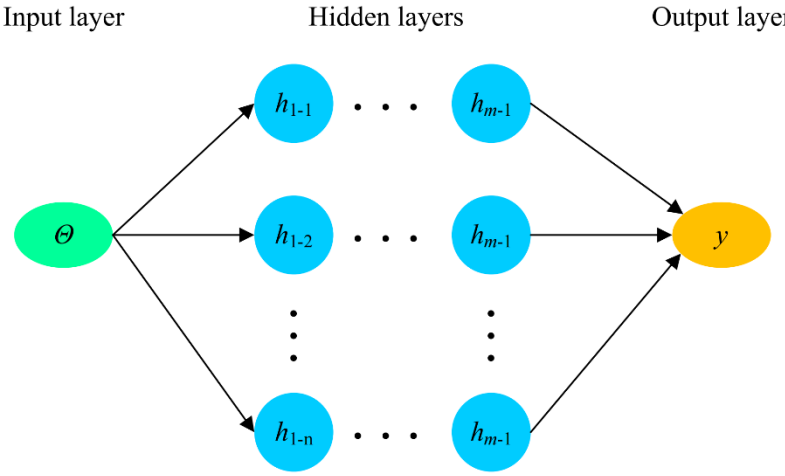

Figure 2. MLP approach structure diagram.

A neural network with multiple hidden layers has a stronger learning ability for high-dimensional nonlinear systems, such as the groundwater contaminant transport

model, than that with only one hidden layer. The general regression expression of the MLP approach [51,52] is:

$$y = \underbrace{f \cdots f}_{m}\left(W\Theta + b\right) \tag{10}$$

where Θ and y denote the inputs and outputs of the regression relationship, respectively; W and b the weights and bias, respectively; m is the number of hidden layers, and $m = 3$ in this study; f is the activation function, including Sigmoid, Tanh, and Relu. The activation function used in this study is Tanh.

More details related to MLP are available in the study by Noriega [53].

3. Numerical Applications

The above theoretical framework was applied to two hypothetical case studies, including a homogeneous case and a heterogeneous case. The advantage of using hypothetical case studies is that the performance of the proposed approaches can be verified clearly and accurately, which is because the calculated results can be directly compared with the theoretical results [54].

3.1. Case Studies

3.1.1. Case 1

The range of the groundwater flow field is 240 L $\times$ 180 L in the first case study, and the discretization graph is shown in Figure 3. The upper and lower boundaries are impervious, and the left and right boundaries are linearly varying heads, as shown in Figure 3. The hydraulic conductivity, porosity, and dispersivities are homogeneous and known, as shown in Table 1. It is assumed that the unknown GCSPs of Case 1 are the intensity (S), release duration (D), and the location of the groundwater contamination source (X and Y). The unknown GCSPs $\phi = [S, D, X, Y]$ are assumed to obey a uniform distribution. The prior ranges and true values of the unknown GCSPs are presented in Table 2. Furthermore, we assume that there is only one contamination source, and the contamination source releases conservative contaminants from the beginning of the simulation.

Table 1. Parameters values of simulation model for Case 1.

Parameters	Values	Unit
Hydraulic conductivity, K	18.00	LT^{-1}
Porosity, n	0.30	-
Longitudinal dispersivity, α_L	12.00	L
Transverse dispersivity, α_T	3.60	L

Table 2. The prior ranges and true values of unknown GCSPs for Case 1.

Parameters	True Values	Prior Ranges	Unit
S	3600	[2000, 5000]	MT^{-1}
D	480	[450, 550]	T
X	11	[10, 18]	L
Y	5	[4, 9]	L

In this case study, three observation wells need to be designed, but the OWLs are unknown. The concentration measurements of all potential observation wells are obtained at t = 660 T, 720 T, 780 T, 840 T and 900 T . It is assumed that the measurement error ε conforms to an independent normal distribution with a mean of 0 and a variance of 0.05.

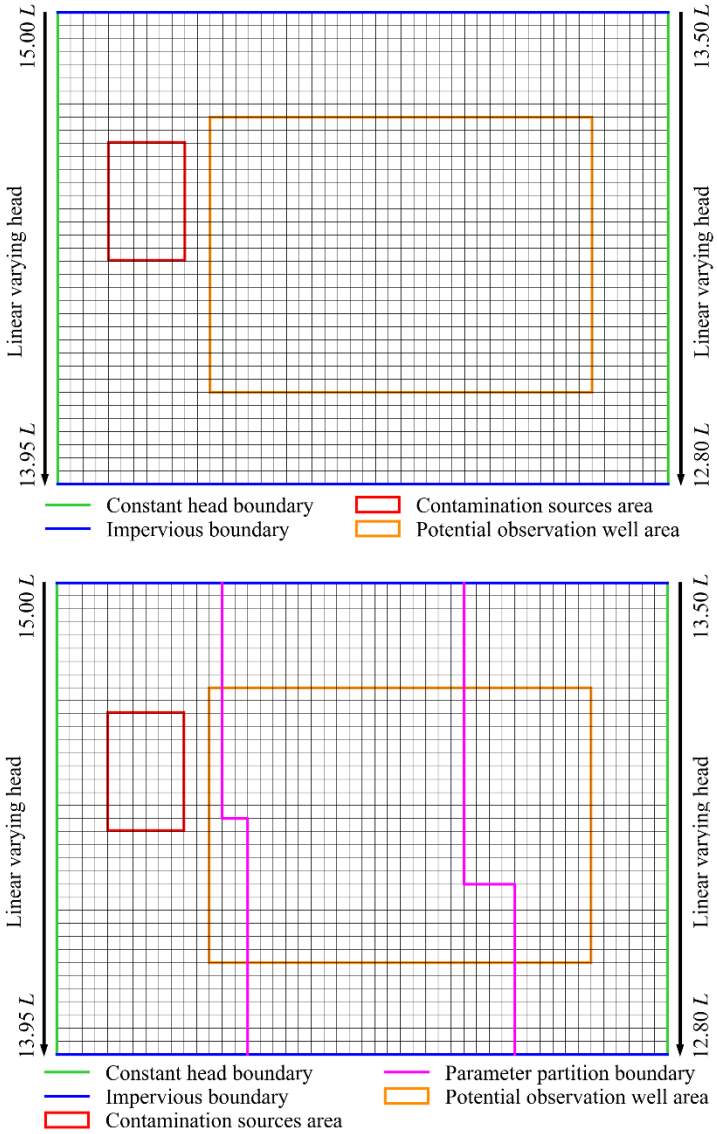

Figure 3. Discretization graph of groundwater flow field for Case 1 (**upper**) and Case 2 (**lower**).

3.1.2. Case 2

The groundwater flow field range of the second case study is 240 L × 180 L. The boundary conditions, porosity (n), dispersivities (α_L and α_T), unknown parameters, observation well conditions, etc., are the same as in Case 1. However, the conductivity field has three zones with hydraulic conductivity K of 6, 15, and 22 (LT^{-1}) from left to right, respectively, as shown in Figure 3.

3.2. Application of the Surrogate Model

The MLP approach is used to construct a surrogate model for the original simulation model to reduce the computational load of the optimal design of OWLs and the identification of GCSPs. The first step of constructing the MPL surrogate model is to draw the necessary parameter samples from the prior range of unknown GCSPs to obtain training samples. The obtained parameter samples are used as the input data for the original simulation model, and then the original simulation model is run to obtain the corresponding concentration outputs. Next, the MLP surrogate model is constructed based on the input–output datasets. In this study, the MLP surrogate model is constructed for the original simulation model according to 1000 sets of input–output datasets, and 200 sets of new parameter samples are randomly selected from the prior range of unknown GCSPs as test

samples to test the accuracy of the newly constructed MLP surrogate model. The 200 sets of parameter samples are used as the input of both the original simulation model and the newly constructed MLP surrogate model. Then, 200 sets of corresponding concentration outputs of the original simulation model are compared with those of the MLP surrogate model to analyze the accuracy of the constructed MLP surrogate model. We compared all of the outputs of the original simulation model with those of the constructed MLP surrogate model one by one in this study to test the accuracy of the constructed MLP surrogate model more accurately. Relative error (*RE*) is used to quantify the accuracy of the MLP surrogate model:

$$RE = \left| \frac{\text{output of original simulation model} - \text{output of surrogate model}}{\text{output of original simulation model}} \right| \times 100\% \quad (11)$$

3.3. Optimal Observation Well Location Design for Case Studies

According to the constructed MLP surrogate model and Equations (5)–(7), the values of $E(\varpi)$ are obtained for all of the potential OWLs. According to the meaning of IE, the observation well design scheme with the minimum value of $E(\varpi)$ is regarded as the optimal observation well design scheme. Three optimal OWLs were obtained to suit the needs of the two case studies.

3.4. Computational Time Analysis

One of the main objectives of this study is to improve the efficiency of the iteration process for the optimal design of OWLs and the identification of GCSPs. Therefore, the execution time of the original simulation model and the MLP surrogate model was measured and compared throughout the computational process. Furthermore, the purpose of simultaneously applying the proposed approach to Cases 1 and 2 is to test whether the approach is suitable for both homogeneous and inhomogeneous media.

4. Results and Discussion

4.1. Analysis of the Surrogate Model

The performance of the constructed MLP surrogate model of Cases 1 and 2 in this study is illustrated in Figures 4 and 5, respectively. To evaluate the performance of the MLP surrogate model more objectively, the contamination concentration outputs of the MLP surrogate model and the original simulation model and the average relative errors between them are displayed randomly, which are obtained by any one parameter sample of the 200-set test samples at each moment (t = 660 T, 720 T, 780 T, 840 T and 900 T). Figures 4 and 5 show that the contamination concentration outputs of the MLP surrogate model are very close to those of the original simulation model, and the relative error value is very small, which indicates that the MLP surrogate model achieves high accuracy and is very close to the original simulation model. The results also show that the MLP approach is suitable for constructing a surrogate model for both the homogeneous and heterogeneous groundwater simulation models. Therefore, the MLP surrogate model can be directly used for the optimal design of OWLs and the identification of GCSPs in the two case studies, which not only enhances the overall computing efficiency but also maintains high precision.

However, the surrogate model has some limitations, such as the inability to analyze the precise mechanism, which is because it is a black box model. Generally, the surrogate model is applied to the situation in which the simulation model must be invoked thousands of times, such as simulation optimization and uncertainty analyses.

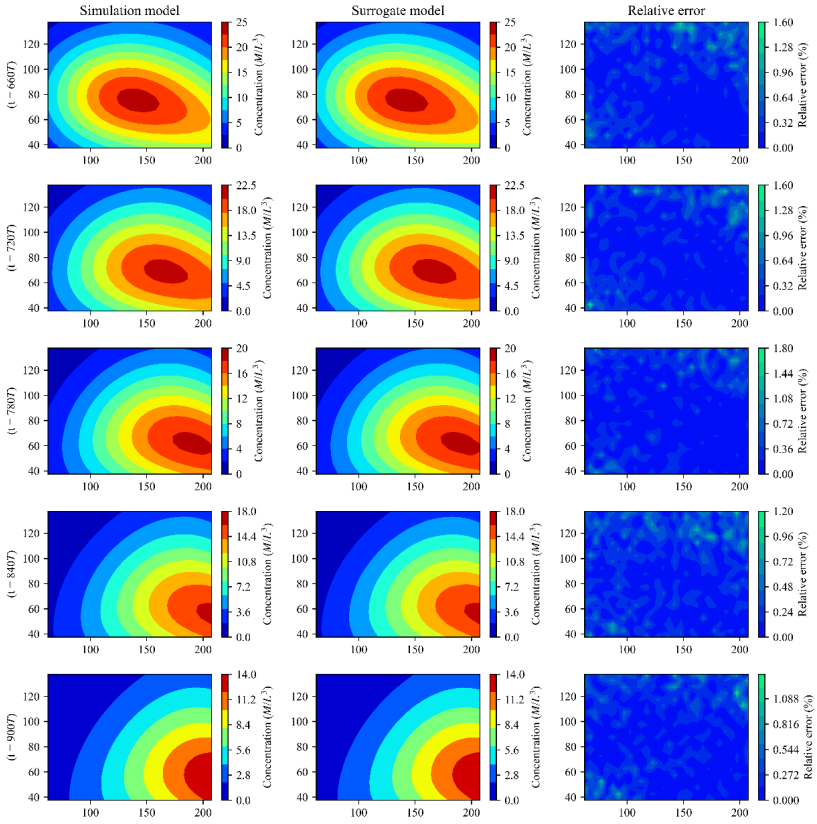

Figure 4. Outputs and relative error of simulation model and surrogate model for Case 1 (the x-axis and y-axis represent the length and width of the flow field, respectively).

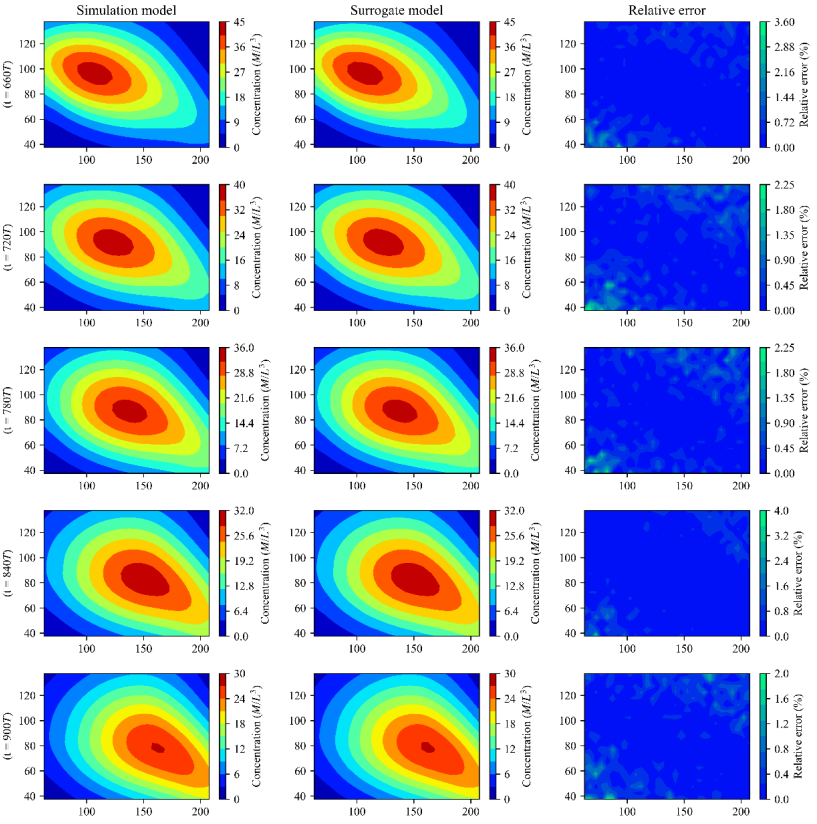

Figure 5. Output and relative error of simulation model and surrogate model for Case 2 (the x-axis and y-axis represent the length and width of the flow field, respectively).

4.2. Analysis of the Optimal Observation Well Locations

With the values of $E(\varpi)$ obtained for two case studies, the optimal OWLs were determined, and the three optimal OWLs are shown in Figure 6.

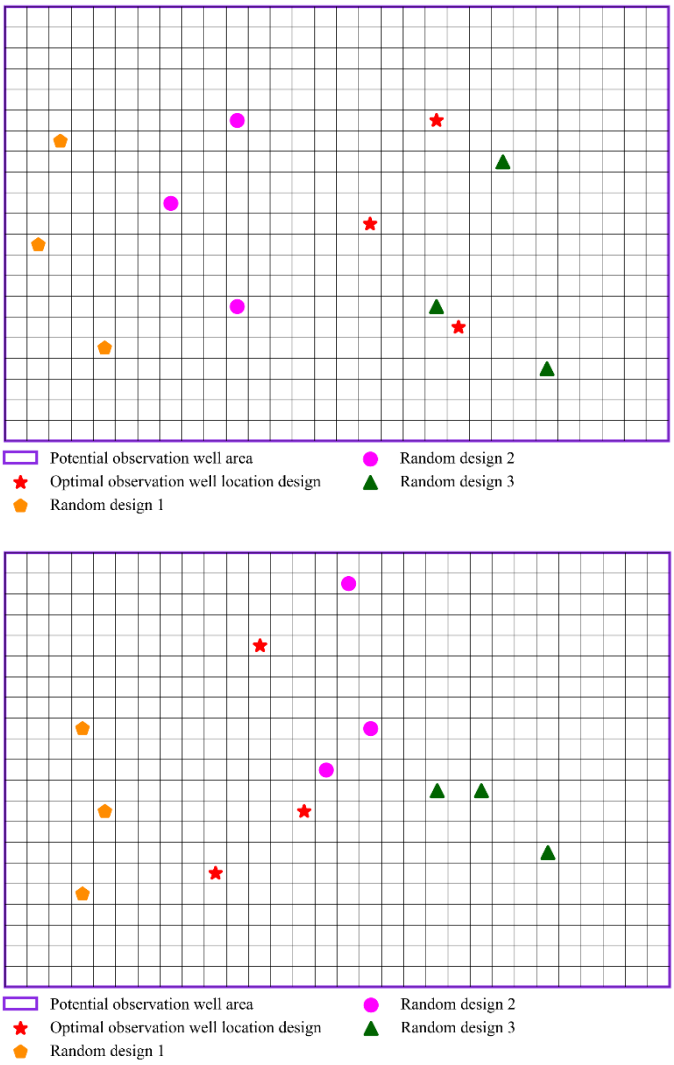

Figure 6. The OWLs of the optimal design and 3 random designs for Case 1 (**upper**) and Case 2 (**lower**).

Most of the optimal observation wells determined by Bayesian design and IE are located along the flow direction of groundwater. This shows that more valuable information about unknown GCSPs can be obtained by sampling along the flow direction.

4.3. Analysis of the Parameter Identification Results

We compared the identification accuracy of GCSPs of the optimal design with that of three other random designs to verify the feasibility and effectiveness of the optimal design of OWLs proposed in Section 4.2. The OWLs of the optimal design and the other three random designs are shown in Figure 6. The corresponding trace plots of the MCMC simulation for the two case studies are shown in Figures 7 and 8.

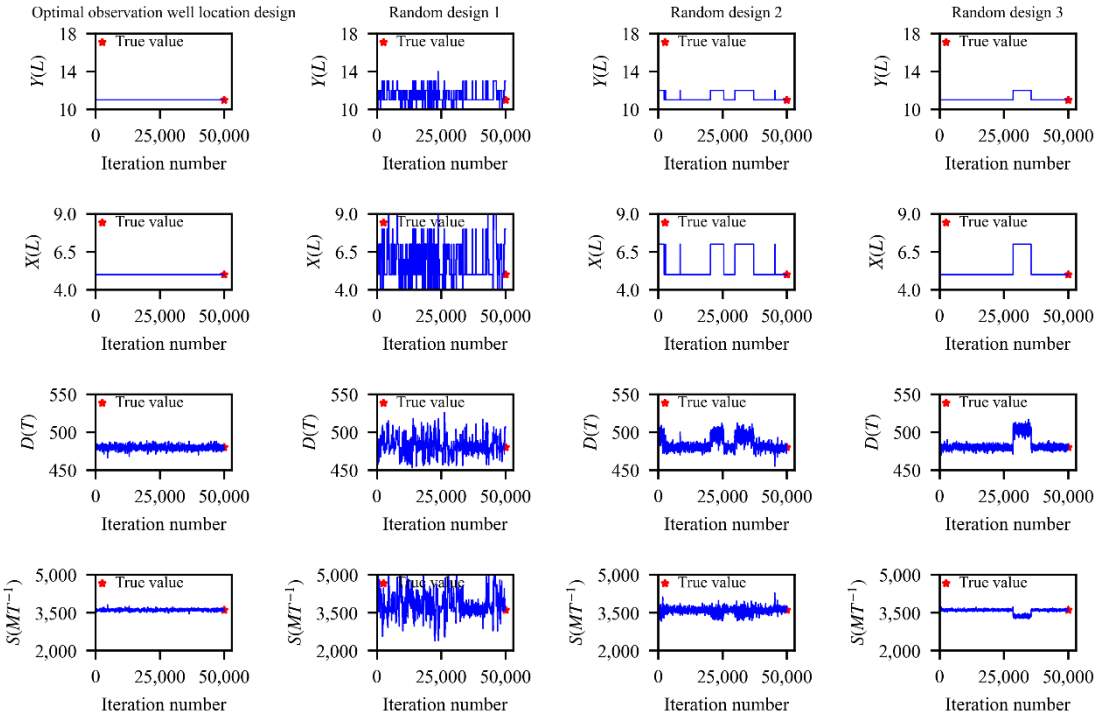

Figure 7. Trace plots of the MCMC simulation for Case 1.

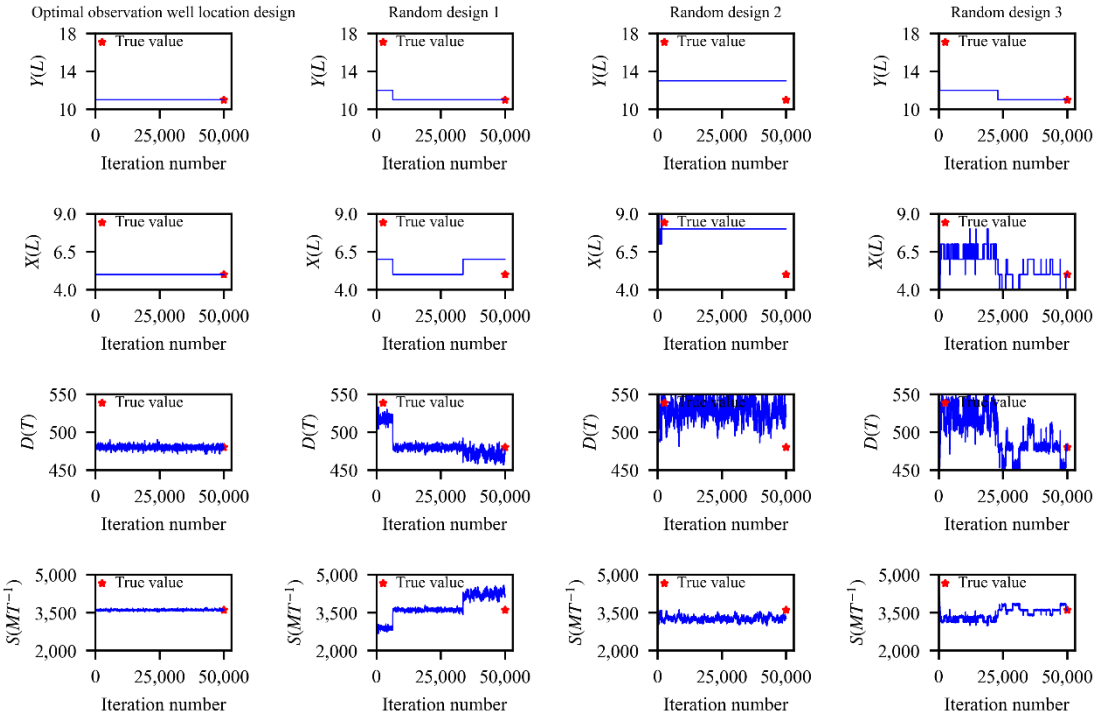

Figure 8. Trace plots of the MCMC simulation for Case 2.

Figures 7 and 8 show that compared with the other three random schemes, the trace plots corresponding to the optimal design of OWLs can converge to near the true value faster.

To further ensure the stable convergence of the Markov chains, the Gelman–Rubin approach was used for convergence diagnosis, which is a variance ratio approach proposed by Gelman and Rubin [55]. The index of convergence diagnosis can be expressed as:

$$R = \sqrt{\frac{g-1}{g} + \frac{q+1}{q} \cdot \frac{B}{W}} \tag{12}$$

where R denotes the diagnostic index, which is the scale reduction score; g denotes the length of Markov chain in the DREAM$_{(D)}$ algorithm; q is the number of Markov chains; B represents the variance of the average value of the q Markov chains; W denotes the average value of the intrachain variance of the q Markov chains. Generally, if the value of R is less than 1.2, it is considered that the Markov chain has attained a stable convergence state; that is, the sampling process of the algorithm has converged.

The last 20,000 sets of samples in the stable convergence stage were used to estimate the statistical characteristics of the posterior parameter distribution. The comparison results of the posterior probability distributions of GCSPs obtained by the optimal design of OWLs and the other three random designs are shown in Figure 9. It should be noted that only the delayed probability distributions of parameters D and S are shown in Figure 9. This is because parameters D, X, and Y are treated as discrete variables in the identification process of GCSPs in this study; only the parameter S is treated as a continuous variable, while X and Y are sensitive to the obtained contamination concentration measurements, and their posterior probability distribution is one or several vertical lines, so their posterior probability distribution is not displayed in this study.

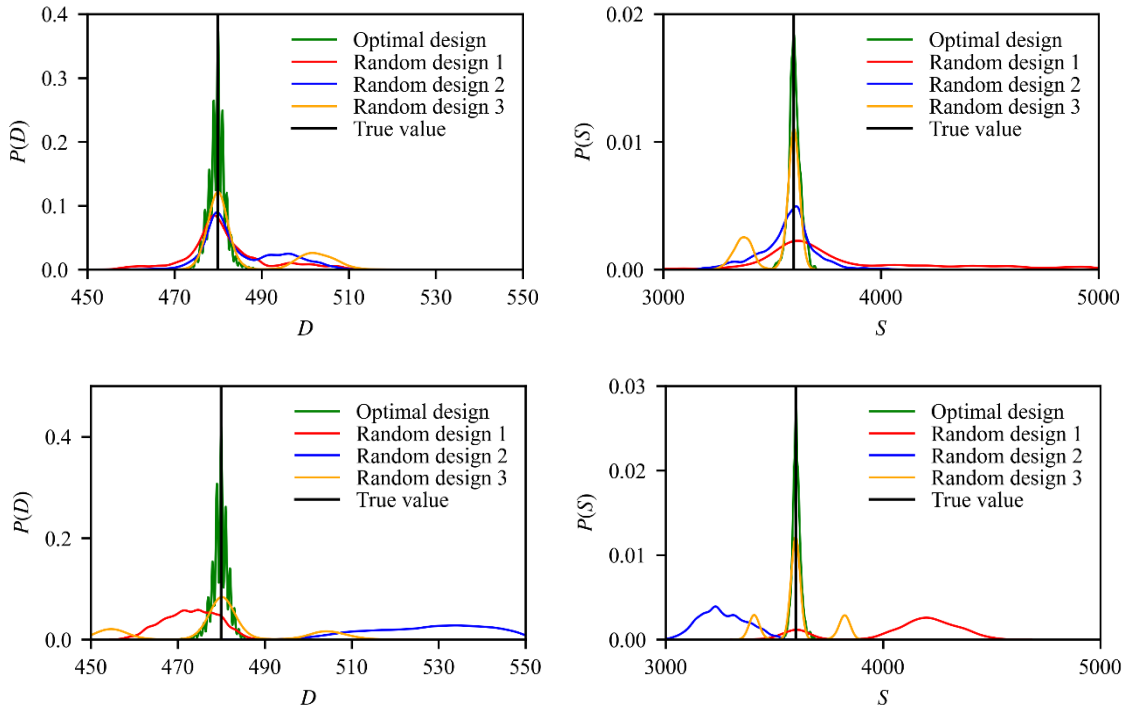

Figure 9. Comparison results of the posterior probability distributions for the optimal design and 3 other random designs for Case 1 (**Upper**) and Case 2 (**Lower**).

Figure 9 shows that compared with the other three random designs, the optimal design of OWLs obtains the maximum a posteriori probability (MAP) and mean values of parameters D and S that are closer to their true value. The uncertainties of posterior parameter distributions obtained by the optimal design of OWLs are also lower than those obtained by the other three random designs. This shows that the optimal design of OWLs

can obtain more accurate identification results. Therefore, the optimal design of OWLs and the approaches for GCSP identification are feasible and efficient.

It is noted that although the MLP surrogate model and the simulation model have a highly similar input–output relationship, the error between them still exists. The error still has a slight impact on identification results. We will study this aspect in detail in future research.

Moreover, the constructed MLP surrogate model effectively reduces the calculation time in the processes of the optimal design of OWLs and the identification of GCSPs. Completing the same 50,000 simulations only takes 175 min when using the MLP surrogate model, while it takes 1400 min to use the simulation model. These operations were run on a PC platform with an AMD R5-3600X 3.80 GHz processor and 16 GB RAM. The MLP surrogate model is characterized by short computation time and high accuracy, which promotes its application in solving the identification problem of GCSPs.

5. Conclusions

This study developed an integrated Bayesian inversion and machine learning approach that combines MCMC, IE, MLP, and surrogate modeling for the optimal design of OWLs and the identification of GCSPs. The expected IE is used to quantify valuable information related to GCSPs carried by concentration measurements by linking it with Bayesian design so as to determine the optimal OWLs. After determining the optimal OWLs and obtaining the corresponding contamination concentration measurements, the posterior distributions of the unknown GCSPs are obtained by using the DREAM$_{(D)}$-MCMC approach.

The MLP, which is a type of machine learning approach, is applied to construct a surrogate model for the original simulation model, which is directly invoked for the optimal design of OWLs and the high-precision identification of GCSPs. The constructed MLP surrogate model can significantly reduce the computing load and computing time, which shows that the processes of design and identification can be greatly accelerated by integrating surrogate modeling and machine learning. Overall, the accuracy and effectiveness of the proposed approach were verified through case studies of homogeneous and heterogeneous media in this study. In summary, this study highlights that the integrated Bayesian and machine learning approach provides a promising solution for the high-precision identification of GCSPs. In future research, a numerical case of groundwater contamination source identification that is closer to the real situation will be established to further illustrate the feasibility and effectiveness of the approach proposed in this paper.

Author Contributions: Conceptualization and methodology, Y.Z.; validation and analysis, Y.A. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (No. 42102287), the China Postdoctoral Science Foundation (No. 2020M683399), and the Key Research and Development Program of Shaanxi (Program No. 2021ZDLSF05-01).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are included within the article.

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

References

1. Egbueri, J.C.; Agbasi, J.C. Combining data-intelligent algorithms for the assessment and predictive modeling of groundwater resources quality in parts of southeastern Nigeria. *Environ. Sci. Pollut. Res.* **2022**, 1–25. [CrossRef]
2. Egbueri, J.C.; Unigwe, C.O.; Omeka, M.E.; Ayejoto, D.A. Urban groundwater quality assessment using pollution indicators and multivariate statistical tools: A case study in southeast Nigeria. *Int. J. Environ. Anal. Chem.* **2021**, *4*, 1–27. [CrossRef]
3. Egbueri, J.C. Groundwater quality assessment using pollution index of groundwater (PIG), ecological risk index (ERI) and hierarchical cluster analysis (HCA): A case study. *Groundw. Sustain. Dev.* **2019**, *10*, 100292. [CrossRef]

4. Omeka, M.E.; Egbueri, J.C. Hydrogeochemical assessment and health-related risks due to toxic element ingestion and dermal contact within the nnewi-awka urban areas, Nigeria. *Environ. Geochem. Health* **2022**, 1–29. [CrossRef]
5. Ayvaz, M.T. A hybrid simulation–optimization approach for solving the areal groundwater pollution source identification problems. *J. Hydrol.* **2016**, *538*, 161–176. [CrossRef]
6. Datta, B. Optimal unknown pollution source characterization in a contaminated groundwater aquifer—Evaluation of a developed dedicated software tool. *J. Geosci. Environ. Prot.* **2014**, *2*, 41. [CrossRef]
7. Zeng, L.; Shi, L.; Zhang, D.; Wu, L. A sparse grid based Bayesian method for contaminant source identification. *Adv. Water Resour.* **2012**, *37*, 1–9. [CrossRef]
8. Atmadja, J.; Bagtzoglou, A.C. State of the art report on mathematical methods for groundwater pollution source identification. *Environ. Forensics* **2001**, *2*, 205–214. [CrossRef]
9. Datta, B.; Chakrabarty, D.; Dhar, A. Identification of unknown groundwater pollution sources using classical optimization with linked simulation. *J. Hydro-Environ. Res.* **2011**, *5*, 25–36. [CrossRef]
10. Sun, A.Y.; Painter, S.L.; Wittmeyer, G.W. A constrained robust least squares approach for contaminant release history identification. *Water Resour. Res.* **2006**, *42*, 1–13. [CrossRef]
11. Amirabdollahian, M.; Datta, B. Identification of pollutant source characteristics under uncertainty in contaminated water resources systems using adaptive simulated anealing and fuzzy logic. *Int. J. GEOMATE* **2014**, *6*, 757–762. [CrossRef]
12. Huang, L.; Wang, L.; Zhang, Y.; Xing, L.; Hao, Q.; Xiao, Y.; Yang, L.; Zhu, H. Identification of groundwater pollution sources by a SCE-UA algorithm-based simulation/optimization model. *Water* **2018**, *10*, 193. [CrossRef]
13. Jha, M.K.; Datta, B. Linked simulation-optimization based dedicated monitoring network design for unknown pollutant source identification using dynamic time warping distance. *Water Resour. Manag.* **2014**, *28*, 4161–4182. [CrossRef]
14. Butera, I.; Tanda, M.G. A geostatistical approach to recover the release history of groundwater pollutants. *Water Resour. Res.* **2003**, *39*, WR002314. [CrossRef]
15. Gzyl, G.; Zanini, A.; Frączek, R.; Kura, K. Contaminant source and release history identification in groundwater: A multi-step approach. *J. Contam. Hydrol.* **2014**, *157*, 59–72. [CrossRef]
16. Yan, X.; Dong, W.; An, Y.; Lu, W. A Bayesian-based integrated approach for identifying groundwater contamination sources. *J. Hydrol.* **2019**, *579*, 124160. [CrossRef]
17. Alapati, S.; Kabala, Z.J. Recovering the release history of a groundwater contaminant using a non-linear least-squares method. *Hydrol. Process.* **2000**, *14*, 1003–1016. [CrossRef]
18. Bagtzoglou, A.C.; Atmadja, J. Marching-jury backward beam equation and quasi-reversibility methods for hydrologic inversion: Application to contaminant plume spatial distribution recovery. *Water Resour. Res.* **2003**, *39*, 1–14. [CrossRef]
19. Woodbury, A.D.; Ulrych, T.J. Minimum relative entropy inversion: Theory and application to recovering the release history of a groundwater contaminant. *Water Resour. Res.* **1996**, *32*, 2671–2681. [CrossRef]
20. Neupauer, R.M.; Borchers, B.; Wilson, J.L. A Comparison of Two Methods for Recovering the Release History of a Groundwater Contamination Source. *Water Resour. Res.* **2000**, *36*, 2469–2475. [CrossRef]
21. Skaggs, T.H.; Kabala, Z.J. Recovering the release history of a groundwater contaminant. *Water Resour. Res.* **1994**, *30*, 71–79. [CrossRef]
22. Ma, X.; Zabaras, N. An efficient Bayesian inference approach to inverse problems based on an adaptive sparse grid collocation method. *Inverse Probl.* **2009**, *25*, 35013–35027. [CrossRef]
23. Zhang, J.; Zeng, L.; Chen, C.; Chen, D.; Wu, L. Efficient Bayesian experimental design for contaminant source identification. *Water Resour. Res.* **2015**, *51*, 576–598. [CrossRef]
24. 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]
25. Beven, K.; Binley, A. The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.* **1992**, *6*, 279–298. [CrossRef]
26. Morse, B.S.; Pohll, G.; Huntington, J.; Rodriguez, R. Stochastic capture zone analysis of an arsenic-contaminated well using the generalized likelihood uncertainty estimator (GLUE) methodology. *Water Resour. Res.* **2003**, *39*, 1–9. [CrossRef]
27. Rojas, R.; Feyen, L.; Dassargues, A. Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resour. Res.* **2008**, *44*, 1–16. [CrossRef]
28. Blasone, R.S.; Vrugt, J.A.; Madsen, H.; Rosbjerg, D.; Robinson, B.A.; Zyvoloski, G.A. Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. *Adv. Water Resour.* **2008**, *31*, 630–648. [CrossRef]
29. Montanari, A. Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. *Water Resour. Res.* **2005**, *41*, 1–13. [CrossRef]
30. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97–109. [CrossRef]
31. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [CrossRef]
32. An, Y.; Lu, W.; Cheng, W. Surrogate model application to the identification of optimal groundwater exploitation scheme based on regression kriging method—A case study of Western Jilin Province. *Int. J. Environ. Res. Public Health* **2015**, *12*, 8897–8918. [CrossRef]

33. Haario, H.; Saksman, E.; Tamminen, J. Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Stat.* **1999**, *14*, 375–395. [CrossRef]
34. Haario, H.; Saksman, E.; Tamminen, J. An adaptive Metropolis algorithm. *Bernoulli* **2001**, *7*, 223–242. [CrossRef]
35. Haario, H.; Laine, M.; Mira, A.; Saksman, E. DRAM: Efficient adaptive MCMC. *Stat. Comput.* **2006**, *16*, 339–354. [CrossRef]
36. Vrugt, J.A.; Ter Braak, C.J.F.; Diks, C.G.H.; Robinson, B.A.; Hyman, J.M.; Higdon, D. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Int. J. Nonlinear Sci. Numer. Simul.* **2009**, *10*, 273–290. [CrossRef]
37. Vrugt, J.A.; Ter Braak, C.J. DREAM (D): An adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 3701–3713. [CrossRef]
38. Laloy, E.; Rogiers, B.; Vrugt, J.A.; Mallants, D.; Jacques, D. Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion. *Water Resour. Res.* **2013**, *49*, 2664–2682. [CrossRef]
39. Wang, H.; Jin, X. Characterization of groundwater contaminant source using Bayesian method. *Stoch. Environ. Res. Risk Assess.* **2013**, *27*, 867–876. [CrossRef]
40. Datta, B.; Chakrabarty, D.; Dhar, A. Optimal dynamic monitoring network design and identification of unknown groundwater pollution sources. *Water Resour. Manag.* **2009**, *23*, 2031–2049. [CrossRef]
41. Prakash, O.; Datta, B. Sequential optimal monitoring network design and iterative spatial estimation of pollutant concentration for identification of unknown groundwater pollution source locations. *Environ. Monit. Assess.* **2013**, *185*, 5611–5626. [CrossRef] [PubMed]
42. Michalak, A.M.; Kitanidis, P.K. A method for enforcing parameter nonnegativity in Bayesian inverse problems with an application to contaminant source identification. *Water Resour. Res.* **2003**, *39*, 1–14. [CrossRef]
43. Huan, X.; Marzouk, Y.M. Simulation-based optimal Bayesian experimental design for nonlinear systems. *J. Comput. Phys.* **2013**, *232*, 288–317. [CrossRef]
44. An, Y.; Yan, X.; Lu, W.; Qian, H.; Zhang, Z. An improved Bayesian approach linked to a surrogate model for identifying groundwater pollution sources. *Hydrogeol. J.* **2022**, *30*, 601–616. [CrossRef]
45. 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]
46. 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]
47. Ruck, D.W.; Rogers, S.K.; Kabrisky, M.; Maybeck, P.S.; Oxley, M.E. Comparative analysis of backpropagation and the extended Kalman filter for training multilayer perceptrons. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 686–691. [CrossRef]
48. Harbaugh, A.W. *MODFLOW-2005, The U.S. Geological Survey Modular Groundwater Model—The GroundWater Flow Process*; U.S. Geological Survey Techniques and Methods 6-A16; U.S. Geological Survey: Reston, VA, USA, 2005.
49. Zheng, C.; Wang, P.P. *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; U.S. Army Corps of Engineers—Engineer Research and Development Center: Vicskburg, MS, USA, 1999.
50. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
51. Agirre-Basurko, E.; Ibarra-Berastegi, G.; Madariaga, I. Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. *Environ. Model. Softw.* **2006**, *21*, 430–446. [CrossRef]
52. Egbueri, J.C.; Agbasi, J.C. Performances of MLR, RBF-NN, and MLP-NN in the evaluation and prediction of water resources quality for irrigation purposes under two modeling scenarios. *Geocarto Int.* **2022**, 1–28. [CrossRef]
53. Noriega, L. *Multilayer Perceptron Tutorial*; School of Computing, Staffordshire University: Stoke-on-Trent, UK, 2005.
54. Zhao, Y.; Lu, W.; Xiao, C. A Kriging surrogate model coupled in simulation–optimization approach for identifying release history of groundwater sources. *J. Contam. Hydrol.* **2016**, *185*, 51–60. [CrossRef] [PubMed]
55. Gelman, A.; Rubin, D.B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **1992**, *7*, 457–472. [CrossRef]