Simultaneous Identification of Number, Location, and Release History of Groundwater Contamination Sources

Jiuhui Li
Northeast Normal University
Wenxi Lu (luwx999@163.com)
Jilin University
Zhengfang Wu
Northeast Normal University
Hongshi He
Northeast Normal University

Research Article

Keywords: Groundwater contamination, Mixed integer nonlinear programming, Source number, Surrogate model

Posted Date: November 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-956495/v1

Simultaneous identification of number, location, and release history of groundwater contamination sources

Jiuhui Li[1], Wenxi Lu[2], Zhengfang Wu[3], Hongshi He[4]

[1]Ph.D. Key Laboratory of Geographical Processes and Ecological Security in Changbai Mountains, Ministry of Education, School of Geographical Sciences, Northeast Normal University, Changchun, China. E-mail: lijh801@nenu.edu.cn

[2]Professor, Key Laboratory of Groundwater Resources and Environment, Ministry of Education, Jilin Univ., Changchun 130021, China; Ph.D. Candidate, College of New Energy and Environment, Jilin Univ., Changchun 130021, China. E-mail: luwx999@163.com

[3]Professor, Key Laboratory of Geographical Processes and Ecological Security in Changbai Mountains, Ministry of Education, School of Geographical Sciences, Northeast Normal University, Changchun, China, (corresponding author). E-mail: wuzf@nenu.edu.cn

[4]Professor, Key Laboratory of Geographical Processes and Ecological Security in Changbai Mountains, Ministry of Education, School of Geographical Sciences, Northeast Normal University, Changchun, China; School of Natural Resource, University of Missouri, Columbia, MO, USA, China. E-mail: hehs100@nenu.edu.cn

Abstract: In previous studies, a 0-1 mixed integer nonlinear programming optimization model (0-1MINLPOM) could only identify the location and release intensity for groundwater contamination sources (GCSs), and the location of each GCS was regarded as a 0-1 integer variable, selected from several locations determined in advance. However, in actual situations, the locations usually cannot be accurately isolated to a few GCSs and the number of GCSs is often unknown, so 0-1MINLPOM was improved in this study. Based on the estimation that there is a maximum of three GCSs in the study area, an improved 0-1 MINLPOM was established to simultaneously identify the number of GCSs (treated as 0-1 integer variable), the location (treated as integer variable) and release history of GCS (treated as continuous variables).

The simulation model was constructed as an equality constraint embedded improved 0-1 MINLPOM. In the improved 0-1 MINLPOM solution process, repeatedly calling the simulation model would have incurred a massive computational load and taken a long time. Thus, a surrogate model based on kriging and extreme learning machine (ELM) was established respectively for the simulation model to avoid this shortcoming. The results show that the accuracy of the kriging surrogate model (Krig-SM) was higher compared with the ELM surrogate model (ELM-SM). The improved 0-1 MINLPOM could identify the number, location, and release history of GCSs simultaneously. The accuracy of identifying the number of GCSs was 100%, and the accuracies of identifying the locations and release history were above 91.67% and 90.14%, respectively.

Keywords: Groundwater contamination, Mixed integer nonlinear programming, Source number, Surrogate model.

Introduction

Groundwater contamination beneath the surface of the earth is characterized by concealment and hysteresis of discovery, which leads to a lack of understanding of GCSs, including their number, location, and release history in aquifers (Mahinthakumar and Sayeed, 2005; Jha and Datta, 2013). The release history refers to the contaminant release intensity in each release period (Atmadja and Bagtzoglou, 2001; Sun et al., 2006). As a consequence, the design of groundwater contamination remediation schemes is difficult, and the responsibility for contamination liabilities and the assessment of contamination risks are also problematic (Lapworth et al., 2012; Om and Bithin, 2013). Developing an effective method for identifying GCSs would address these

problems, and have important significance (Bagtzoglou and Atmadja, 2005).

The identification of GCSs roughly started in the 1980s and many methods have been proposed for identifying GCSs, including optimization approaches (Gorelick et al., 1983; Mahar and Datta, 2000; Zhao et al., 2015), direct approaches (Skaggs and Kabala, 1994), probabilistic and geostatistical simulation approaches (Neupauer and Wilson, 2000), and analytical solution and regression approaches (Sidauruk et al., 2010).

Among those methods, simulation optimization methods have been used widely to characterize GCSs. Gorelick et al. (1983) applied a method based on least squares regression and linear programming to identify the location and release intensity of GCSs. Woodbury and Ulrych (1996) employed the minimum relative entropy method to recover the release and evolution history of contamination plumes in one-dimensional systems. Mahar and Datta (2000) used a nonlinear optimization model to identify the location and release intensity for GCSs in an unsteady flow system. Mahinthakumar and Sayeed (2006) employed hybrid optimization method to reconstruct release histories of GCSs. Jha and Datta (2011) applied a simulation-optimization based a variant of simulated annealing algorithm to identify the unknown source flux magnitude, duration and timing. Zhao et al. (2015) employed optimization approaches to identify the location and release intensity for GCSs. Xu and Gómez-Hernández (2016) used an ensemble Kalman filter to identify the source location, release time, and release concentration of GCSs.

Various preconditions are considered when applying simulation optimization methods to identify GCSs including: (1) the locations and number of GCSs are known to identify the release history (Zhao et al., 2016); (2) the suspected locations (3-4 specific locations) of the GCSs are known to identify the true location and release history (Ayvaz, 2010; Guo et al. 2018); (3) the

number (usually set to one) and all locations where the contaminant may be released are known to identify the true location and release history (Yeh et al., 2014; Xu and Gómez-Hernández, 2016).

Although good research results have been achieved in the application of simulation optimization methods for pollution source identification, there is no study have reported the application of simulation optimization methods for simultaneously identifying the number, location, and release history of GCSs. Guo et al. (2018) used the 0-1 MINLPOM to identify the location (treated as a 0-1 integer variable) and release intensity (treated as a continuous variable) of GCSs, but the number of GCSs was not identified. Thus, in the present study, an improved 0-1 MINLPOM was developed to simultaneously identify the number, location, and release history of GCSs. In the improved 0-1 MINLPOM, the number of GCSs is regarded as a 0-1 integer variable, and the locations and release history of GCSs are regarded as an integer variable and continuous variable, respectively.

When solving the improved 0-1 MINLPOM, the simulation model must be called repeatedly for calculations, thereby incurring a large calculation load and requiring a long time. This problem can be solved by establishing a surrogate model of the simulation model. The previously studies to identify GCSs based on surrogate models include kriging (Simpson et al., 2001; Luo and Lu, 2014a), radial basis function, support vector regression and extreme learning machines (Jiang et al., 2015; Hou and Lu, 2018).

The kriging and ELM methods were used to establish surrogate models for the simulation model, and the surrogate model with higher accuracy was embedded in the improved 0-1 MINLPOM. The improved 0-1 MINLPOM based on the surrogate model was then used to simultaneously identify the number, location, and release history of GCSs.

Methodology

Kriging

The kriging method is also known as the spatial local interpolation method and it is based on variation function theory and structural analysis for obtaining unbiased optimal estimates of regionalized variables in a finite region (Bargaoui and Chebbi., 2009; Kleijnen., 2009). In recent years, the kriging method has been extended as a surrogate modeling method with applications in many fields of engineering (Ryu et al., 2002; Kleijnen and van Beers, 2005; Coetzee et al., 2012). A kriging surrogate model is established according to the following principle:

$$Y(x)=\sum_{i=1}^{p} f_i(x)\cdot\beta_i+Z(x)=f^T\cdot\beta+Z\ , \tag{1}$$

where: x is the input value of the training samples, $f(x)=\left[f_1(x),f_2(x),\mathrm{L}\ ,f_p(x)\right]^{\mathrm{T}}$ are known determinate regression functions, $\beta=(\beta_1,\beta_2,\mathrm{L}\ ,\beta_p)^{\mathrm{T}}$ is a regression coefficient matrix estimated from the training samples, p is the number of determinate regression functions, and $Z(x)$ is a random part deviation of the regression model, which must satisfy the following conditions:

$$\begin{cases} E(z(x))=0 \\ D(z(x))=\sigma^2 \\ \mathrm{cov}[z(x_i),z(x_j)]=\sigma^2R(x_i,x_j) \end{cases}, \tag{2}$$

where $R(x_i,x_j)$ is the spatial correlation function between any two sampling points x_i and x_j, σ^2 is variance of $Z(x)$:

$$R(x_i,x_j)=\exp(-\sum_{k=1}^{n}\theta_k\left|x_k^i-x_k^j\right|)\quad(i=1,2,\mathrm{L}\ n,j=1,2,\mathrm{L}\ n), \tag{3}$$

where θ_k is the undetermined coefficient and x_k^i is the k-dimensional coordinate of the ith sample.

Using the input and output data for n known sample points, the output value corresponding to

any point x^* in the predicted feasible domain is:

$$Y(x) = f^T \beta^* + r^T(x) R^{-1}(y - f\beta^*), \tag{4}$$

where $r(x)$ is the correlation vector of point x^* and n sampling points $\{x_1, x_2 \mathrm{L}\ , x_n\}$, y is the matrix $n \times m$, n is the number of sampling points, m is the dimension of the output value, and β^* is the undetermined coefficient of the regression part, which can be obtained by the optimal linear unbiased estimation:

$$r^{\mathrm{T}}(x) = \begin{bmatrix} R(x_1^*, x_1) & \mathrm{L} & R(x_1^*, x_n) \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ R(x_n^*, x_1) & \mathrm{L} & R(x_n^*, x_n) \end{bmatrix} \tag{5}$$

$$\beta^* = (f^T R^{-1} f)^T f^T R^{-1} y, \tag{6}$$

where R is the correlation matrix comprising the correlation coefficients of n sampling points:

$$R = \begin{bmatrix} R(x_1, x_1) & \mathrm{L} & R(x_1, x_n) \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ R(x_n, x_1) & \mathrm{L} & R(x_n, x_n) \end{bmatrix}. \tag{7}$$

The variance estimate value σ^2 is determined by:

$$\sigma^2 = \frac{1}{n}(y - f\beta^*)^T R^{-1}(y - f\beta^*) \tag{8}$$

$$\max_{\theta_k} \left\{ -[n \ln \sigma^2 + \ln |R|] \right\}. \tag{9}$$

The surrogate model can be established by solving the nonlinear unconstrained optimization problem defined above.

Extreme learning machine

Huang et al. (2004) proposed the ELM method to improve backward propagation neural networks in order to improve the low efficiency of learning, while also simplifying the settings for the learning parameters. Compared with the backward propagation neural network algorithm, the ELM method has advantages in terms of its rapid learning speed and good generalization performance (Huang et al., 2006). ELM is a type of machine learning method based on a

feedforward neural network (Huang et al., 2015), as illustrated in Fig. 1.

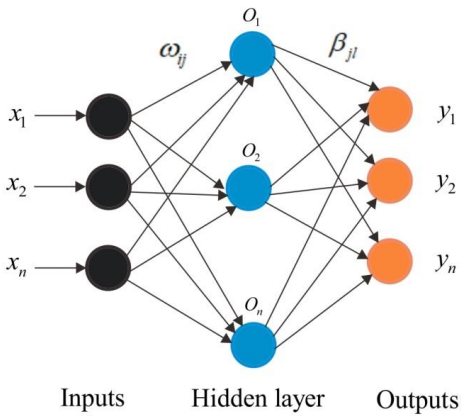


Fig. 1. Schematic diagram showing the single hidden layer feedforward neural network of ELM.

Thus, ELM was used to establish the surrogate model for the simulation model because of these advantages. A single hidden layer feedforward neural network comprises an input layer, hidden layer, and output layer. The input layer is fully connected to the hidden layer, and the hidden layer is fully connected to the output layer. The input layer contains m neurons corresponding to each input variable, the hidden layer has n neurons, and the output layer has k neurons corresponding to each output variable. The connection weight matrix between the input layer and the hidden layer is ω. The connection weight matrix between the hidden layer and output layer is β. The threshold value matrix for the hidden layer neurons is b. These matrices are expressed as follows:

$$\omega = \begin{bmatrix} \omega_{11} & \mathrm{L} & \omega_{1m} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \omega_{1n} & \mathrm{L} & \omega_{nm} \end{bmatrix}_{n\times m} \tag{10}$$

$$\beta = \begin{bmatrix} \beta_{11} & \mathrm{L} & \beta_{1k} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \beta_{1n} & \mathrm{L} & \beta_{nk} \end{bmatrix}_{n\times k} \tag{11}$$

$$b = \begin{bmatrix} b_1 \\ \mathrm{M} \\ b_n \end{bmatrix}_{n\times 1}, \tag{12}$$

where ω_{ij} is the connection weight between the i-th neuron in the input layer and the j-th neuron in the hidden layer, and β_{jl} is the connection weight between the j-th neuron in the

hidden layer and the l -th neuron in the input layer.

The input matrix X and output matrix Y comprise training data sets containing M samples, as follows.

$$\mathrm{X}=\begin{bmatrix} x_{11} & \mathrm{L} & x_{1M} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ x_{m1} & \mathrm{L} & x_{mM} \end{bmatrix}_{m\times M} \tag{13}$$

$$\mathrm{Y}=\begin{bmatrix} y_{11} & \mathrm{L} & y_{1M} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ y_{l1} & \mathrm{L} & y_{lM} \end{bmatrix}_{l\times M} \tag{14}$$

The activation function for the hidden layer neurons is $g(x)$ and the output T from the network is:

$$\mathrm{T}=\begin{bmatrix} t_1 & \mathrm{L} & t_M \end{bmatrix}_{l\times M} \tag{15}$$

$$t_j=\begin{bmatrix} t_{1j} \\ \mathrm{M} \\ t_{lj} \end{bmatrix}=\begin{bmatrix} \sum_{i=1}^{n}\beta_{i1}g(\omega_i x_j+b_i) \\ \mathrm{M} \\ \sum_{i=1}^{n}\beta_{il}g(\omega_i x_j+b_i) \end{bmatrix}_{l\times 1} \quad (j=1,2\mathrm{L}\ ,M)\,. \tag{16}$$

The objective when training a single hidden layer neural network is to minimize the error between the output data and training data, which can be expressed as follows.

$$\sum_{j=1}^{M}\left\|t_j-y_j\right\|=0 \tag{17}$$

$$\omega_i=\begin{bmatrix}\omega_{i1} & \mathrm{L} & \omega_{im}\end{bmatrix};\ x_j=\begin{bmatrix}x_{1j} & \mathrm{L} & x_{mj}\end{bmatrix}^{\mathrm{T}};\ y_j=\begin{bmatrix}y_{1j} & \mathrm{L} & y_{mj}\end{bmatrix}^{\mathrm{T}} \tag{18}$$

Thus, we calculate ω, β, and b while ensuring that Eq. (19) holds:

$$\begin{bmatrix} \sum_{i=1}^{n}\beta_{i1}g(\omega_i x_1+b_i) & \mathrm{L} & \sum_{i=1}^{n}\beta_{i1}g(\omega_i x_M+b_i) \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \sum_{i=1}^{n}\beta_{il}g(\omega_i x_1+b_i) & \mathrm{L} & \sum_{i=1}^{n}\beta_{il}g(\omega_i x_M+b_i) \end{bmatrix}=\mathrm{Y}^{\mathrm{T}}\,. \tag{19}$$

The input weight and the hidden layer threshold are determined randomly before training and they remain unchanged during the training process, so the hidden layer's output matrix H is uniquely determined. Training a single hidden layer neural network can be transformed into solving the linear equations in Eq. (21) to obtain the output weight matrix. The final output layer

connection weight matrix can be obtained by solving the least squares solution of Eq. (20):

$$H\beta = Y^{\mathrm{T}} \tag{20}$$

$$\min_{\beta} \left\| H\beta - \mathrm{Y}^{\mathrm{T}} \right\| \tag{21}$$

$$H = \begin{bmatrix} g(\omega_1 x_1 + b_1) & \mathrm{L} & g(\omega_i x_1 + b_i) \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ g(\omega_1 x_M + b_1) & \mathrm{L} & g(\omega_i x_M + b_i) \end{bmatrix}_{M \times n} \tag{22}$$

$$\hat{\beta} = H^{+}\mathrm{Y}^{\mathrm{T}}, \tag{23}$$

where H^{+} is the Moore-Penrose generalized inverse matrix of the hidden layer output matrix H. The output corresponding to any input can be predicted after calculating the output layer connection weight matrix.

Numerical applications

Site overview

A hypothetical contaminated site was considered as a case study in order to analyze the application of the research methods for the identification of groundwater contamination sources. The hypothetical contaminated site was designed to represent the conditions that may occur in most actual problems, including the aquifer geometry, boundary conditions, initial conditions, hydrogeological parameters, and groundwater flow characteristics (Prakash and Datta, 2014; Datta et al., 2017). Contaminant is considered to be divalent manganese ion which do not undergo chemical and biological conversion. The study area comprised a two-dimensional, heterogeneous, isotropic, submerged aquifer with irregular boundaries, where the groundwater flow was transient. There three parameter zones in the aquifer and their medium types were as follows: I, coarse sand, II, medium sand and III, fine sand. The boundary conditions, locations of observation wells and potential contamination source area (all locations where the contaminant might have been released) are shown in Fig. 2, and the other parameters related to the aquifer are presented in Table 1. The hydraulic head of the northwest specific head boundary was 27 m and the hydraulic head of the

southeast specific head boundary was 25 m. The vertical direction of the study area received uniform recharge from atmospheric rainfall with a recharge amount of 730 mm/a. The initial concentration of the contaminant in the study area was 0 g/L.

In order to test the effectiveness of the identification method developed in this study, two hypothetical cases are designed in the paper. The aquifer parameters of hypothetical contaminated site were the same for the two cases, as shown in Table 1.

Table 1. Parameters for the aquifers in the hypothetical contaminated site

Parameters Values	Value		
	I	II	III
Hydraulic conductivity in x-direction, Kxx (m/d)	42	30	18
Hydraulic conductivity in y-direction, Kyy (m/d)	42	30	18
Effective porosity, θ	0.25	0.2	0.18
Longitudinal dispersivity, αL (m)	40	40	40
Transverse dispersivity, αT (m)	8	8	8
Saturated thickness, b(m)	35	35	35
Grid spacing in x-direction, Δx (m)	3	3	3
Grid spacing in y-direction, Δy (m)	3	3	3
Length of the simulation period, Δt (month)	6	6	6
Volume flux per unit area, W (m/d)	0.0004	0.00035	0.0003
Initial concentration (g/L)	0	0	0

The study area was discretely divided into 5262 grids. Since the divided grid is treated as an integer variable, the location information of the potential contamination source area in the study area is represented by grid numbers. The horizontal and longitudinal locations of the potential contamination source areas and their corresponding grid numbers are shown in Table 2.

Table 2. The horizontal and longitudinal locations and their corresponding grid numbers

(a) horizontal

Number	Location	Number	Location	Number	Location	Number	Location
19	55.5	32	94.5	45	133.5	58	172.5
20	58.5	33	97.5	46	136.5	59	175.5
21	61.5	34	100.5	47	139.5	60	178.5
22	64.5	35	103.5	48	142.5	61	181.5
23	67.5	36	106.5	49	145.5	62	184.5
24	70.5	37	109.5	50	148.5	63	187.5
25	73.5	38	112.5	51	151.5	64	190.5
26	76.5	39	115.5	52	154.5	65	193.5
27	79.5	40	118.5	53	157.5	66	196.5
28	82.5	41	121.5	54	160.5	67	199.5
29	85.5	42	124.5	55	163.5	68	202.5

30	88.5	43	127.5	56	166.5	69	205.5
31	91.5	44	130.5	57	169.5	70	208.5

(b) longitudinal

Number	Location	Number	Location	Number	Location
18	157.5	31	196.5	44	235.5
19	160.5	32	199.5	45	238.5
20	163.5	33	202.5	46	241.5
21	166.5	34	205.5	47	244.5
22	169.5	35	208.5	48	247.5
23	172.5	36	211.5	49	250.5
24	175.5	37	214.5	50	253.5
25	178.5	38	217.5	51	256.5
26	181.5	39	220.5	52	259.5
27	184.5	40	223.5	53	262.5
28	187.5	41	226.5	54	265.5
29	190.5	42	229.5	55	268.5
30	193.5	43	232.5		

The real number, location, and release history of GCSs for the two cases are designed as follows:

Case one: the number of GCSs was one, the location of the GCSs is as shown in Fig. 2a and release intensities of the GCSs during two release periods are as shown in Table 3a.

Case two: the number of GCSs was two, the locations of the two GCSs are shown in Fig. 2b and release intensities of the two GCSs during two release periods are as shown in Table 3b.

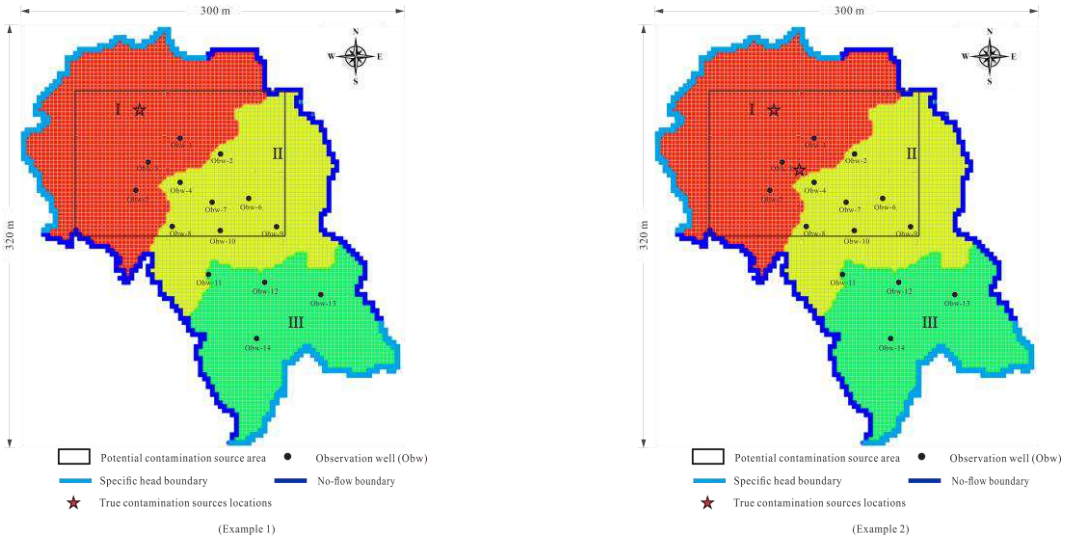
Table 3. Details of the real GCSs for case two: (a) case one; (b) case two.

(a) case one

Number of sources	Grid number		Release intensity (g/d)	
	Horizontal	longitudinal	SP1	SP2
1	35	52	260	140

(b) case one

Number of sources	Sources	Grid number		Release intensity (g/d)	
		Horizontal	longitudinal	SP1	SP2
2	S1	35	52	190	120
	S2	41	36	170	160



(a) (b)

Fig.2. Boundary conditions, location of observation wells, potential contamination source area and true location of GCSs in the study area of case two: (a) case one; (b) case two.

During the inverse identification of GCSs process, the number, location, and release history of the GCSs for the two cases were regarded as unknown, what known is only the potential contamination source area (shown in Fig.2a and Fig.2b). For hypothetical case one and case two it is estimated that the maximum number of GCSs is 3. The number, location, and release intensity of the GCSs were then identified based on optimization model.

All the GCSs are assumed to have same activity initiation times (the first day of simulation) and the time lag between the initial release contaminant of the GCSs and first contaminant concentration measurement of the two cases are shown in Table 4. The activity duration of two cases is 300 days. The entire activity duration of the sources is divided into two equal stress periods of 150 days. Concentration is measured every 90 days, starting from the first concentration measurement time. The contaminant flux from the sources is assumed to be constant over a stress period. The total simulation time for contaminant transport is 900 days. A total of eight periods concentration measurements from each of the fourteen observation wells are utilized for the two cases.

Table 4. Test case scenarios

Case	Activity duration (d)	Time lag between the initial release contaminant of the source and first concentration measurement (d)	Concentration measurement intervals (d)
one	2×90	270	90

Numerical simulation model

In order to identify the GCSs, a numerical simulation model of the groundwater flow and contaminant transport was established based on the specific conditions of the study area. The governing partial differential equations for the groundwater flow and the contaminant transport of the transient flow in the two-dimensional aquifer system are defined as follows (Pinder and Bredehoeft, 1968; Singh and Datta, 2006):

The governing partial differential equation for the groundwater flow is as follows:

$$\frac{\partial}{\partial x_i}(K_{ij}\frac{\partial H}{\partial x_j})+W=\mu\frac{\partial H}{\partial t} \quad (x,y)\in\Omega \quad i,j=1,2 \quad t\geq 0, \tag{25}$$

where μ is the specific yield, H is the hydraulic head, K_{ij} is the hydraulic conductivity, Ω is the simulated area range, and W is the volumetric flux per unit volume.

The governing partial differential equations for the contaminant transport are as follows:

$$\frac{\partial C}{\partial t}=\frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial C}{\partial x_j}\right)-\frac{\partial}{\partial x_i}(u_i C)+\frac{R}{\theta} \quad (x,y)\in\Omega \quad i,j=1,2 \quad t\geq 0 \tag{26}$$

$$u_i=\frac{K_{ij}}{\theta}\frac{\partial H}{\partial x_i} \quad i,j=1,2, \tag{27}$$

where Ω is the simulated area range, θ is the effective porosity of the aquifer medium, c is the contamination concentration, D_{ij} is the dispersion coefficient, u_i is the average linear velocity of the groundwater flow determined by Darcy's law, and R is the source or sink term.

After establishing the groundwater flow and contaminant transport numerical simulation model, the GMS software was used to solve the simulation model.

In contrast to the actual problem, the hypothetical example had no actual data measurements. Therefore, it was necessary to forward run the contaminant transport simulation model and obtain the contaminant concentration data for the observation wells during each simulation period for use as the data measurements during the identification process. Fig. 3 shows the contaminant plume distributions on day 450 and day 900 of each case. Fig. 4 show the values measured for each

observation well in each simulation period for the two contaminated cases.

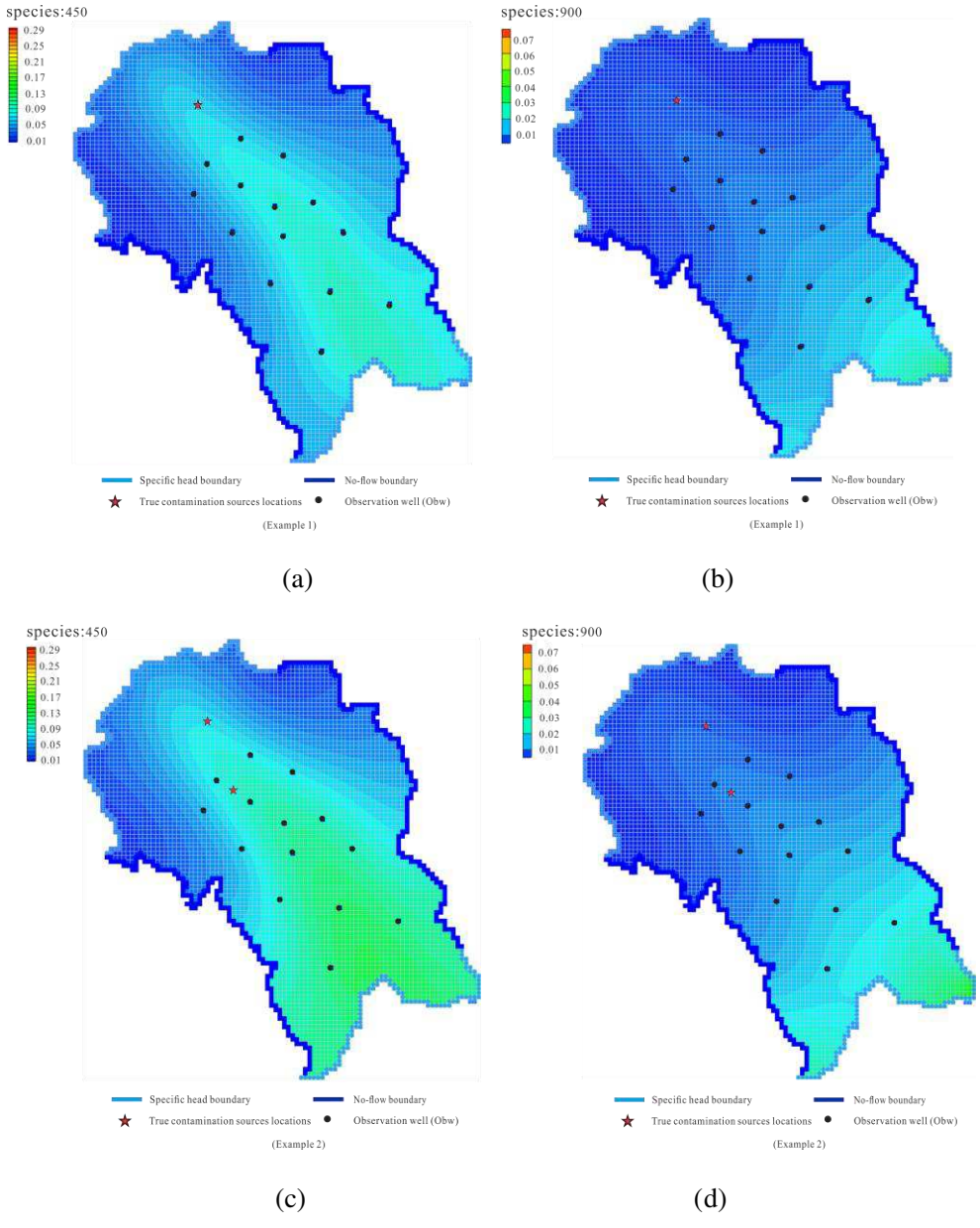
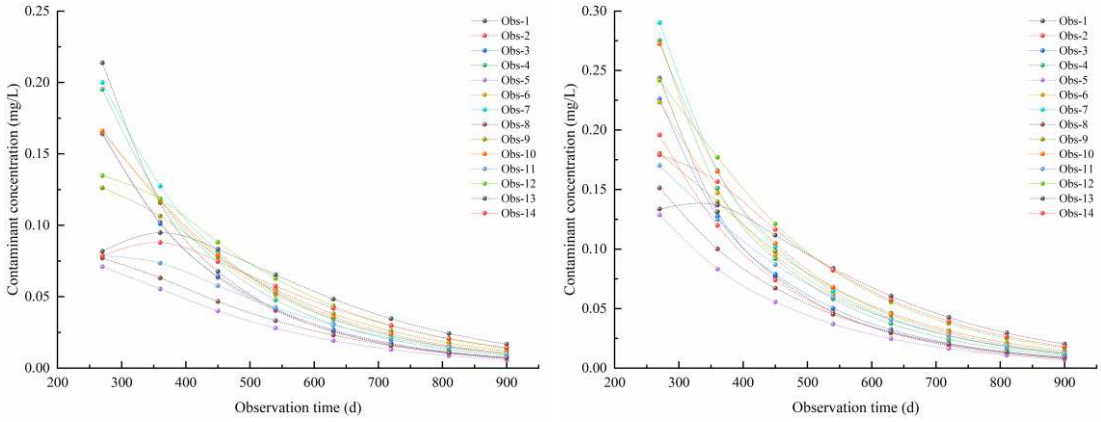


Fig. 3. Contaminant plume distributions: (a), (b) 450 d, 900 d for case one, respectively; (c), (d), 450 d, 900 d for case two, respectively.



(a) (b)

Fig. 4. Values measured of the observation wells for two cases: (a) case one; (b) case two.

Surrogate models of the numerical simulation model

A surrogate model with satisfactory accuracy has almost the same input-output relationship as the simulation model. The advantage of applying a surrogate model instead of a simulation model is that the surrogate model is not only simpler to call, but it can reduce a lot of computational load and time required when solving the optimization model (Simpson et al., 2001; Luo and Lu, 2014a Luo et al., 2014b; Jiang et al., 2015).

The horizontal and longitudinal grid numbers and release history corresponding to a maximum number of the GCSs were used as the input variables for the surrogate model (six grid numbers variables and six release intensity variables comprising a total of 12 input variables). The contaminant concentrations in each observation well during each simulation period were treated as the output variables in the surrogate model. Using the Latin hypercube method, the grid numbers and release history of each zone were sampled in their feasible domains. The feasible domain ranges are shown in equations (30). In total, 420 groups were sampled where the grid numbers and release intensities for the 420 groups were used sequentially as inputs for the simulation model. The corresponding outputs comprising the concentrations in all of the observation wells in each period were obtained by running the contaminant transport numerical simulation model. 360 groups of input–output data were selected as the training samples for the surrogate model and 60 groups of input-output data were selected as the testing samples for the surrogate model. The kriging and ELM methods were coded with Matlab, and the Krig-SM and ELM-SM were then established based on the training samples.

The accuracy of the two surrogate models was tested based on three evaluative coefficients comprising the certainty coefficient (R^2), root mean square error (RMSR), and mean relative error (MRE).

R^2 was calculated as follows.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{28}$$

RMSR was calculated as follows.

$$RMSR = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{29}$$

MRE was calculated as follows.

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{30}$$

In Eqs 28–30, y_i is the output value of the *i*th sample obtained from the contaminant transport numerical simulation model, $\hat{y}_i$ is the output value of the *i*th sample obtained from the surrogate model, and $\overline{y}_i$ is the average of the output values for *n* samples obtained from the contaminant transport simulation model. Smaller values for the MRE and RMSR coefficients, as well as values of R^2 closer to 1 indicate that the surrogate model is more accurate at simulating the output of the simulation model, thereby demonstrating that the surrogate model can be used to replace the simulation model.

0-1 MINLPOM

The optimization model is the basis for the identification of GCS. Therefore, establishing the optimization model is an essential step in the research process (Cristo et al., 2008; Guneshwor et al., 2018). The optimization model comprises three important components: an objective function, decision variables, and constraints.

Based on measured values of groundwater contaminant the optimization model for identifying the characteristics of GCSs was established. The number, location, and release history of the GCSs were decision variables in the optimization model. The number of GCSs was treated as integer variable, which could only be a value of 0 or 1. The location (grid numbers) and release history were treated as integer variables and continuous variables, respectively. The fitting error between the measured and simulated contaminant concentrations in the observation wells during each simulation period was minimized as the objective function. The contaminant transport numerical simulation model (replace with the surrogate model) was embedded in the optimization model as an equality constraint to ensure that the optimization model satisfied the contaminant transport law in a groundwater system during the optimization process. The feasible domains for the number, location, and release history of GCSs comprised the inequality constraints. The objective function and constraints constituted the 0-1 MINLPOM used to identify the GCSs. The 0-1 MINLPOM of two cases are expressed as follow:

$$\min z(\varphi_i, x_i, y_i, q_m) = \sum_{t=1}^{8}\sum_{k=1}^{14}(C_k^t(t) - C_k^t(0))^2$$

$$\begin{cases} \varphi_i = \begin{cases} 0 \\ 1 \end{cases} & i = 1, 2, 3 \\ \sum_{i=1}^{3} \varphi_i \le 3 \\ \varphi_1 \le \varphi_2 \le \varphi_3 \\ 19 \le x_i \le 70, 18 \le y_i \le 55 \quad i = 1, 2, 3 \quad x_i, y_i \in N^* \\ y_1 \le y_2 \le y_3 \\ 0 \le q_m \le 300 \quad m = 1, 2, \mathrm{L}, 6 \\ C_k^t(t) = f(q_m) \end{cases} \tag{31}$$

where φ_i is the 0-1 integer variable representing whether the current location is GCS, 1 indicates that a real GCS is present in the current location, 0 indicates that no GCS is present in the current location, x_i is the horizontal grid numbers, y_i is the longitudinal grid numbers, i represents the i-th potential GCS, q_m is the release intensity of the GCSs during each release

period, m represents the m-th release intensity variable (each GCS contains two release intensity variables), $C_k^t(t)$ is the simulated concentration of the contaminant at the observation point, and $C_k^t(0)$ is the measured value of the contaminant concentration at the observation point.

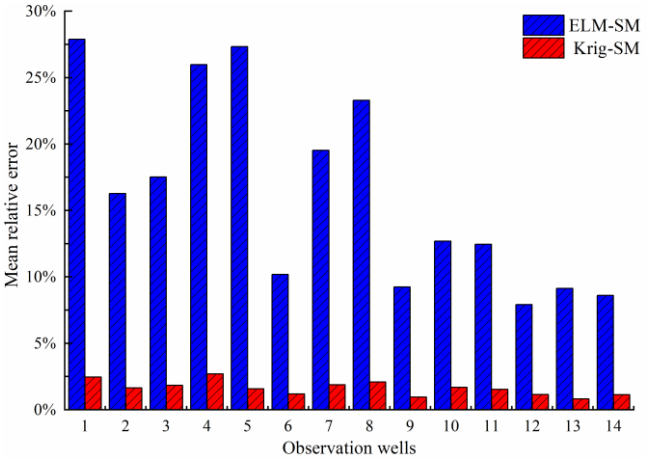
The 0-1MINLPOM for identifying the GCSs in each case was exactly the same, except the contaminant concentrations values measured in observation wells corresponding to the two cases were different.

Results and discussion

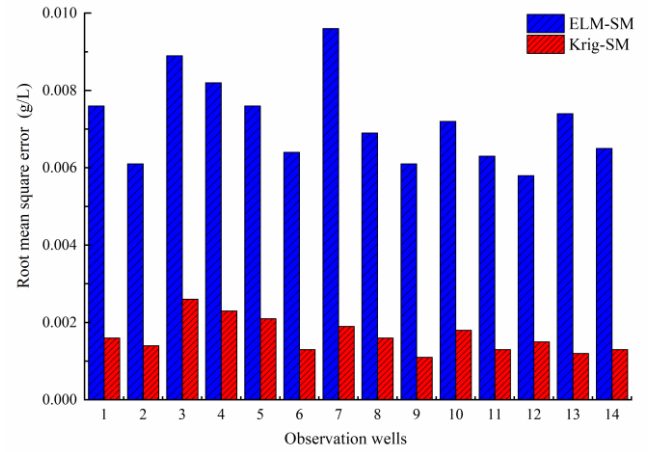
Comparative analysis of the accuracy of the surrogate models

The accuracies of ELM-SM and Krig-SM were compared, and the surrogate model with higher accuracy was selected and embedded in the 0-1 MINLPOM to be called when solving the 0-1 MINLPOM, thereby allowing the number, location (grid numbers), and release history of the GCSs to be identified simultaneously.

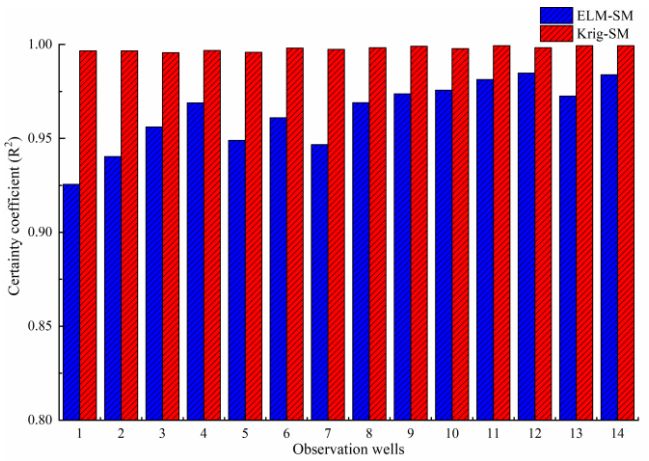
In order to test the accuracy of the two surrogate models, 60 groups of input test samples were entered in the simulation model and 60 groups of output samples were obtained. The same 60 groups of input test samples were then entered in the two surrogate models and 60 groups of output samples were obtained from the two surrogate models. The accuracies of the two surrogate models were evaluated by comparing the output samples from the simulation model and those from the two surrogate models. In order to clearly compare the accuracies of the two surrogate models, the contrast histograms were plotted for MRE, RMSR, and R^2 (Fig. 5).



(a)



(b)



(c)

Fig. 5. Evaluative coefficients of surrogate for the two cases: (a) mean relative error (MRE);(b) Root mean square error (RMSR);(c) Certainty coefficient (R^2).

The minimum certainty coefficients with Krig-SM of case one and case two were all above

0.99. The maximum and minimum certainty coefficients with ELM-SM were 0.9934 and 0.9255, respectively (show in Fig. 5),which demonstrate that Krig-SM obtained a higher R^2 than ELM-SM. Fig. 5 show that the RMSR and MRE values were smaller with Krig-SM than ELM-SM. Thus, Krig-SM obtained a better approximation of the simulation model and it had higher accuracy. Therefore, Krig-SM was embedded in the 0-1 MINLPOM for simultaneously identifying the number, location, and release history of GCSs.

Analysis of GCSs identification results

Krig-SM is embedded in the 0-1 MINLPOM and the GA was used to solve the 0-1 MINLPOM to identify the number, location, and release history of GCSs for two cases. A detailed introduction to GAs was provided by Guo et al. (2018) and Zwickl (2006). The parameter settings for the GA are shown in Table 5.

When solving the 0-1 MINLPOM, it takes about 2200 hours to call the simulation model for 240000 times, while it only takes about 0.21 hours, to call the surrogate model for 240000 times. Thus, using the surrogate model instead of the simulation model for calculation could save 99% of the computational load and the computational time required.

The identification results of GCSs for the two cases are shown in Table 6 and Fig.6.

Case one: The identification results for GCSs are shown in Fig.6a and Table 6a. A value “1” (shown in Fig. 6a) in the identification results indicates that the number of GCSs in the study area was one. The location and the release intensities of GCSs are shown in Table 6a. The identification accuracy of the number and location was 100%, and the identification accuracy of release history was above 93.59%.

Case two: The identification results for GCSs are shown in Fig.6b and Table 6b. Those two values of "1" (shown in Fig. 6b) in the identification results indicate that the number of GCSs was two. The location and the release intensities of the two GCSs are shown in Table 6b. The identification accuracy of the number was 100%, and the identification accuracy of location and release history was above 91.67% and 90.14%, respectively.

Table 5. Parameter settings for the genetic algorithm

Parameter	Setting
Population size	200
Scaling function	Rank
Selection function	Stochastic uniform
Mutation function	Constraint dependent
Crossover function	Scattered
Direction	Forward
Stall generations	1200
Other parameter settings	Default

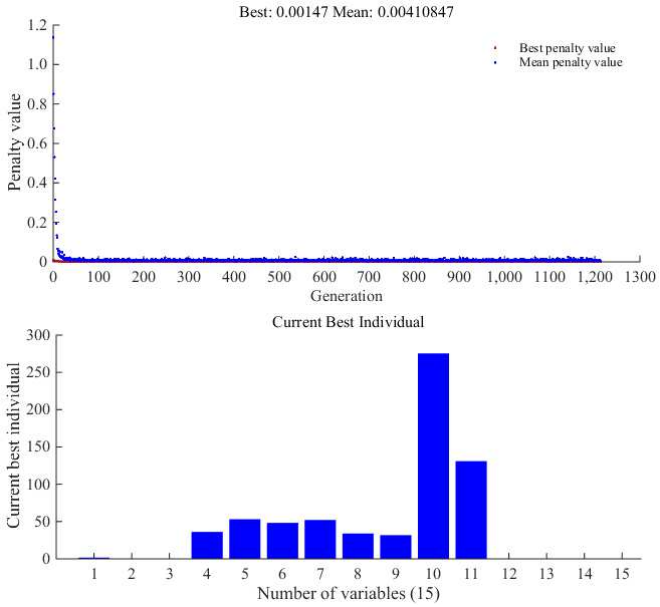
Table 6. Source number, location, and release history results for case two

(a) case one

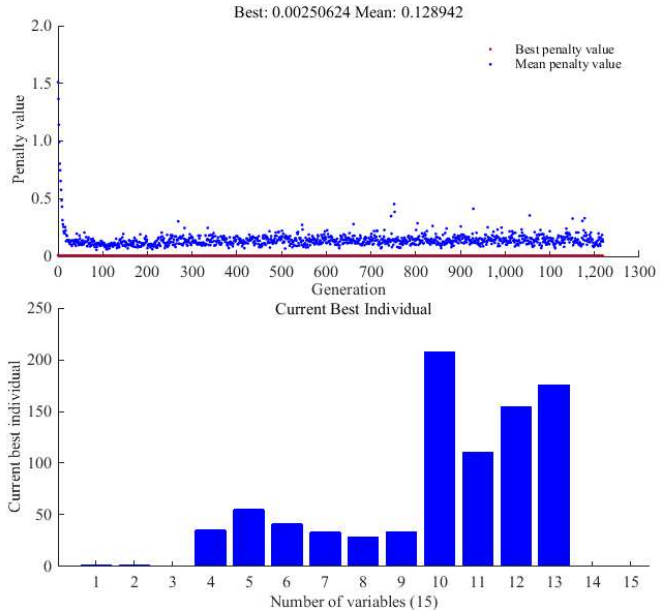
Example 1	Number of sources	S-location (m)		S-Release intensity (g/d)	
		X	Y	SP1	SP2
Real value	1	35	52	260	140
Identified results	1	35	52	271.73	131.03
Relative error	0.00%	0.00%	0.00%	4.51%	6.41%

(b) case two

Example 2	Number of sources	S1-location (m)		S2-location (m)		S1-Release intensity (g/d)		S2-Release intensity (g/d)	
		X	Y	X	Y	SP1	SP2	SP1	SP2
Real value	2	35	52	41	36	190	120	170	160
Identified results	2	35	55	41	33	207.89	110.42	154.73	175.78
Relative error	0.00%	0.00%	5.77%	0.00%	8.33%	9.41%	7.98%	8.98%	9.86%



(a)



(b)

Fig. 6. The convergence diagram of GA for the two cases: (a) case one; (b) case two.

The identification results of GCSs for the two cases showed that the proposed method performed well at simultaneously identifying the number, location, and release history of the GCSs.

Discussion

Comparing with the 0-1 MINLPOM proposed by previous researchers, the improved 0-1 MINLPOM proposed in this study is more applicable to the identification of GCSs, and can simultaneously identify the number, location, and release history of the GCSs.

In this study, the actual number of GCSs was less than the estimated maximum number of three. If the actual number of GCSs is greater than or equal to the maximum number of GCSs, the solution is increasing the original estimated maximum number of GCSs. Then establishing a new surrogate model and optimization model to identify the information for the GCSs until the number of GCSs identified is less than the estimated maximum number of GCSs (in a feedback correction process).

However, the number of input variables of the numerical simulation model will increase as the estimated maximum number of GCSs increases and the nonlinear relationships will become more complicated. When faced with simulation model with more complex nonlinear relationship, the accuracy of the surrogate model established by the traditional shallow learning method (Kriging and extreme learning machine, etc) may be unsatisfactory. Thus, methods for establishing surrogate model with better accuracy should be studied. Deep learning neural networks and deep reinforcement learning neural networks have strong fitting capabilities for simulation models with complex nonlinear relationships and have great potential to establish surrogate model in future research.

Conclusions

In this study, an improved 0-1 MINLPOM was developed to simultaneously identify the

number, location, and release history of GCSs. Based on the study two conclusions were obtained:

First, the results showed that the surrogate model established using the kriging method was more accurate than the ELM. Thus, Krig-SM can be used instead of the simulation model to embed in the improved 0-1 MINLPOM during solution process. The surrogate model provided a highly accurate approximation of the simulation model, and applying Krig-SM instead of the simulation model in calculation can also save 99% of the computational load and the computational time required.

Second, in previous methods employed for the identification of GCSs, the improved 0-1 MINLPOM was used to identify only the location and release history of the GCSs. In the present study, the number of GCSs was treated as integer variable with a value of 0 or 1, the location was treated as integer variable and release history of GCSs were treated as continuous variables. The number, location, and release history of GCSs were simultaneously identified using the 0-1 improved MINLPOM. The results showed that the identification results were very similar to the true values of the GCSs characteristics. The accuracy of identifying the number of GCSs was 100%, and the accuracies of identifying the locations and release intensities were over 91.67% and 90.14%, respectively. This method can effectively solve the problem of identifying the number of GCSs, as well as simultaneously identifying their locations and release history.

Data Availability Statement

All of the data, models, or code generated or used in this study are available from the corresponding author by request.

Acknowledgments

The authors acknowledge support provided by the National Key R&D Program of China (Grant Nos. 2019YFC0409101) and the National Nature Science Foundation of China (Grant Nos. 41972252). Special gratitude is extended to the journal editors for their efforts in evaluating this study. The valuable comments provided by the anonymous reviewers are also gratefully acknowledged.

Declarations

The authors have no relevant financial or non-financial interests to disclose.

References

Atmadja, J., Bagtzoglou, A.C. (2001). State of the art report on mathematical methods for groundwater pollution source identification. Environmental Forensics., 2(3), 205–214. https://doi.org/10.1006/enfo.2001.0055

Ayvaz M.T. (2010). A linked simulation–optimization model for solving the unknown groundwater pollution source identification problems. Journal of Contaminant Hydrology., 117(1-4),46–59. https://doi.org/10.1016/j.jconhyd.2010.06.004

Mahsa, A., Datta, B. (2013). Identification of contaminant source characteristics and monitoring network design in groundwater aquifers: An overview. Journal of Environmental Protection., 4(5),26-41.

Bagtzoglou, A.C., Atmadja, J. (2005). Mathematical methods for hydrologic inversion: The case

453 of pollution source identification. Springer Berlin Heidelberg. https://doi.org/10.1007/b11442
454 Bargaoui, Z.K., Chebbi, A. 2009. “Comparison of two kriging interpolation methods applied to
455 spatiotemporal rainfall”. Journal of Hydrology., 365(1-2), 56-73.
456 https://doi.org/10.1016/j.jhydrol.2008.11.025
457 Cristo, C.D., Leopardi, A. (2008). Pollution source identification of accidental contamination in
458 water distribution networks. Journal of Water Resources Planning & Management., 134,
459 197-202. https://doi.org/ 10.1061/(ASCE)0733-9496(2008)134:2(197)
460 Coetzee, W., Coetzer, R.L., Rawatlal, R. (2012). Response surface strategies in constructing
461 statistical bubble flow models for the development of a novel bubble column simulation
462 approach. Comput. Chem. Eng., 36, 22–34.
463 https://doi.org/10.1016/j.compchemeng.2011.07.014
464 Datta, B., Petit, C., Palliser, M., Esfahani, H. K., & Prakash, O. (2017). Linking a simulated
465 annealing based optimization model with PHT3D simulation model for chemically reactive
466 transport processes to optimally characterize unknown contaminant sources in a former mine
467 site in Australia. Journal of Water Resource and Protection., 9(5), 432-454.
468 Gorelick, S.M., Evans, B., Remson, I. (1983). Identifying sources of groundwater pollution: An
469 optimization approach. Water Resources Research., 19(3).
470 https://doi.org/10.1029/WR019i003p00779
471 Guo, J. Y., Lu, W. X., Yang, Q. C., et al. (2018). The application of 0–1 mixed integer nonlinear
472 programming optimization model based on a surrogate model to identify the groundwater
473 pollution source. Journal of Contaminant Hydrology.,
474 https://doi.org/10.1016/j.jconhyd.2018.11.005

Guneshwor, L., Eldho, T.I., Kumar, A.V. (2018). Identification of groundwater contamination sources using meshfree rpcm simulation and particle swarm optimization. Water Resources Management., 32(4), 1517-1538. https://doi.org/10.1007/s11269-017-1885-1

Huang, G. B., Zhu, Q. Y., Siew, C.K. (2004). Extreme learning machine: a new learning scheme of feedforward neural networks. Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. IEEE. https://doi.org/10.1109/IJCNN.2004.1380068

Huang, G.B., Zhu, Q.Y., Siew, C.K. (2006). Extreme learning machine: Theory and applications. Neurocomputing., 70 (1-3) :489-501. https://doi.org/ 10.1016/j.neucom.2005.12.126

Huang, G., Huang, G. B., Song, S., et al. (2015). Trends in extreme learning machines: A review. Neural Networks., 61, 32–48. https://doi.org/10.1016/j.neunet.2014.10.001

Hou, Z.Y., Lu, W.X. (2018). Comparative study of surrogate models for groundwater contamination source identification at DNAPL-contaminated sites. Hydrogeology Journal. 26 (3):923-932. https://doi.org/ 10.1007/s10040-017-1690-1

Jiang, X., Lu, W.X., Hou, Z.Y., et al. (2015). Ensemble of surrogates-based optimization for identifying an optimal surfactant-enhanced aquifer remediation strategy at heterogeneous DNAPL-contaminated sites. Computers & Geosciences., 84, 37–45. https://doi.org/10.1016/j.cageo.2015.08.003

Kleijnen, J.P. (2009). Kriging metamodeling in simulation: A review. European journal of operational research., 192(3), 707-716. https://doi.org/10.1016/j.ejor.2007.10.013

Kleijnen, J.P.C., Beers, W.C.M.V. (2005). Robustness of Kriging when interpolating in random simulation with heterogeneous variances: some experiments. European Journal of Operational Research., 165(3), 826–834. https://doi.org/10.1016/j.ejor.2003.09.037

497 Lapworth, D.J., Baran, N., Stuart, M.E., et al. (2012). Emerging organic contaminants in
498 groundwater: A review of sources, fate and occurrence. Environmental Pollution., 163(4):
499 287-303. https://doi.org/10.1016/j.envpol.2011.12.034
500 Luo, J.N., Lu, W.X. (2014a). Comparison of surrogate models with different methods in
501 groundwater remediation process. Journal of Earth System Science., 123(7),1579–1589.
502 https://doi.org/10.1007/s12040-014-0494-0
503 Luo, J.N., Lu, W.X. (2014b). A mixed-integer non-linear programming with surrogate model for
504 optimal remediation design of NAPLs contaminated aquifer. International Journal of
505 Environment and Pollution., 54 (1), 1–16. https://doi.org/10.1504/IJEP.2014.064047
506 Mahar, P.S., Datta, B. (2000). Identification of pollution sources in transient groundwater systems.
507 Water Resources Management., 14.3, 209–227. https://doi.org/10.1023/A:1026527901213
508 Mahinthakumar, G.K., Sayeed, M. (2006). Reconstructing groundwater source release histories
509 using hybrid optimization approaches. Environmental Forensics, 7 (1): 45-54. https://doi.org/
510 10.1080/15275920500506774
511 Jha, MK., Datta, B. (2011). Simulated annealing based simulation-optimization approach for
512 identification of unknown contaminant sources in groundwater aquifers. Desalination and
513 Water treatment., 32(1-3):79-85. https://doi.org/10.5004/dwt.2011.2681
514 Jha, M., Datta, B. (2013). Three-dimensional groundwater contamination source identification
515 using adaptive simulated annealing. Journal of Hydrologic Engineering., 18(3), 307-317.
516 https://doi.org/10.1061/(ASCE)HE.1943-5584.0000624
517 Neupauer, R.M., Borchers, B., Wilson, J.L. (2000). Comparison of inverse methods for
518 reconstructing the release history of a groundwater contamination source. Water Resources

519 Research., 36(9), 2469-2475. https://doi.org/10.1029/2000wr900176

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

524 Pinder, G.F., Bredehoeft, J.D. (1968). Application of the digital computer for aquifer evaluations".
525 Water Resources Research., 4 (5), 1069–1093. https://doi.org/10.1029/WR004i005p01069

526 Prakash, O., Datta, B. (2014). Characterization of groundwater pollution sources with unknown
527 release time history. Journal of Water Resource and Protection., 6(4), 337-350.
528 https://doi.org/10.4236/jwarp.2014.64036

529 Ryu, J.S., Kim, M.S., Cha, K.J., et al. (2002). Kriging interpolation methods in geostatistics and
530 DACE model. Journal of Mechanical Science and Technology., 16 (5), 619–632.
531 https://doi.org/10.1007/BF03184811

532 Simpson, T.W., Mauery, T.M., Korte, J.J., Mistree, F. (2001). Kriging models for global
533 approximation in simulation-based multidisciplinary design optimization. AIAA J., 39 (12),
534 2233–2241. https://doi.org/10.2514/3.15017

535 Singh, R.M., Datta, B. (2006). Identification of groundwater pollution sources using GA-based
536 linked simulation optimization model. Journal of Hydrologic Engineering., 11(2), 101–109.
537 https://doi.org/10.1061/9780784413623.118

538 Sun, A.Y., Painter, S.L., Wittmeyer, G.W. (2006). A constrained robust least squares approach for
539 contaminant release history identification". Water Resources Research., 42(4),263–269.
540 https://doi.org/10.1029/2005WR004312

Sidauruk, P., Cheng, H.D., Ouazar, D. (2010). Ground Water Contaminant Source and Transport Parameter Identification by Correlation Coefficient Optimization". GroundWater., 36 (2), 208–214. https://doi.org/10.1111/j.1745-6584.1998.tb01085.x

Woodbury, A.D., Ulrych, T.J. (1996). Minimum relative entropy inversion: Theory and application to recovering the release history of a groundwater contaminant. Water Resources Research., 32.9, 2671-2681. https://doi.org/10.1029/95wr03818

Xu, Teng., Gómez-Hernández, J.J. (2016). Joint identification of contaminant source location, initial release time and initial solute concentration in an aquifer via ensemble Kalman filtering. Water Resources Research. https://doi.org/10.1002/2016WR019111

Yeh, H.D., Lin, C.C., Yang, B.J. (2014). Applying hybrid heuristic approach to identify contamination source information in transient groundwater flow systems. Mathematical Problems in Engineering. 369369:1–13. https://doi.org/10.1155/2014/369369

Zwickl, D.J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.

Zhao, Y., Lu, W.X., An, Y.K. (2015). Surrogate model-based simulation-optimization approach for groundwater source identifcation problems. Environmental Forensics., 16 (3), 296–303. https://doi.org/10.1080/15275922.2015.1059908

Zhao,Y., Lu, W.X., Xiao, N.C. (2016). A Kriging surrogate model coupled in simulation–optimization approach for identifying release history of groundwater sources. Journal of Contaminant Hydrology., 185–186, 51-60. https://doi.org/10.1016/j.jconhyd.2016.01.004

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Data.xlsx