Scientific Research Publishing

Journal of Water Resource and Protection, 2017, 9, 1612-1633
http://www.scirp.org/journal/jwarp
ISSN Online: 1945-3108
ISSN Print: 1945-3094

Evaluation of Unknown Groundwater Contaminant Sources Characterization Efficiency under Hydrogeologic Uncertainty in an Experimental Aquifer Site by Utilizing Surrogate Models

Shahrbanoo Hazrati-Yadkoori[1*], Bithin Datta[1,2]

[1]Discipline of Civil Engineering, College of Science and Engineering, James Cook University, Townsville, Australia
[2]CRC for Contamination Assessment and Remediation of the Environment, CRC CARE, University of Newcastle, Callaghan, Australia
Email: *shahrbanoo.hazratiyadkoori@my.jcu.edu.au, bithin.datta@jcu.edu.au

How to cite this paper: Hazrati-Yadkoori, S. and Datta, B. (2017) Evaluation of Unknown Groundwater Contaminant Sources Characterization Efficiency under Hydrogeologic Uncertainty in an Experimental Aquifer Site by Utilizing Surrogate Models. *Journal of Water Resource and Protection*, **9**, 1612-1633.
https://doi.org/10.4236/jwarp.2017.913101

Received: November 29, 2017
Accepted: December 25, 2017
Published: December 28, 2017

Abstract

Characterization of unknown groundwater contaminant sources is an important but difficult step in effective groundwater management. The difficulties arise mainly due to the time of contaminant detection which usually happens a long time after the start of contaminant source(s) activities. Usually, limited information is available which also can be erroneous. This study utilizes Self-Organizing Map (SOM) and Gaussian Process Regression (GPR) algorithms to develop surrogate models that can approximate the complex flow and transport processes in a contaminated aquifer. The important feature of these developed surrogate models is that unlike the previous methods, they can be applied independently of any linked optimization model solution for characterizing of unknown groundwater contaminant sources. The performance of the developed surrogate models is evaluated for source characterization in an experimental contaminated aquifer site within the heterogeneous sand aquifer, located at the Botany Basin, New South Wales, Australia. In this study, the measured contaminant concentrations and hydraulic conductivity values are assumed to contain random errors. Simulated responses of the aquifer to randomly specified contamination stresses as simulated by using a three-dimensional numerical simulation model are utilized for initial training of the surrogate models. The performance evaluation results obtained by using different surrogate models are also compared. The evaluation results demonstrate the different capabilities of the developed surrogate models.

These capabilities lead to development of an efficient methodology for source characterization based on utilizing the trained and tested surrogate models in an inverse mode. The obtained results are satisfactory and show the potential applicability of the SOM and GPR-based surrogate models for unknown groundwater contaminant source characterization in an inverse mode.

Keywords

Surrogate Models, Unknown Groundwater Contamination Sources, Source Characterization, Experimental Site, Contaminated Aquifers

1. Introduction

Groundwater is a valuable natural resource and its consumption has increased over the years. As a result, the environmental problems associated with groundwater have increased due to widespread improper and unplanned groundwater management worldwide. Groundwater contamination in an aquifer becomes more difficult to remedy as the contamination spreads. The challenge arises due to insufficient information regarding the contaminated aquifers and especially, often lack of knowledge regarding the sources of contamination and its history of activity. Usually, the contaminations are accidentally detected long time after the first contaminant source activities started. As a result, limited and sparse data are available and generally several contaminant sources are considered as the potential contaminant sources. Therefore, developing an efficient methodology for source characterization is essential.

The most frequently applied methodology for source characterization is linked simulation-optimization approach. This approach consists of numerical simulation models and optimization models, with the linked simulation model embedded or implicitly embedded within the optimization model [1]. The main drawback of this approach is that its applications in real-world cases are computationally very intensive. To overcome this drawback, simulation models are replaced by surrogate models to develop Surrogate Models based Optimization (SMO). In the SMO, the optimization model instead of linking to a complex and time-consuming simulation model is linked to a simpler and faster surrogate model. This surrogate model can efficiently decrease the computational time once the surrogate models are developed after training and testing. Two different algorithms with different capabilities for comparison purpose are utilized to develop surrogate models in this study. Self-Organizing Maps (SOM) and Gaussian Process Regression (GPR) are utilized in this study to develop two types of surrogate models. The SOM algorithm is selected as the surrogate model type because of its capabilities in classifying nonlinear multidimensional data. The GPR algorithm is also utilized as the other surrogate model type because it can reveal the relationships of high dimensional data. The performance evaluation results of the developed surrogate models for source characterization are com-

pared. The developed surrogate models utilizing SOM and GPR can approximate the groundwater flow and transport simulation models. These surrogate models are also applicable for source characterization. Comparisons of evaluation results of these surrogate models do not show a significant difference in terms of accuracy for source characterization. However, the SOM-based surrogate model could identify inactive contaminant sources more precisely than the GPR-based Surrogate model.

The methodologies proposed earlier for unknown groundwater source characterization can be subdivided into two main categories. 1. Methodologies based on statistical and deterministic approaches which mainly solved this problem in an inverse mode. 2. The approaches based optimization algorithm which integrate the groundwater flow and transport simulation models with an optimization algorithm [2]. In the first group, some of the proposed methodologies are random walk particle method [3], the Minimum Relative Entropy (MRE) [4], Tikhonov Regularization (TR) [5], the Backward Beam Equation (BBE) [6]. These developed methodologies applied mostly to characterize one or two-dimensional homogeneous contaminated study areas. In these studies, usually, the contaminant source(s) also considered being a single contaminant source. Their evaluation results also demonstrated that the applied methodologies can be effective in the presence of sufficient and accurate measurements data.

In the second group, consisting of the embedding technique, response matrix and linked simulation-optimization approaches were utilized to incorporate simulation models with optimization models [7]. For example, the embedded technique was used in [8]. They utilized least square regression and linear programming technique which each of them combined with groundwater solute transport simulation models for source identification. A nonlinear optimization model incorporating finite difference discretized governing equations of groundwater flow and transport for unknown contamination source characterization was utilized in [1] [9] and [10]. The embedded techniques have some limitations. For instance, for obtaining the optimal solutions, repeated solutions of the set of discretized groundwater and transport governing equations are required. As a result, these procedures are computationally intensive and inefficient. The main disadvantages of the response matrix approach are that the approach needs relatively large information of the aquifer system and the aquifer responses are generally assumed linear. The approach is also highly sensitive to measurement errors [1] [7] and [11]. The linked simulation-optimization approach which is the most commonly used approach is externally linked the groundwater flow and transport simulation models with an optimization model. Some of the prominent techniques which were utilized in this procedure are: Genetic Algorithm (GA) [12] [13], the Artificial Neural Network (ANN) [14] [15], Simulated Annealing (SA) [16] [17] [18] and [19] and Adaptive Simulated Annealing (ASA) [20], ASA in conjunction with uncertainty modelling [21]. The main advantages of the linked simulation-optimization approach compared to

the other ones are: 1. in this approach, some complex groundwater flow and transport simulation models such as MODFLOW and MT3DMS can be used, and 2. the number of decision variables of the optimization model can be decreased in this approach by eliminating the embedded equations as binding constraints [1], so the solutions can be easier and less intensive in terms of feasibility. However, the main disadvantage of the developed linked simulation-optimization approaches is their computational times which are very high. For example, for solving a real-world case, they may need several days of the iterative solution.

In this study, collected field data from an experimental aquifer site located in the Botany Basin aquifer, New South Wales, Australia are used to evaluate the performance of the developed methodology. The hydrogeologic characteristics of this experimental site are investigated through a few tests [22]. As a result, some measured values for hydraulic conductivity and contaminant concentrations are available. The performance evaluation results at this experimental site demonstrate the potential applicability of the developed surrogate models for source characterization in terms of contaminant source location, magnitudes, and release history.

2. Methodology

2.1. Groundwater Flow and Transport Simulation Models

The MODFLOW [23] and MT3DMS [24] are groundwater flow and transport numerical simulation codes utilized in this study. MODFLOW [23] which is a finite-difference based groundwater flow model is utilized for numerical flow simulation. The general governing equation of the groundwater flow through porous media is described by Equation (1) [23].

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_{zz}\frac{\partial h}{\partial z}\right)\pm W=S_s\frac{\partial h}{\partial t} \quad (1)$$

where, K_{xx}, K_{yy} and K_{zz} are the hydraulic conductivity values along the *x, y*, and *z* coordinate axes (*L/T*), *h* is the potentiometric head (*L*), S_S is the specific storage of the porous media (L^{-1}), *t* is time (*T*) and *W* is a volumetric flux per unit volume from aquifer as sources (sinks); the negative value represents withdrawal of the groundwater system and vice versa (T^{-1}).

The MT3DMS [24] is the numerical mass transport simulation model used in this study. This model has the capability of simulating the advection, dispersion, and chemical reaction processes of the groundwater contaminants transport. The governing equation of this model is described in Equation (2) [24]:

$$\frac{\partial\left(\theta C^k\right)}{\partial t}=\frac{\partial}{\partial x_j}\left(\theta D_{ij}\frac{\partial C^k}{\partial x_j}\right)-\frac{\partial}{\partial x_i}\left(\theta v_i C^k\right)+q_s C_s^k+\sum R_n \quad (2)$$

where, θ is the subsurface porous media porosity (dimensionless), C^k is the dissolved concentration of species *k* (ML^{-3}), *t* is time (*T*), x_i, x_j represents the distances along the Cartesian coordinate axis (*L*), D_{ij} is the hydrodynamic

dispersion coefficient tensor (L^2T^{-1}), v_i represents the seepage velocity (LT^{-1}); it is related to the Darcy flux through the relationship; $v_i = \frac{q_i}{\theta}$, q_s is volumetric flow rate per unit volume of the groundwater system which represents fluid source (positive) and sinks (negative) (T^{-1}), C_s^k is the concentration of the source or sink flux for species k (ML^{-3}); and $\sum R_n$ is the chemical reaction term ($ML^{-3}T^{-1}$).

2.2. Developing Surrogate Models for Source Characterization

Generally, implementation of the simulation models for real-world cases is complex and extensively time-consuming. Therefore, to decrease the high computational cost of the complex simulation models, these computationally intensive simulation models have been replaced by response surface methodologies. It is supposed that by accurately constructing these models, the behavior of more sophisticated simulation models can be approximately emulated with much reduced computational time [25]. Several types of surrogate models have been developed based on Kriging, ANN, MARS and Gaussian Process (GP) as approximate simulators of the physical processes [26]. Surrogate Models based Optimization (SMO) is one of the popular surrogate models which has been suggested to reduce computational burden. This approach replaces the computationally intensive simulation models with a cheaper to run trained surrogate model. Therefore, for obtaining global optimal solution there is no need to run the computational intensive simulation models tens of thousands times [27].

In this study, for characterization of unknown groundwater contaminant sources, Self-Organizing Map (SOM) and Gaussian Process Regression (GPR) algorithms for comparison purpose are used to construct the surrogate models (**Figure 1**). These models mimic the behavior of the groundwater flow and transport simulation models, MODFLOW and MT3DMS, respectively. Also, the developed surrogate models are applied to characterize unknown groundwater sources in terms of contaminant source locations, magnitudes and activity times. The main steps involved to develop a surrogate model for characterization of unknown groundwater contaminant sources are presented in **Figure 1**. These steps are also explained as follows:

1) Problem definition and sampling plan: this stage is a crucial stage and has essential effects on the accuracy of results. First, the problem and the most important variables of the system which are highly dependent on the complexity of origin system are defined. These variables are constituted of known variables and decision variables. Then, for generating qualified sampling points for training and testing surrogate models a suitable random generating methodology need to be selected and utilized. In this study, Latin Hypercube Sampling (LHS) is utilized to generate the training and testing sample data. For source identification problem, LHS is used to generate adequate random contaminant source fluxes. It is also suggested that the sampling size be 15 - 20 times of the dimensions of the problem [27].

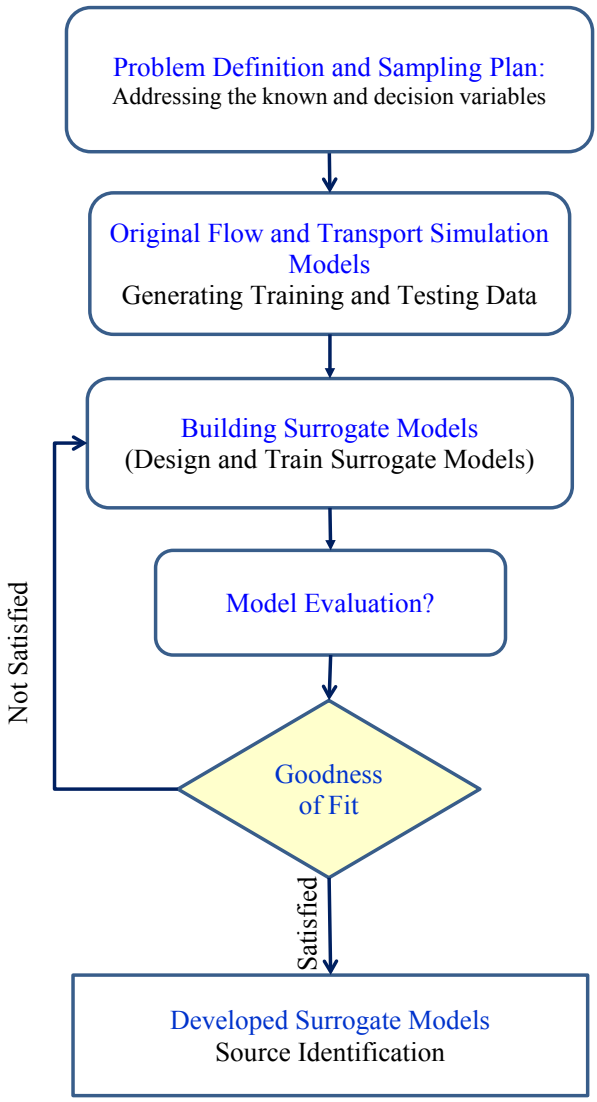

Figure 1. Flowchart of the main steps for developing surrogate models for characterization of unknown groundwater contaminant Sources.

2) Solving the simulation models: at this stage, the flow and groundwater simulation models for the contaminated aquifer site are solved. These models are solved to randomly generated contaminant source fluxes at stage 1. As a result, the contaminant concentration values are obtained as the solution of the groundwater flow and transport simulation models.

3) Solving the simulation models: at this stage, the flow and groundwater simulation models for the contaminated aquifer site are solved. These models are solved to randomly generated contaminant source fluxes at stage 1. As a result, the contaminant concentration values are obtained as the solution of the groundwater flow and transport simulation models.

4) Building surrogate models: in this stage, at least one important question should be addressed, the tool(s) which are to be used for constructing the surro-

gate model(s) [28]. In this study, SOM and GPR are selected as the surrogate models' types. The design or architecture of the surrogate model also can be addressed at this stage.

5) Model evaluation: in this stage, the performances of the developed surrogate models are evaluated by using a new sample data set which are independent of the training data. The model evaluation results can be used to change the surrogate model types or designs.

6) Source characterization solution/step 3: if the goodness of fit is achieved, source characterization results are obtained and stop. Otherwise, go to step 3.

2.3. Self-Organizing Map

The Self-Organizing Map (SOM) is an unsupervised learning method that was introduced by T. Kohonen in 1982 [29]. This algorithm is widely utilized to visualize and cluster multidimensional data due to its easy implementation. The main features of this algorithm are its capability of transforming complex non-linear high-dimensional data space into a simple geometric relationship (**Figure 2(a)**).Usually, SOM results are represented in two dimensions by preserving the topological structure of the input data [29] and [30]. The main processes of SOM algorithm can be summarized as initialization, competition, cooperation, and adaptation which are briefly described at below. The more detailed information of this algorithm can be found in [31].

1) Initialization: a group of high-dimensional inputs data is quantized by a few weight vectors to a discrete space usually two-dimensional grid [32] and [33]. If X is an m-dimensional continues input data pattern $\{X = x_1, x_2, \cdots, x_m\}$, these data are mapped to output neurons which usually is a two-dimensional discrete space by the weight matrix $\{W = w_{j1}, w_{j2}, \cdots, w_{jm}\}$, where m is the size of the input data and $j = 1, \cdots, n$, which n defines the number of the output space neurons.

2) Competition: for each random sample of input space, the output neurons

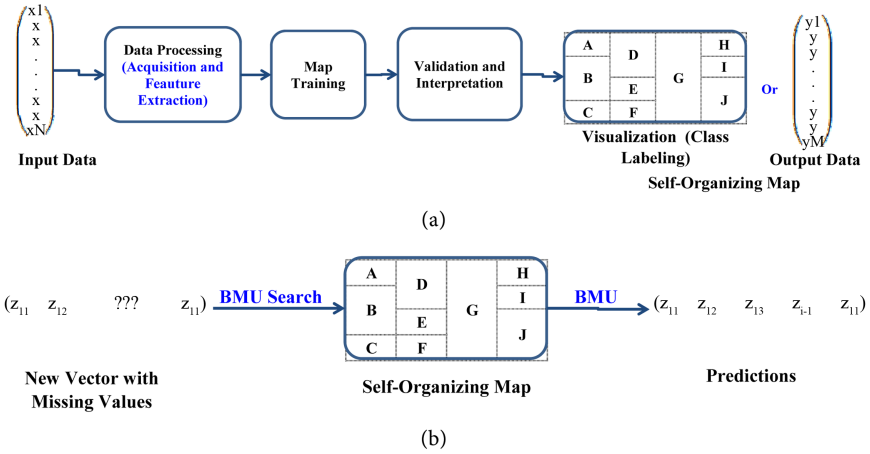

Figure 2. (a) SOM algorithm's stages in clustering and visualization process; (b) prediction of missing components of new input vector [31].

compete to be the winner neuron. The winning neuron which has the most similarity to the input data is called Best Matching Unit (BMU). The distance between the random sample of input space and all weight vectors are calculated by using Equation (3) or Euclidian distance measure.

$$d_j(x) = \sum_{i=1}^{m} \left(x_i - w_{ji}\right)^2, \quad \forall i = 1, \cdots, m \tag{3}$$

BMU command in SOM algorithm by searching to find the most similar output neuron to the input vector can be used for finding missing values of an input vector (**Figure 2(b)**). This command in this study is used to characterize unknown groundwater contaminant sources [31].

3) Cooperation: once the winner neuron is obtained, the weight vector of the winning neuron and all other neurons are updated according to Equation (4) to minimize the local error [32].

$$W_{ji} = w_{ji}(t) + \eta(t) K(j,t) \left[X_i - W_{j,i}(t)\right] \tag{4}$$

where $\eta(t)$: is the learning rate at iteration t; and $K(j,t)$ is a suitable neighborhood function. This neighborhood function has the responsibility of preserving topological of input data [32].

4) Adaptation: The weight adjusting is repeated until a stable map is obtained or the map is converged [33].

Moreover, SOM Map quality could be assessed by various methods. In this study, Quantization Error (QE) which is a widely used criterion for evaluation of SOM Maps is utilized. The QE gradually decreases with increasing map sizes. The earlier studies indicate that the suitable number of neurons have an essential role in the accuracy and performance of the SOM algorithm [30]. The "SOM Toolbox for Matlab 5" is used in this study for constructing the SOM-based surrogate model [34].

2.4. Gaussian Process Regression

Gaussian Process Regression (GPR) models are nonparametric kernel-based probabilistic models. These models are flexible nonlinear interpolating techniques which are based on the training data [35]. This technique is capable of exploring unknown functions of multi-dimensional data which maps inputs data to output data (explore their interactions) [36]. This technique is capable of approximating any multi-dimensional data [37]. These capabilities make GPR a popular and widely used surrogate models technique. The GPR models are defined by two functions: mean function $m\left(\vec{X}\right)$ and covariance function $k\left(\vec{X}, \vec{X}'\right)$, these functions can be described as [27]:

$$m\left(\vec{X}\right) = E\left[f\left(\vec{X}\right)\right] \tag{5}$$

$$k\left(\vec{X}, \vec{X}'\right) = E\left[\left(f\left(\vec{X}\right) - m\left(\vec{X}\right)\right)\left(f\left(\vec{X}'\right) - m\left(\vec{X}'\right)\right)\right] \tag{6}$$

The mean function represents the expected function value for input X [37]. The covariance function models the interactions between the function values at

different inputs points $X$and X' [36]. And a GP model can be written as [27]:

$$f\left(\vec{X}\right) \sim GP\left(m\left(\vec{X}\right), k = \left(\vec{X}, \vec{X}'\right)\right) \tag{7}$$

3. Performance Evaluations of the Developed Surrogate Models

3.1. Site Description, Eastlake Experimental Site

The performance of the developed methodology is evaluated by using the data from an experimental site. A natural gradient tracer experiment carried out at the Eastlakes Experimental Site, located at the Botany Basin, New South Wales, Australia [38]. This site is located in the upper part of the Botany Sands aquifer next to pond 5 of Lachlan ponds in an area about 80 m^2 [22] and [39]. **Figure 3** illustrates this site in the Botany Sands aquifer. Although this aquifer is a homogeneous and isotropic on a macroscopic scale, it is heterogeneous and anisotropic on a microscopic scale [38]. This site was founded in 1992 for research studies at the University of New South Wales (UNSW) Groundwater Center [22].

Figure 4 shows the most important features of this study area. A network of 49 piezometers was installed on a 7 × 11 m in this part of the aquifer [22] [38]. These piezometers penetrated up to 6 m into the underlying sediments to investigate geological and hydrogeological characteristics of this experimental site. This experimental site is known as East Lake Experimental Site (ELE Site) [22].

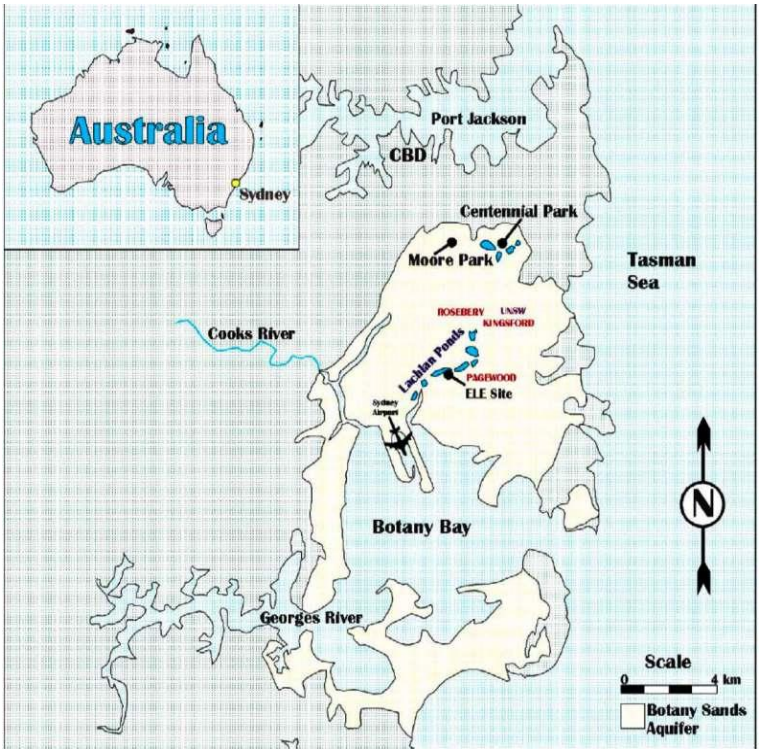

Figure 3. The east lake experimental site location (ELE site) at the Botany Sands aquifer [22].

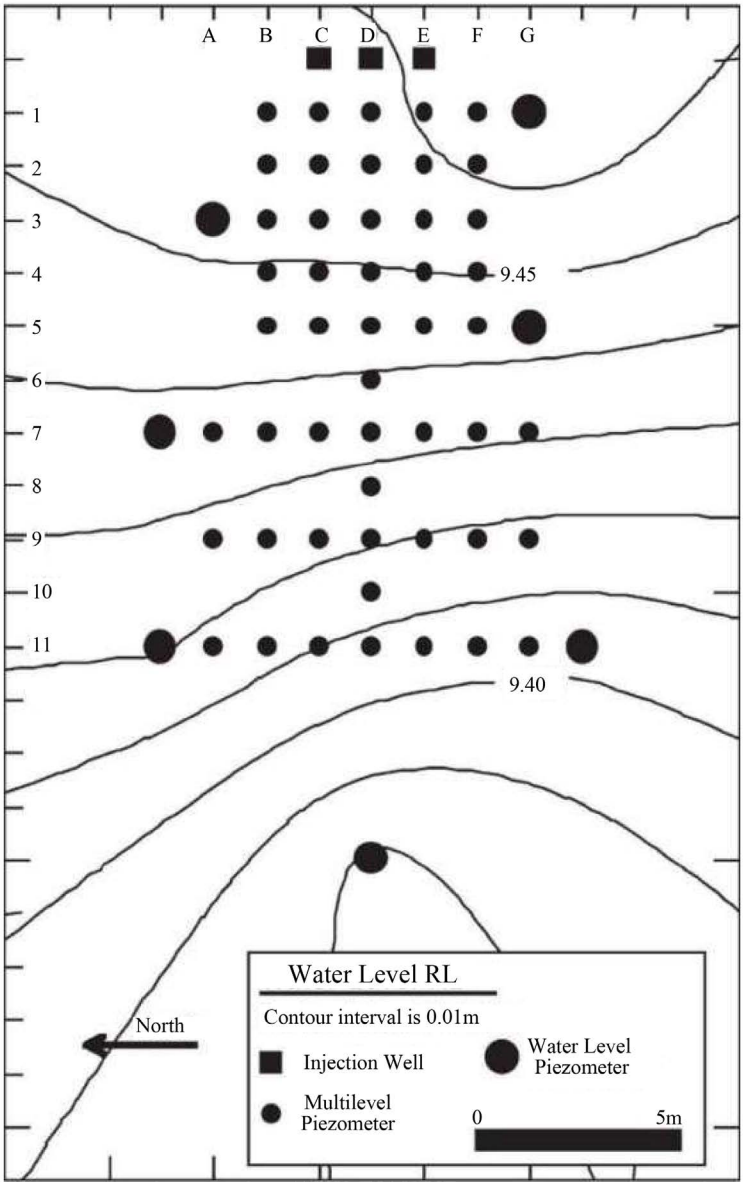

Figure 4. Layout of the ELE site showing injection well locations, multilevel piezometer and water level piezometer [39].

The dimension and characteristic values of the study area are presented in **Table 1** [39]. This information is obtained from the previous studies reports at this experimental site [22] and [38]. In this study area, the east and west boundaries are considered as specified head boundaries, due to the location of this site on the side of the pond 5 of Lachlan ponds that provides hydraulic continuity with the pond (**Figure 3**). The north and south boundaries are variable heads. The initially specified head distributions are based on the specified contours in **Figure 4**. Rainfall is the main source of recharge for the Botany Sands aquifer.

According to the results of previous geological investigations, the experimental site consists of five sedimentological distinct layers (**Figure 5**): 1. Medium sand with silt/clay content of up to 5%; 2. Waterloo Rock; 3. Organic silty sand;

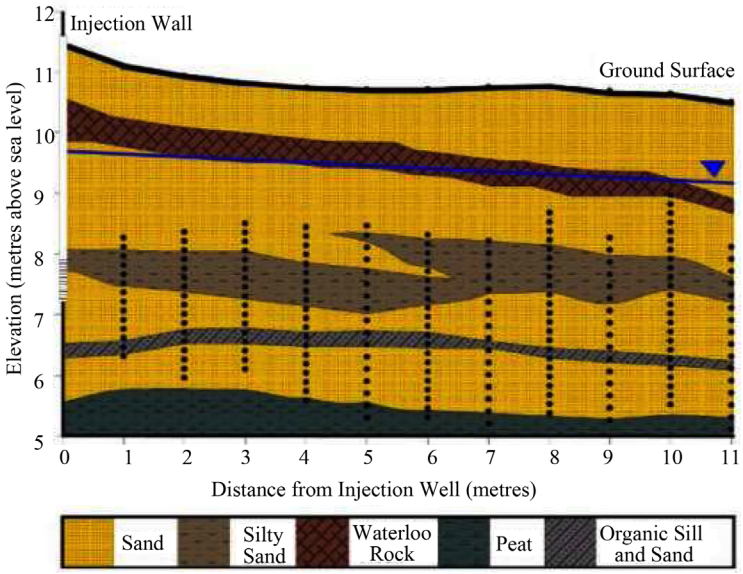

Figure 5. Geological cross section of the ELE site along the line D [22].

Table 1. Hydrogeological information of the study area.

Parameter	Unit	Value
Maximum length of study area	m	15.00
Maximum width of study area	m	13.00
Thickness of study area	m	3.50
Grid spacing in x-direction	m	1.00
Grid spacing in y-direction	m	1.00
Porosity (layer 1, layer 2, layer 3 and layer 4)	Dimensionless	(0.39, 0.41, 0.36 and 0.41)
Longitudinal dispersivity (all layers)	m	0.03
Ratio: H/L dispersivity	Dimensionless	0.10
Specific storage (all layers)	1/m	0.20
Specific Yield (all layers)	Dimensionless	0.20
Recharge	m/day	0.00
Flow rate in injection wells	m^3/day	4.40
Initial contaminant release concentrations	mg/l	0 - 300

4. Peat material; and 5. Silty/clay sand unit [38]. According to this information, for simulation of the study area from 6 meters above sea level to the groundwater level; the study area can be divided into four distinct layers. The thickness of the top layer which extends from the top of the silty sand layer to the groundwater level is 1.5 meters. This layer is mainly comprised of sand. The second layer has 0.4 meters depth and it is mainly comprised of silty sand. The third layer with injection wells located in has 0.6 meters depth. This layer is mainly comprised of sand. The thickness of the bottom layer is one meter and it is situated on the top of peat layer [39].

In the tests carried out in the ELE site, the injected tracer solutions included conservative and reactive inorganic elements such as bromide, calcium, lead, and potassium. Three injection wells, C, D, and E were used in this test. These wells are illustrated in **Figure 4**. The tracer test was conducted by preparing 300 liters of a solution that included boron, bromide, chloride, and lithium as the conservative tracers and six reactive solutes. The concentrations of conservative tracers needed to be three to four times higher than background concentrations to be properly monitored. To analyze the background chemical concentrations of tested elements, 88 groundwater samples were collected. The analysis results indicated that all of the tested elements concentrations were below the analytical detection limit [22]. The detection limit concentration for bromide was 1.8 mg/l [22]. In this study, bromide is considered as a conservative contaminant. The concentration of bromide in the test was 186 mg/l. The containers of tracer solution were injected over 30 minute's period from 13:00 to 13:30 on 2nd July 1996. During the tracer injection, the flow rates of wells were kept low enough to avoid the significant increases in the hydraulic heads at the injection wells [22]. For this reason, in this study, the flow rate of additional potential contaminant source was considered 1 m^3/day to prevent a significant change of the flow system and hydraulic head distribution [39].

The first samples of contaminant concentrations were collected two days after the injection on 4th July 1996. Gathering samplings repeated by nine more sessions 4, 6, 8, 12, 16, 20, 24, 28 and 32 days after injection. Monitoring the transport of tracers plume movements demonstrated that bromide and the other conservative elements transports are mainly controlled by the variability of aquifer's hydraulic conductivity [22]. According to the previous studies at ELE Site, monitoring values until 16 days after the injection showed no noticeable chemical transport processes to effect on the natural tracer behaviors [22]. In other words, advection and dispersion were the dominant physical processes of the bromide tracer transport during the monitoring time. Therefore, the total time of simulation is divided into five different stress periods. The first stress period is the only active stress period and its duration is 30 minutes. The second to fourth stress periods are each of two days duration and the last stress period is of eight days duration. In this study, the monitored contaminant concentrations at nine monitoring locations and totaling to 10 values, and belong to stress periods two to five are utilized as described in **Table 2**. In this study, in addition to the three injection sources, one more potential location is considered as a possible contaminant sources location to evaluate the performance of the developed methodology for identifying unknown contaminant sources in terms of locations and magnitudes.

The hydraulic conductivity values for ELE site were estimated by applying a combination of constant head test and falling head tests [22]. A total of 522 hydraulic conductivity values along the three lines shown C, D, and E are available. The distributions of hydraulic conductivity showed considerable variations from

Table 2. The monitoring locations and observed concentration values.

ID	Monitoring locations (i, j, k)*	Stress Period	Contaminant concentration values (mg/l)
1	M1 (7, 3, 3)		12.20
2	M2 (6, 3, 3)	2	15.50
3	M3 (5, 3, 3)		0.10
4	M4 (8, 3, 3)	3	9.00
5	M2 (6, 3, 3)		19.00
6	M5 (5, 4, 3)	4	0.09
7	M6 (6, 5, 3)		0.09
8	M7 (8, 4, 3)		0.15
9	M8 (6, 4, 3)	5	13.30
10	M9 (7, 6, 3)		0.11

*: (i,j,k) the nodes coordinates in X, Y and Z directions, respectively.

1.8 to 50 m/day. Sometimes these variations are observed in short distances [22] [38]. According to the results of previous studies, the mean hydraulic conductivity value for Botany Sands aquifer is likely around 20 m/day [22]. The simulation of groundwater flow and transport of ELE site needs the hydraulic conductivity values be known throughout the entire study area. Therefore, due to unavailability of the hydraulic conductivity values at all discretization nodes; 240 hydraulic conductivity values (some of these are multiple measurements within the same layer) are used to generate interpolated hydraulic conductivity values for all nodes of the study area. In this study, the Inverse Distance Weighting (IDW) methodology is utilized to interpolate hydraulic conductivity values for the entire study area because of its simplicity, and efficiency [40]. The 240 hydraulic conductivity values are used in three different iterations. As mentioned earlier, in some cases, for a certain location different measured hydraulic conductivity values are available. Therefore, IDW was utilized to interpolate hydraulic conductivity values throughout the whole study area in three different iterations. Then, the average values of these three iterations for all the nodes of the study area are utilized as the inputs of simulation models. **Figure 6** represents the generated hydraulic conductivity values for layer three of the EES aquifer utilizing IWD interpolation method.

3.2. Results and Discussion

In this section, first, the following steps for constructing surrogate models for source characterization are explained. Then, the performance evaluation results of the constructed models are discussed.

1) Problem definition and sampling plan: as previously mentioned, four potential contaminant sources are considered in this study. These four sources are included three injection wells (**Figure 4**) and one dummy source with (10, 2, and 3) coordinates along XYZ directions, respectively. The LHS is used to randomly

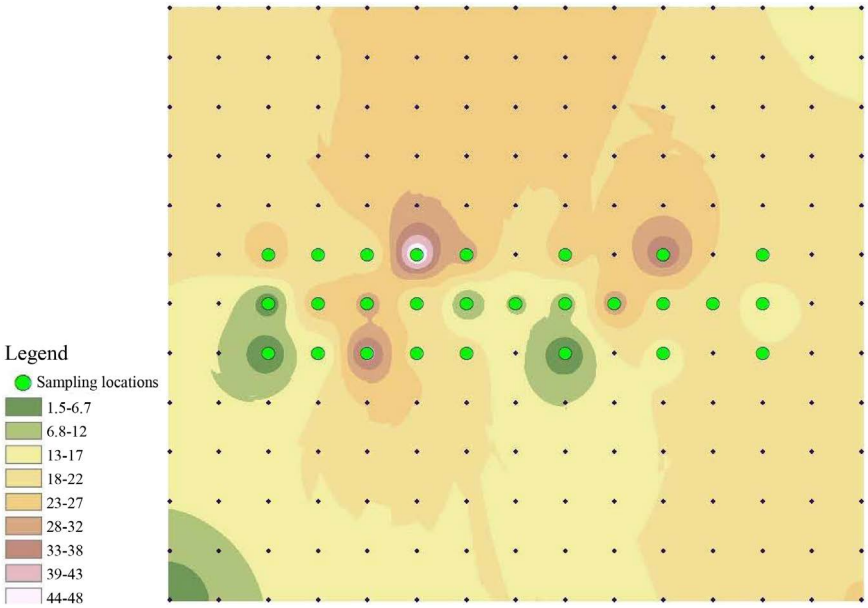

Figure 6. Generated hydraulic conductivity for layer three-iteration two; applying the IDW interpolation algorithm (m/day).

generate 1000 initial sample sets. These sample sets consist of contaminant release concentrations at four potential contaminant sources. The contaminant release concentrations are assumed to be in the range of 0 - 300 mg/l for all potential contaminant sources. The contaminant concentration values at specified locations and times (Table 2) are considered as the rest of important variables of the explained system.

2) Solving the numerical simulation models: The numerical flow and transport simulation models, MODFLOW and MT3DMS, respectively (within GMS 7) are solved for randomly generated contaminant release concentrations at the previous stage. The solutions contained the corresponding contaminant concentration magnitudes at selected monitoring locations at specific stress periods (Table 2).

3) Developing the surrogate models: in this step, SOM and GPR algorithms are utilized to develop surrogate models for source characterization.

Table 3 represents a typical set of inputs for training the surrogate models. This input set consists of five sample sets. Each set consists of randomly generated contaminant release concentration values at potential contaminant sources at first stress period (SP1) and corresponding contaminant concentration magnitudes at nine monitoring locations (M1 to M9) at four stress periods (SP2 to SP5). It is supposed that if the surrogate models are constructed accurately, these models could properly approximate the groundwater flow and transport simulation models.

Same sets of training data are used for constructing the SOM and GPR-based surrogate models. Due to the different natures of the applied tools, different approaches are utilized to design the training data for developing the surrogate

Table 3. A typical input for training a surrogate model.

ID	Source 1	Source 2	Source 3	Source 4	M1	M2	M3	M4	M2	M5	M6	M7	M8	M9
	Contaminant release concentrations (mg/l)				Contaminant concentrations (mg/l)									
	SP1				SP2		SP3		SP4				SP5	
1	290	251	8	146	13.3	36.0	5.7	0.3	55.0	2.7	0.5	0.0	15.8	0.0
2	163	216	245	157	18.3	14.9	3.7	5.4	26.1	1.2	0.1	0.2	12.4	0.0
3	289	0	5	59	0.1	24.9	3.5	0.3	42.3	0.5	0.2	0.0	15.6	0.0
4	16	159	102	269	13.2	1.5	0.4	0.2	3.5	0.1	0.0	0.1	0.3	0.0
5	55	298	52	84	16.8	6.7	0.0	1.6	9.2	0.0	0.1	0.1	1.5	0.0

models. In the SOM-based surrogate models, all the training data is used to develop the surrogate models in one shot or in a single run. At this stage, different SOM-based surrogate models representing different numbers of SOM map units are constructed.

For training and developing GPR-based surrogate models, first, the predictors and target variables of the system need to be addressed. Since, in source characterization problem, just observed contaminant concentrations data is available, unknown groundwater contaminant sources need to be characterized in an inverse mode. Therefore, in the training process of the GPR-based surrogate model, the contaminant concentration values of the training data are addressed as the predictors of the GPR prediction models. The randomly generated contaminant release concentrations at potential contaminant sources at specific times are considered to be the target variables of the GPR prediction models. Each GPR prediction model can only have one target variable. As a result, for each target variable, separate GPR model is developed. Then, after developing all the GPR prediction models, the constructed GPR prediction models are integrated to develop the GPR-based surrogate model. By providing the measured or simulated contaminant concentration values for the GPR-based surrogate model, unknown contaminant sources can be characterized at potential contaminant sources at specific times.

After developing the SOM and GPR-based surrogate models, the developed surrogate models are independently utilized for unknown source characterization without using an explicit optimization model.

4) Validation of the surrogate models: the developed surrogate models are tested by new sample sets. The contaminant release concentrations of these sample sets are randomly generated by using the LHS method in the range of 0 - 300 mg/l. Then, the corresponding concentration values at monitoring locations are obtained by implementing the simulation models.

In order to evaluate the capability and efficiency of the SOM and GPR-based surrogate models to identify the unknown source characteristics, when the field concentration measurements resulting from specified contaminant release concentrations in the study area are specified, the surrogate models are used in in-

verse mode. The simulated contaminant concentration values at specific locations and time of testing data are considered to be the known variables of the system. The developed surrogate models are utilized for source characterization by using information regarding these known variables. Table 4 presents a typical input dataset with missing data for testing the surrogate models.

In the SOM-based surrogate model case, when utilized in the inverse mode, to estimate unknown contaminant sources, the BMU command of the SOM algorithm which searches for the most similar vectors of the SOM-based surrogate model to match the testing input data is utilized for source characterization. The detailed information of the application of this surrogate model was discussed in [31]. While utilizing the GPR based surrogate model, the GPR-based surrogate model acts as a prediction model. This prediction model by using simulated contaminant concentration data at specific monitoring locations and times (testing data) characterize unknown contaminant sources.

The performance of the developed surrogate models is evaluated by utilizing Normalized Absolute Error of Estimation (NAEE) as an error criterion. NAEE can be defined by Equation (8) [20] [41]:

$$\mathrm{NAEE}(\%) = \frac{\sum_{i=1}^{S}\sum_{j=1}^{N}\left|\left(q_i^j\right)_{est} - \left(q_i^j\right)_{act}\right|}{\sum_{i=1}^{S}\sum_{j=1}^{N}\left(q_i^j\right)_{act}} \times 100 \tag{8}$$

where $\left(q_i^j\right)_{act}$ and $\left(q_i^j\right)_{est}$ are the actual and estimated source fluxes at source number i in stress period j, respectively. S and N are the total number of potential contaminant sources and transport stress periods, respectively.

The performance evaluations of the different developed SOM-based surrogate models representing different numbers of SOM map units are illustrated in Figure 7. The obtained results demonstrate that the SOM-based surrogate model with 110 × 110 map units had the least quantization error while the surrogate model with 100 × 100 map units had the lowest error of estimation in terms of NAEE. Therefore, the developed SOM based surrogate model with 100 × 100 map units is considered as the selected surrogate model. This SOM based surrogate model is selected for its best accuracy of estimation.

Table 4. A typical input vector with missing data for testing the developed surrogate models.

ID	Source 1	Source 2	Source 3	Source 4	M1	M2	M3	M4	M2	M5	M6	M7	M8	M9
	Contaminant release concentrations (mg/l)				Contaminant concentrations (mg/l)									
	SP1				SP2		SP3		SP4				SP5	
1					10.1	7.6	0.7	1.7	10.7	0.2	0.0	0.4	7.4	0.0
2					2.8	5.7	0.0	0.4	11.1	0.0	0.1	0.0	3.4	0.0
3					2.9	21.9	5.4	6.2	23.8	1.6	0.2	0.1	16.5	0.0
4					13.1	21.7	0.1	3.3	29.8	0.0	0.2	0.1	18.0	0.0
5					16.7	11.7	0.1	4.0	16.1	0.0	0.1	0.1	2.5	0.0

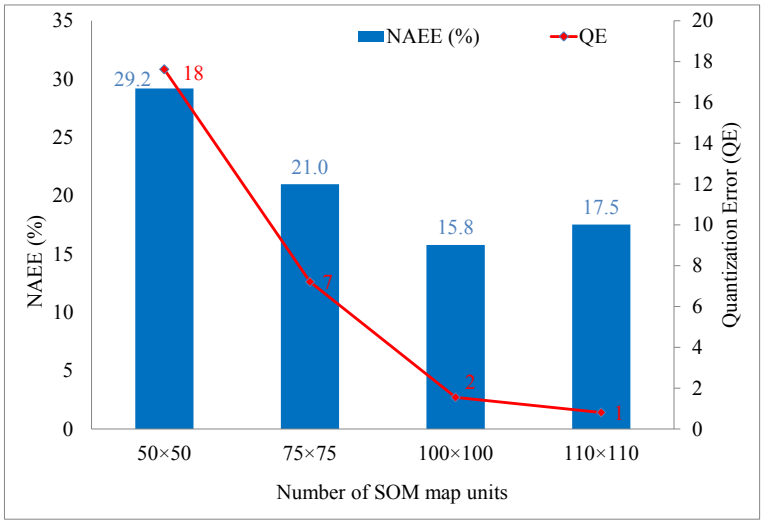

Figure 7. The performance evaluation results of the SOM-based surrogate models for different scenarios representing different numbers of SOM map units in terms of NAEE and QE values by using testing data.

The performance of the developed GPR-based surrogate model for source characterization is also evaluated by using the same testing data. The performance evaluation results of the SOM and GPR-based surrogate models for testing data in terms of NAEE are equal to 15.8% and 16.2%, respectively. The evaluation results show similar accuracy for the selected SOM-based surrogate model compared to the performance evaluation results of the GPR-based surrogate models. Despite on the average similar performance in terms of accuracy of these two surrogate models for source identifications, their abilities in screening dummy sources are different. The SOM-based surrogate model could screen the dummy sources in 98 percent of the cases accurately, against only six percent correct inference by the GPR-based surrogate model. Actually, the approximations of the GPR-based surrogate model for the dummy sources are not unsatisfactory. The developed GPR-based surrogate model could appropriately estimate the dummy sources (not actual sources) as very low magnitudes but not exactly as zero flux values.

The obtained average NAEE for each source for all the developed surrogate models are compared and presented in **Figure 8**. Although, the accuracy of the developed GPR-based surrogate model is higher than the selected SOM-based surrogate model (**Figure 8**); the capability of the SOM algorithm in clustering and subsequently in screening the dummy sources may make the SOM algorithm a potentially powerful tool for the unknown contaminant source identification problems.

2) Source characterization or recovering source injection history: The obtained results at evaluation stage demonstrate that these surrogate models can be utilized for source characterization. Therefore, the developed SOM and GPR-based surrogate models by using the measured bromide concertation data (**Table 2**) are utilized to recover source injection history from the ELE site. The

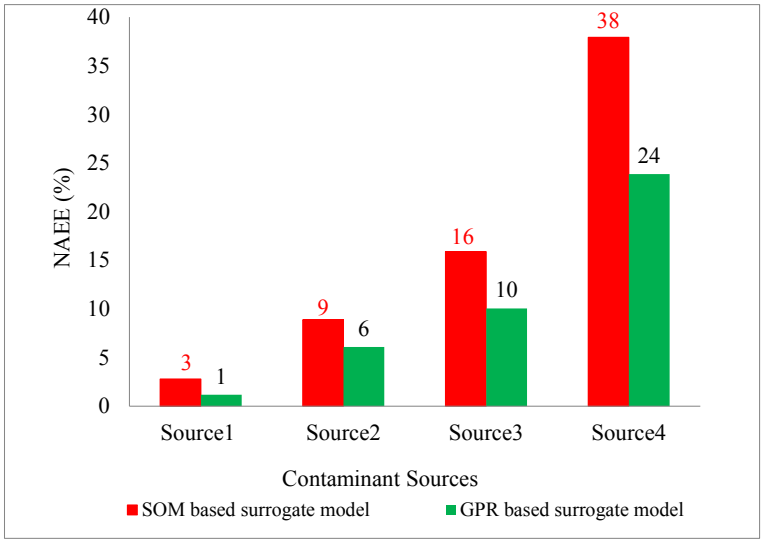

Figure 8. The average obtained results of the developed surrogate models for source characterization of the testing sample setsin terms of NAEE.

results are illustrated in **Figure 9**. The obtained results in terms of NAEE are equal to 24.9% and 24.6% for the SOM and GPR based surrogate models, respectively.

4. Conclusion

In this study, SOM and GPR algorithms for comparison purpose are used to construct the surrogate models for source characterization. Same training data is used to develop SOM and GPR-based surrogate models. Limited performance evaluations of the developed SOM and GPR-based surrogate models are conducted to test their efficiency for source characterization in an experimental contaminated aquifer site. This site constitutes of a portion of a heterogeneous aquifer with uncertainties in hydraulic conductivity values, and errors in measured contaminant concentration values. Main conclusions that can be drawn from these performance evaluation results are:

1) SOM and GPR based surrogate models are potentially effective tools to approximate the groundwater flow and transport simulation processes in a multi-layer heterogeneous experimental contaminated aquifer site.

2) The performance evaluation results demonstrate potential applicability of the SOM and GPR algorithms as the surrogate model types in inverse mode, for unknown groundwater source characterization problems under hydraulic conductivity estimation uncertainty and erroneous contaminant concentration data (**Figure 9**).

3) Comparison of the performance of the developed surrogate models for characterization of each of the potential contaminant sources (**Figure 8**) shows more accuracy for the GPR-based surrogate mode in terms of NAEE.

4) In source characterization problems, SOM algorithm capability in clustering multidimensional input data leads the SOM-based surrogate model to screen

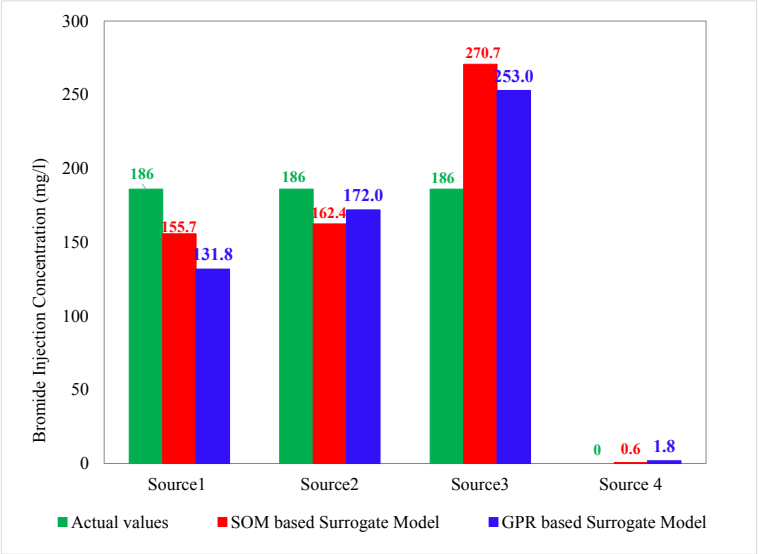

Figure 9. Comparison of the actual injected bromide concentrations with retrieved bromide injection values by utilizing the SOM and GPR-based surrogate models.

dummy sources, *i.e.*, not actual sources but included as potential sources precisely.

5) The most important conclusion is that these surrogate models may provide a feasible methodology for characterization of unknown groundwater contaminant sources in terms of location, magnitude, and duration of source activity, without the necessity of using a linked simulation-optimization model.

However, these performance evaluation results are limited to specific cases and further evaluations are necessary to establish the applicability of the developed methodology.

Acknowledgements

The second author thanks CRC-CARE, Australia for providing financial support for this research through Project No. 5.6.0.3.09/10(2.6.03), CRC-CARE-Bithin Datta which partially funded the Ph.D. scholarship of the first author.

References

[1] Mahar, P.S. and Datta, B. (2000) Identification of Pollution Sources in Transient Groundwater Systems. *Water Resources Management*, **14**, 19.

[2] Datta, B. and Kourakos, G. (2015) Preface: Optimization for Groundwater Characterization and Management. *Hydrogeology Journal*, **23**, 7. https://doi.org/10.1007/s10040-015-1297-3

[3] Bagtzoglou, A.C., Dougherty, D.E. and Tompson, A.F.B. (1992) Application of Particle Methods to Reliable Identification of Groundwater Pollution Sources. *Water Resources Management*, **6**, 9.

[4] Woodbury, A.D. and Ulrych, T.J. (1996) Minimum Relative Entropy Inversion: Theory and Application to Recovering the Release History of a Groundwater Contaminant. *Water Resources Research*, **32**, 2671-2681.

https://doi.org/10.1029/95WR03818

[5] Liu, C.X. and Ball, W.P. (1999) Application of Inverse Methods to Contaminant Source Identification from Aquitard Diffusion Profiles at Dover AFB, Delaware. *Water Resources Research*, **35**, 1975-1985. https://doi.org/10.1029/1999WR900092

[6] Atmadja, J. and Bagtzoglou, A.C. (2001) Pollution Source Identification in Heterogeneous Porous Media. *Water Resources Research*, **37**, 2113-2125. https://doi.org/10.1029/2001WR000223

[7] Amirabdollahian, M. and Datta, B. (2013) Identification of Contaminant Source Characteristics and Monitoring Network Design in Groundwater Aquifers: An Overview. *Journal of Environmental Protection*, **4**, 16. https://doi.org/10.4236/jep.2013.45A004

[8] Gorelick, S.M., Evans, B. and Remson, I. (1983) Identifying Sources of Groundwater Pollution—An Optimization Approach. *Water Resources Research*, **19**, 779-790. https://doi.org/10.1029/WR019i003p00779

[9] Mahar, P.S. and Datta, B. (1997) Optimal Monitoring Network and Ground-Water-Pollution Sources Identification. *Journal of Water Resource Planning and Management*, **123**, 199. https://doi.org/10.1061/(ASCE)0733-9496(1997)123:4(199)

[10] Mahar, P.S. and Datta, B. (2001) Optimal Identification of Ground-Water Pollution Sources and Parameter Estimation. *Journal of Water Resources Planning and Management-ASCE*, **127**, 10. https://doi.org/10.1061/(ASCE)0733-9496(2001)127:1(20)

[11] Das, A. and Datta, B. (1999) Development of Multiobjective Management Models for Coastal Aquifers. *Journal of Water Resources Planning and Management*, **125**, 12. https://doi.org/10.1061/(ASCE)0733-9496(1999)125:2(76)

[12] Aral, M.M., Guan, J.B. and Maslia, M.L. (2001) Identification of Contaminant Source Location and Release History in Aquifers. *Journal of Hydrologic Engineering*, **6**, 225-234. https://doi.org/10.1061/(ASCE)1084-0699(2001)6:3(225)

[13] Singh, R.M. and Datta, B. (2006) Identification of Groundwater Pollution Sources using GA-Based Linked Simulation Optimization Model. *Journal of Hydrologic Engineering*, **11**, 9. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:2(101)

[14] Singh, R.M., Datta, B. and Jain, A. (2004) Identification of Unknown Groundwater Pollution Sources using Artificial Neural Networks. *Journal of Water Resources Planning and Management-Asce*, **130**, 9. https://doi.org/10.1061/(ASCE)0733-9496(2004)130:6(506)

[15] Singh, R.M. and Datta, B. (2007) Artificial Neural Network Modeling for Identification of Unknown Pollution Sources in Groundwater with Partially Missing Concentration Observation Data. *Water Resources Management*, **21**, 557-572. https://doi.org/10.1007/s11269-006-9029-z

[16] Jha, M. and Datta, B. (2012) Simulated Annealing Based Simulation-Optimization Approach for Identification of Unknown Contaminant Sources in Groundwater Aquifers. 3*rd International Conference on Challenges in Environmental Science & Engineering*, Cairns.

[17] Prakash, O. and Datta, B. (2014) Characterization of Groundwater Pollution Sources with Unknown Release Time History. *Journal of Water Resource and Protection*, **6**, 14. https://doi.org/10.4236/jwarp.2014.64036

[18] Prakash, O. and Datta, B. (2014) Optimal Monitoring Network Design for Efficient Identification of Unknown Groundwater Pollution Sources. *International Journal of GEOMATE*, **6**, 785-790.

[19] Prakash, O. and Datta, B. (2015) Optimal Characterization of Pollutant Sources in

Contaminated Aquifers by Integrating Sequential-Monitoring-Network Design and Source Identification: Methodology and an Application in Australia. *Hydrogeology Journal*, **23**, 1089-1107. https://doi.org/10.1007/s10040-015-1292-8

[20] Jha, M. and Datta, B. (2013) Three-Dimensional Groundwater Contamination Source Identification using Adaptive Simulated Annealing. *Journal of Hydrologic Engineering*, **18**, 307-317. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000624

[21] Amirabdollahian, M. and Datta, B. (2014) Identification of Pollutant Source Characteristics under Uncertainty in Contaminated Water Resources Systems using Adaptive Simulated Anealing and Fuzzy Logic. *International Journal of GEOMATE*, **6**, 757-762.

[22] Beck, P.H. (2000) Transport of Conservative and Reactive Inorganic Elements in the Saturated Part of a Hetrogeneous Sand Aquifer, Botany Basin, Sydney, Australia. University of New South Wales, Sydney.

[23] Harbaugh, A.W. (2005) MODFLOW-2005, The U.S. Geological Survey Modular Ground-Water Model—The Ground-Water Flow Process. U.S. Geological Survey Techniques and Methods 6-A16.

[24] Zheng, C. and Wang, P.P. (1999) MT3DMS: A Modular Three-Dimensional Multispecies Transport Model for Simulation of Advection, Dispersion, and Chemical Reactions of Contaminants in Groundwater Systems; Documentation and User's Guide. US Army Corps of Engineers-Engineer Research and Development Center, Contract Report SERDP-99-1, 220.

[25] Gorissen, D., *et al.* (2010) A Surrogate Modeling and Adaptive Sampling Toolbox for Computer Based Design. *Journal of Machine Learning Research*, **11**, 5.

[26] Razavi, S., Tolson, B.A. and Burn, D.H. (2012) Review of Surrogate Modeling in Water Resources. *Water Resources Research*, **48**, 32. https://doi.org/10.1029/2011WR011527

[27] Wang, C., *et al.* (2014) An Evaluation of Adaptive Surrogate Modeling Based Optimization with Two Benchmark Problems. *Environmental Modelling & Software*, **60**, 167-179. https://doi.org/10.1016/j.envsoft.2014.05.026

[28] Queipo, N.V., *et al.* (2005) Surrogate-Based Analysis and Optimization. *Progress in Aerospace Sciences*, **41**, 1-28. https://doi.org/10.1016/j.paerosci.2005.02.001

[29] Kohenon, T., *et al.* (1996) Engineering Applications of the Self-Organizing Map. *IEEE*, **84**, 1358-1384. https://doi.org/10.1109/5.537105

[30] Di Mauro, M., *et al.* (2016) Design Performance Analysis of a Self-Organizing Map for Statistical Monitoring of Distribution-Free Data Streams. CIRP CMS, 6.

[31] Hazrati, Y.S. and Datta, B. (2017) Adaptive Surrogate Model Based Optimization (ASMBO) for Unknown Groundwater Contaminant Source Characterizations using Self-Organizing Maps. *Journal of Water Resource and Protection*, **9**, 23.

[32] Chalasani, R. and Principe, J.C. (2015) Self-Organizing Maps with Information Theoretic Learning. *Neurocomputing*, **147**, 12. https://doi.org/10.1016/j.neucom.2013.12.059

[33] Amauri, H., *et al.* (2015) Regional Models: A New Approach for Nonlinear System Identification via Clustering of the Self-Organizing Map. *Neurocomputing*, **147**, 16.

[34] Vesanto, J., *et al.* (2000) SOM Toolbox for Matlab 5. Report A57, 60. http://www.cis.hut.fi/projects/somtoolbox

[35] Belyaev, M., *et al.* (2016) GTApprox: Surrogate Modeling for Industrial Design. 31.

[36] Schulz, E., Speekenbrink, M. and Krause, A. (2016) A Tutorial on Gaussian Process

Regression with a Focus on Exploration-Exploitation Scenarios. 37.

[37] Retherford, J.Q. and McDonald, M. (2010) Estimation and Validation of Gaussian Process Surrogate Models for MEPDG-Based Sensitivity Analysis and Design Optimization. Vanderbilt University: Transportation Research Board Annual Meeting.

[38] Jankowski, J. and Beck, P. (2010) Aquifer Heterogeneity: Hydrogeological and Hydrochemical Properties of the Botany Sands Aquifer and Their Impact on Contaminant Transport. *Australian Journal of Earth Sciences*, **47**, 20.

[39] Amir Abdollahian, M. (2016) Development of Integrated Methodologies for Optimal Monitoring and Source Characterization in Contaminated Groundwater Systems under Uncertainty. College of Science and Engineering, James Cook University.

[40] Boman, G.K., Molz, F.J. and Guven, O. (1995) An Evaluation of Interpolation Methodologies for Generating Three-Dimensional Hydraulic Property Distribution from Measured Data. *Ground Water*, **33**, 12. https://doi.org/10.1111/j.1745-6584.1995.tb00279.x

[41] Hazrati, Y.S. and Datta, B. (2017) Self-Organizing Map Based Surrogate Models for Contaminant Source Identification under Parameter Uncertainty. *International Journal of GEOMATE*, **13**, 8.