International Journal of GEOMATE, Aug, 2017, Vol.13, Issue 36, pp.11-18
Geotec., Const. Mat. & Env., ISSN:2186-2990, Japan, DOI: http://dx.doi.org/10.21660/2017.36.2750

SELF-ORGANIZING MAP BASED SURROGATE MODELS FOR CONTAMINANT SOURCE IDENTIFICATION UNDER PARAMETER UNCERTAINTY

* Shahrbanoo Hazrati. Y [1] and Bithin Datta [1,2]

[1]College of Science and Engineering, James Cook University, Australia;
[2]CRC for Contamination Assessment and Remediation of the Environment, CRC CARE, University of Newcastle, Australia

*Corresponding Author, Received: 13 May 2016, Revised: 11 July 2016, Accepted: 17 Nov 2016

ABSTRACT: Identification of unknown groundwater contaminant sources is a complex problem. The complexities arise mainly due to the uncertainties related to the hydrogeologic information, sparsity of measurement data and unavoidable concentration measurement errors. The process of contaminant source identification with sparse and limited concentration measurement data especially when the hydrogeologic parameters are uncertain requires an efficient procedure. The existing methodologies to tackle this problem in real world cases usually require huge computational time and the solutions may be non-unique. The goal of this study is to evaluate a developed methodology to characterize the groundwater contamination sources in a heterogeneous, multi layered aquifer. This developed methodology utilizes the Self Organizing Maps (SOM) algorithm to design the surrogate models for source characterization. The most important advantages is that in this methodology, the trained SOM based surrogate models is directly utilized for groundwater contaminant source characterization without the necessity of using a separate linked simulation optimization model. The performance of the developed methodology is evaluated by using deterministic hydraulic conductivity values, and uncertain hydraulic conductivity values. These results indicate that the developed methodology could efficiently approximate groundwater flow and transport simulation models, and also characterize unknown groundwater contaminant sources in terms of location, magnitude and release history.

Keywords: Self-Organizing Maps; Surrogate Models; Groundwater Contaminant Source Identification, Hydrogeologic Uncertainty.

1. INTRODUCTION

Widespread human activities and improper management practices have caused widespread deterioration of groundwater quality worldwide, and have seriously threatened its beneficial use in recent decades. However, usually when groundwater contamination is detected after a long time, often there is not enough information regarding the characteristics of groundwater contamination sources as well as the hydrogeologic parameters of the system. On the other hand, the efficiency and reliability of contaminant source identification depends on the availability, adequacy and accuracy of hydrogeologic information and contaminant concentration measurements data. For instance, the crux of previous approaches is highly vulnerable to the accuracy and adequacy of contaminant concentration measurements and hydrogeologic data. A significant number of previously proposed approaches considered that all the hydrogeological parameter values are known. These approaches include: embedded optimization method [1], [2] and linked simulation optimization method which is the most effective approach to contaminant source identification. In linked simulation optimization approach different optimization algorithms were utilized such as Genetic Algorithm (GA) [3], [4], Simulated Annealing (SA) [5] and Adaptive Simulated Annealing (ASA) [4], [6]. Only a few of previously developed methodologies such as [4], [6] were evaluated under uncertain hydrogeological parameter conditions.

In this study, to characterize the unknown characteristics of contaminant sources a new approach is developed and evaluated for potential applicability in practical scenarios. In this new methodology, a trained Self Organizing Map (SOM) based surrogate model for source characterization approximates the flow and transport simulation models as well as an optimization algorithm. In other words, this model independently provides a procedure to characterize 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, in this methodology and other methods the accurate

analysis of the process of groundwater flow and transport requires accurate and adequate information of hydrogeologic parameters and contaminant concentration values. On the other hand, the simulation of groundwater flow and solute transport involves intrinsic uncertainties due to the sparsity or lack of enough hydrogeologic information of the porous medium. For example, hydraulic conductivity plays a main role in the process of groundwater flow and transport and this parameter may be the most uncertain parameter in the groundwater flow and transport models. It is not possible to measure this parameter in every location or discretization nodes, where the ground water flow and transport simulation model needs hydraulic conductivity values. Generally, in real world cases limited numbers of measured hydraulic conductivity are available and the values of this parameter for other locations are subject to uncertainty and these values need to be estimated.

Therefore, utilizing a proper method to estimate the unknown hydrogeologic parameters based on limited available data is essential in any contaminant source identification strategy. If these estimations do not approximate the hydrogeologic parameters accurately, it will result in the propagation of more errors and uncertainty in the groundwater flow and transport simulation models. Thus, the specific, main objective of this study is to develop an efficient methodology to characterize unknown contaminant sources. Also, this developed methodology is evaluated especially where contaminant concentration measurement data are missing for long time intervals, and hydraulic conductivity is only known at limited sample points. These evaluation results demonstrate the potential applicability of this methodology to contaminant source identification in real world cases.

2. METHODOLOGY

2.1 Surrogate Model

Surrogate models or Response Surface Models (RSM) are compact analytical models. These compact models are based on limited numbers of input and output sets obtained from computationally extensive simulation models. If these models are precisely constructed, surrogate models are able to approximate the behavior of complex system at reduced computational time [7]. Surrogate Model Based on Optimization (SMBO) is one of the most practical types of surrogate models which is utilized to solve the nonlinear complex problems. The main steps of constructing a SMBO are explained in the following paragraphs [8].

1. Sampling plan: in this step, the relevant input and output variables are selected as per their degree of relevance [8]. Furthermore, Latin Hypercube Sampling (LHS) technique is suggested to utilize for generating adequate number of sample sets of these variables.
2. Implementation of the numerical simulation models: the solution results of simulation models for randomly generated input variables in previous step are obtained.
3. Construction of surrogate model: in this step, the type of surrogate model and the architecture of it should be addressed.
4. Model evaluation: this step assesses the eligibility and predictive accuracy of the surrogate model. The results can be used in model selection and selection of the model architecture.
5. Stop/step 3: If the termination criteria are satisfied stop; otherwise, go to step 3.

2.2 Simulation Models

To solve the flow equation, numerical simulation model MODFLOW is used. MODFLOW was developed by the United States Geological Survey (USGS). The three-dimensional equation of groundwater flow through porous media is utilized by MODFLOW which is a partial-differential equation that represents the groundwater flow in non-equilibrium, anisotropic and heterogeneous conditions [9].

In addition, a Modular Three-Dimensional Multi species Transport Model (MT3DMS) is used in this study. This model is used for the groundwater system in order to simulate the advection, dispersion, and chemical reactions process of contaminants to calculate the contamination concentration values. The governing equation is a partial differential equation and considers the fate and transport of contaminants of species in a 3-D transient groundwater flow system[10].

2.3 Self-Organizing Maps

The Self-Organizing Maps (SOM) algorithm transforms complex non-linear statistical multidimensional data problems in to simple geometric relationships [11]. The SOM abstracts the main information and their topologic relationships on a visual display. Therefore, reducing the dimensions and visualizing of data are the two main characteristics of SOM that enable SOM to be practically utilized in different complex fields of sciences [11]. In this study, due to the SOM's ability to abstract the nonlinear

relationships of high-dimensional complex system; SOM is utilized as the tool to construct the surrogate model. The main steps of SOM algorithm are initialization, competition, cooperation and adaptation which are described below:

a. Initialization: all the units in the output domain connect with the input units with an initial network weights value.

b. Competition: for each input pattern, the output neurons compete to declare the winner neuron for each input vector. The winner neuron or Best Matching Unit (BMU) is that one which has the most similarity with the input vector.

c. Cooperation: The winning neuron calculates the spatial distance of exited neighborhood neurons to cooperate with them and update all weights of the winning neuron.

d. Adaptation: For this step, the process repeats steps b to d until the desired iteration is reached, or the changes in the map for two consecutive iterations are less than a specified target value.

In this study, the software "SOM Toolbox for Matlab 5" is used to construct the SOM based surrogate models [12].

2.4 Definition of a Generic Objective Function

The implicit objective function of source identification problem can be defined as by Eq. (1) which minimizes the difference between the estimated and the observed concentration values at specific monitoring locations at specific time [13], [4].

Minimize:

$$E = \sum_{k=1}^{nk} \sum_{iob=1}^{nob} \left(Cest_{iob}^{k} - Cobs_{iob}^{k}\right)^2 . w_{iob}^{k} \quad (1)$$

Where $Cest_{iob}^{k}$ and $Cobs_{iob}^{k}$ are estimated concentration values and observed concentration values at observation well location iob and at the end of time period k, respectively. nk and nob are total number of concentration observations time and total number of observation wells, respectively. It is also possible to normalize the objective function using weights w_{iob}^{k}, where is weight corresponding to observation location iob at the end of time period k, this parameter can be defined as:

$$w_{iob}^{k} = \frac{1}{\left(Cobs_{iob}^{k} + \eta\right)^2} \quad (2)$$

Where η is a constant and should be sufficiently large to prevent the denominator become near zero at very low concentration [1]. The main constraints of the optimization model are

$$Cest_{iob}^{k} = f(x, y, z, v_x, q_s, C_s, t) \quad (3)$$

Where $f(x, y, z, v_x, q_s, C_s, t)$ represents the simulation models or SOM based surrogate model at time step t. x, y, z is Cartesian coordinates of the monitoring locations, v_x is groundwater velocity along the x coordinate axis, q_s is volumetric flux of water per unit volume of aquifer (T^{-1}), C_s is concentration of the sources or sinks (ML^{-3}) and $q_s C_s$ is contaminant source fluxes ($ML^{-3}T^{-1}$). This approach is the linked simulation optimization approach as proposed by [3], [5].

However, in this new approach using the SOM based surrogate model, the surrogate model is utilized in an inverse mode to characterize the unknown sources of contamination from concentration measurement data. The SOM based surrogate model is first trained and tested to approximate the flow and transport processes in the aquifer study area. Once the training and testing processes are complete, the surrogate model represents the approximate simulation model.

In the traditional approach, the surrogate model once developed can be linked to the optimization model represented by objective functions and constraints (1)-(3). In the developed methodology in this study, the optimization model is not solved, and the source characterization based on concentration measurement data is accomplished by running the SOM based surrogate model in inverse mode. The SOM based surrogate model is used to estimate the contaminant source characteristics as the output, while the concertation measurements resulting from the unknown contaminant sources are used as inputs. Therefore, the optimization for an objective similar to the defined objective (1) is actually implicitly carried out by using the developed SOM based surrogate model in inverse mode. The limited performance evaluation results presented here for an illustrative contaminated aquifer study area, utilizing synthetic hydrogeologic data, and simulated concentration measurements establish the potential applicability of this approach.

2.5 Performance Evaluation

The performance of the developed methodology is evaluated for an illustrative contaminated aquifer study area (Fig. 1), with simulated concentrations measurements. The performance evaluation is carried out for two different scenarios based on two different assumptions as stated below.

1. All the hydrogeologic parameters of the model are precisely known; and

2. Uncertainties are associated with the hydraulic conductivity of the study area; and these parameter values are known only at limited sparse locations.

As for the first assumption, the study area considered is heterogeneous and the actual hydraulic conductivity values are assumed to be a random variable. Therefore, in order to generate hydraulic conductivity throughout the entire study area the values of hydraulic conductivity (K) are assumed to follow the Lognormal distribution [14]. Thus, it is possible to define a new parameter such as Y = log K which is normally distributed. Also, the LHS method is utilized to randomly generate the hydraulic conductivity field throughout the study area following the method used in [15].

The second assumption implies that the hydraulic conductivity measurements are available only at limited locations, while the simulation models need this parameter values at all its nodes. Therefore, hydraulic conductivity should be estimated at other nodes. According to [16] the Inverse Distance Weighting (IDW) methodology could be the most suitable method to generated hydraulic conductivity because of its simplicity, and thus, associated computational ease. This study also demonstrated that the more complicated interpolation methods such as Kriging or fractal-based methods perform little better compared to simplified method such as the IDW. Also, with very sparse measurement data these two methods are not suitable. Therefore, in this study IDW is utilized to generate hydraulic conductivity values at locations where these values are unknown.

Moreover; to quantify the performance evaluation of the developed procedure, Normalized Absolute Error of Estimation (NAEE) is used as a criterion in this study. This parameter calculates a normalized error of estimation. Equation (4) represents NAEE [4]:

$$\boldsymbol{NAEE}(\%) = \frac{\sum_{i=1}^{S}\sum_{j=1}^{N}\left|\left(q_i^j\right)_{est} - \left(q_i^j\right)_{act}\right|}{\sum_{i=1}^{S}\sum_{j=1}^{N}\left(q_i^j\right)_{act}} \times \mathbf{100} \qquad (4)$$

Where S is number of pollution source(s); N is number of transport stress periods; $\left(q_i^j\right)_{act}$ is actual source flux at potential source number i in stress period j; and $\left(q_i^j\right)_{est}$ is estimated source flux at source number i in stress period j.

3. RESULTS AND DISCUSSION

The illustrative study area utilized for the performance evaluation of the proposed methodology is a heterogeneous aquifer which consists of three unconfined layers. This study area is shown in Fig. 1. Table 1 shows the aquifer characteristic values and dimensions of this study area.

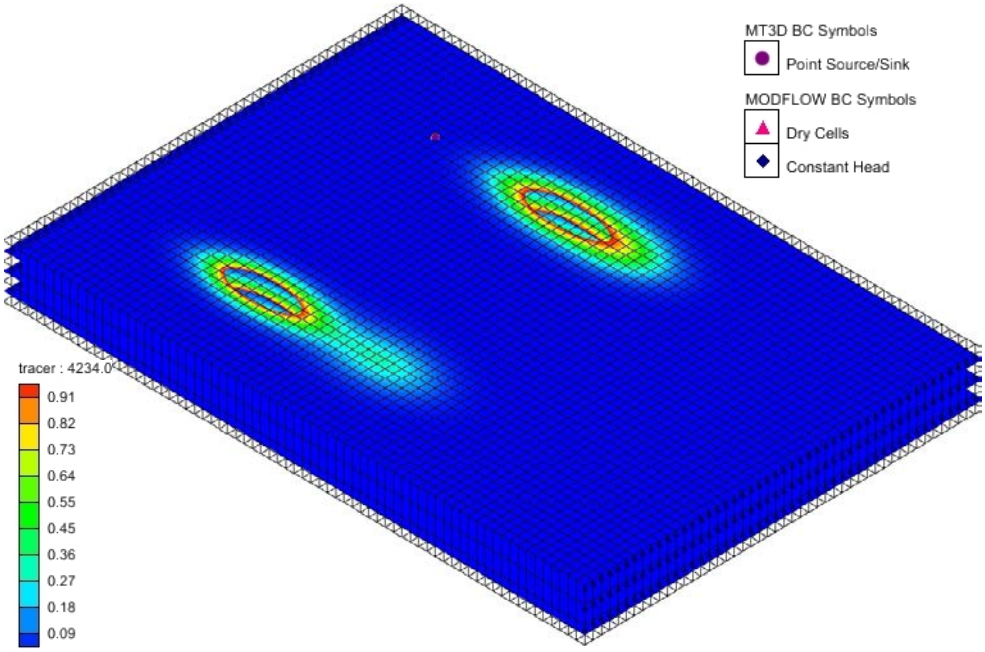

Fig. 1 Illustrative study area representing typical concentration plumes 4234 days after start of first source activity (concentration values g/l)

In this study area, the north and south boundaries are considered as no-flow boundaries. Whereas, the east and west boundaries are considered as specified head boundaries. Only a

conservative contaminant and two potential contaminant source locations (S1 and S2) are considered. S1 and S2 are located in layer 1 and layer 2, respectively. Table 2 shows the locations and flux magnitudes of the actual contaminant sources. There are five monitoring wells with their locations shown in table 3. The total time of simulation is divided into 5 different stress periods. The first four stress periods are each of two years duration and the last stress period is of 12 years duration. Potential contaminant sources are assumed to be active only in the first four stress periods. It is specified that the contamination is detected just two years after the contaminant sources had stopped their activity. It is also specified that the five monitoring locations are monitored over the last 10 years at an interval of 73 days.

Table 1 Hydrogeologic characteristics of the study area

Parameter	Unit	Value
Maximum length of study area	m	2100
Maximum width of study area	m	1500
Saturated thickness, b	m	30
Grid spacing in X-direction	m	30
Grid spacing in Y-direction	m	30
Grid spacing in Z-direction	m	10
Hydraulic Conductivity in X-direction	m/d	20
Hydraulic Conductivity in Y-direction	m/d	20
Vertical anisotropy		5
Hydraulic gradient		0.00238
Porosity		0.3
Longitudinal Dispersivity	m/d	15
Transvere Dispersivity	m/d	3
Initial Contaminant Flux	g/s	0-10

Table 2 Locations and flux magnitudes of actual contaminant sources

Source	Row	Column	Stress Period (SP)	Contaminant fluxes (g/s)
1	12	15	1	6.25
			2	4.63
			3	9.03
			4	5.56
			5	0.00
2	38	9	1	6.69
			2	9.35
			3	6.10
			4	7.28
			5	0.00

Table 3 Locations of monitoring wells

Monitoring Location	Layer	Row	Column
1	1	12	21
2	1	12	35
3	1	26	28
4	1	38	16
5	1	38	29

3.1 Validation of the Model

In order to solve the source identification problem and evaluate the performance of the SOM based surrogate model, the following steps are followed.

1. Scenarios for sampling plan: LHS is utilized to produce three groups of 250, 500, and 1000 initial sample sets with 2 potential contaminant sources with contaminant fluxes in the range of 0-10 g/s.
2. Implementing simulation models: groundwater flow and transport simulation models MODFLOW and MT3DMS (within GMS 7) are solved for three randomly generated groups of source fluxes. The simulation results provide the contaminant concentration values at the five monitoring locations resulting from these contaminated sources as specified.
3. Construction of the surrogate models: SOM algorithm is applied to create surrogate models representing the relationship between the aquifer stresses in the form of contaminant injection and the resulting impacts in terms of the concentration values at specified different locations and times. The randomly generated potential source fluxes and their corresponding contaminant concentration magnitudes at specified monitoring locations at specified time are used as the inputs for training of the SOM based surrogate models. Then, to find the unknown characteristics (magnitude, location and duration) of potential contaminant sources, the BMU of SOM algorithm that satisfies similar criterion as the implicit objective function (1) of this problem is utilized. Therefore, this capability of SOM based surrogate models eliminates the necessity of using any complex and explicit optimization model.
4. Validation of the model: A group of 100 randomly generated sample sets of potential contaminant source fluxes and

corresponding simulated measured concentrations are utilized to test the performance of the developed model once it has been adequately trained. In this step, the performance of the trained SOM models for different scenarios representing different numbers of initial sample sizes and SOM map units are evaluated in terms of NAEE defined by Eq. (4).

Different surrogate models using different numbers of initial sample sets i.e., 250, 500, 1000, 1500 and 1750 are constructed. The randomly generated source fluxes at two potential contaminant sources and corresponding concentration measurement data at 5 selected monitoring locations at an interval of 73 days over the last 10 years of simulation are used to construct these surrogate models. In these scenarios, the numbers of SOM map units are maintained constant (100×100 units). The best results among these SOM based surrogate models are obtained by using 1500 initial sample sets. Also, different SOM based surrogate models representing different numbers of SOM map units are constructed. In these scenarios, the number of initial sample sets is maintained constant at 1500. The best solution result for source identification is obtained by utilizing 100×100 map units. An important constraint in these evaluations of different scenarios is the required CPU time, which significantly increases when the numbers of SOM map units are more than 14400 (Fig.2).

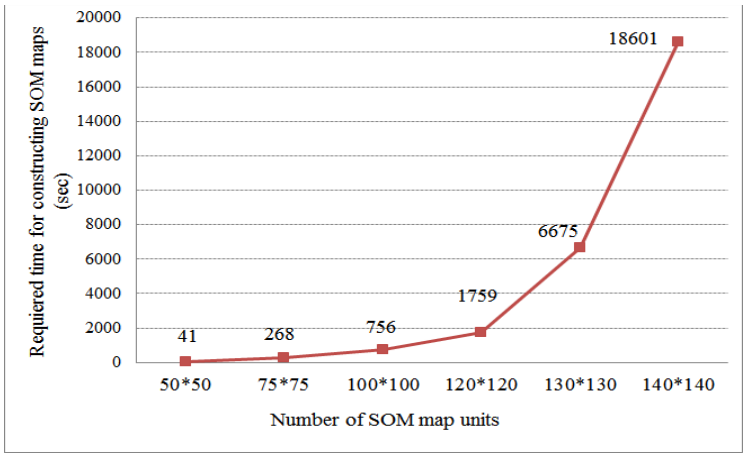

Fig. 2 Computational times for constructing different SOM based surrogate models representing different numbers of SOM map units

5. Finally, based on the validation results obtained in the previous stage, the best candidate SOM based surrogate model is selected. Once the best candidate SOM model is validated and chosen, it is used for further performance evaluation of the developed methodology with the hydraulic conductivity values are assumed to be uncertain.

When utilizing the first assumption, the hydraulic conductivity field for the whole study area is generated by assuming that the mean of hydraulic conductivity in each of the three layers (layer 1, 2 and 3) are 20, 17, and 21 m/day and the standard deviation are 0.1, 0.08, and 0.12, respectively. In the second assumption, it is assumed that the hydraulic conductivity measurements are available only at 20 locations. The distances between any two locations along the maximum length and minimum length of the study area are 300 and 450 meters, respectively. Therefore, to generate hydraulic conductivity values at other locations; the IWD method is utilized as the interpolation method, due to its efficiency and simplicity [16]. Figure 3 represents the generated hydraulic conductivity field for layer 1 using IWD interpolation method.

The obtained NAEE values for source identification based on the first and second assumptions are equal to 14 and 16 percent, respectively. These values are averaged over the 5 stress periods (sp1, sp2, sp3, sp4 and sp5) for two actual contaminant sources (S1 and S2). Figure 4 represents the results of source identification for both the assumptions. This figure compares the estimated source flux values with the actual source flux values.

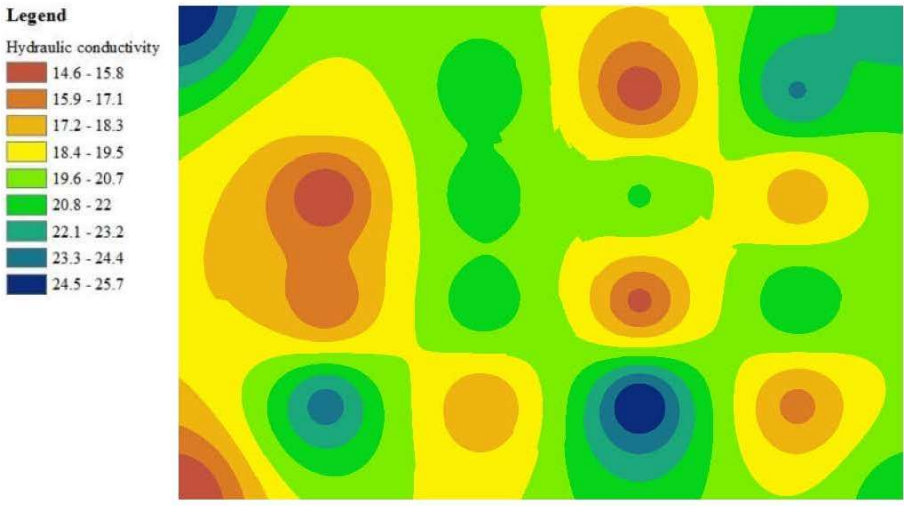

Fig. 3 Generated hydraulic conductivity for layer 1

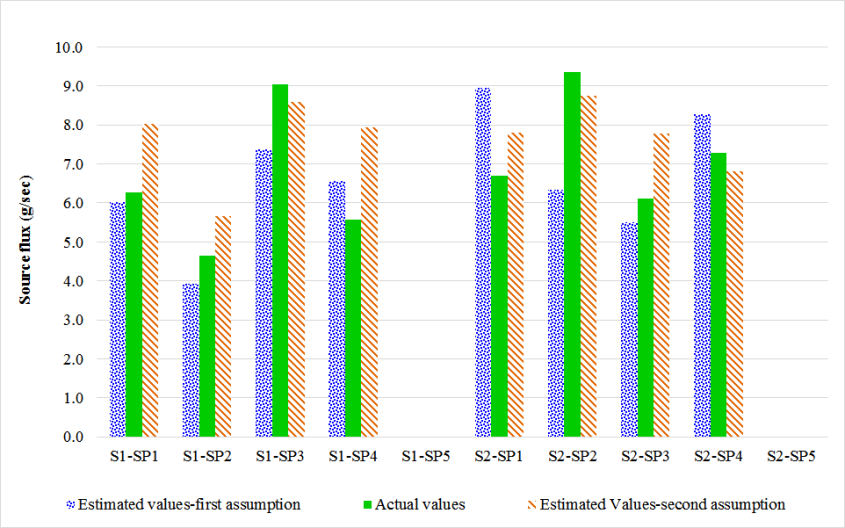

Fig.4 Source identification results

4. CONCLUSION

In this study, in order to develop a SOM based surrogate model, different scenarios representing different numbers of initial sample sizes and SOM map units are considered. Also, the performance of the developed methodology is evaluated by considering two scenarios representing two assumptions: first, the hydrogeologic parameters, i.e., hydraulic conductivity are assumed to be known. Second: hydraulic conductivity values are uncertain and it is assumed that measurement values are known only at 20 locations. Main conclusions that can be obtained from these limited performance evaluation results are:

1. The SOM based surrogate model could approximate groundwater flow and transport simulation models adequately. Also, this developed methodology provides an alternative methodology to identify unknown characteristics of unknown contaminant source in terms of location, magnitude and duration of source activity, without explicitly using a linked simulation-optimization approach.
2. The initial sample size used for training has crucial role on the efficiency of the SOM based surrogate models. This size should be sufficient to properly cover the whole range of potential contaminant source fluxes and corresponding contaminant concentration values. However, very larger number of initial sample sizes may also sometimes decrease the accuracy of the solution results.
3. The optimal numbers of SOM map units are important. This parameter relates to

the memory of the PC used and initial sample sizes.

4. The most important conclusion is that the SOM based surrogate models independently provide a procedure for contaminant source identification, without the necessity of using a linked simulation optimization model.
5. The performance evaluation results are based on very limited scenarios and therefore restricted in scope. Further performance evaluations are required to fully establish the applicability of the proposed methodology.

5. 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.

6. REFERENCES

[1] Mahar, P.S. and B. Datta, Optimal monitoring network and ground-water-pollution sources identification. Journal of Water Resource Planning and Management, 1997. 123 (4): p. 199-207.

[2] Mahar, P.S. and B. Datta, Identification of pollution sources in transient groundwater systems. Water Resources Management, 2000. 14(3): p. 209-227.

[3] Singh, R.M. and B. Datta, Identification of groundwater pollution sources using GA-based linked simulation optimization model. Journal of Hydrologic Engineering, 2006. 11(2): p. 101-109.

[4] Jha, M. and B. Datta, Three-Dimensional Groundwater Contamination Source Identification Using Adaptive Simulated Annealing. Journal of Hydrologic Engineering, 2013. 18(3): p. 307-317.

[5] Prakash, O. and B. Datta, 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, 2015. 23(6): p. 1089-1107.

[6] Amirabdollahian, M. and B. Datta, Reliability Evaluation of Groundwater Contamination Source Characterization under Uncertain Flow Field. International Journal of Environmental Science and Development, 2015. 6(7): p. 512-518.

[7] Gorissen, D., et al., A Surrogate Modeling and Adaptive Sampling Toolbox for Computer Based Design. Journal of Machine Learning Research, 2010. 11: p. 2051-2055.

[8] Forrester, A.I.J. and A.J. Keane, Recent advances in surrogate-based optimization. Progress in Aerospace Sciences, 2009. 45(1-3): p. 50-79.

[9] Harbaugh, A.W., MODFLOW-2005, The U.S. Geological Survey Modular Ground-Water Model-the Ground-Water Flow Process. 2005: U.S. Geological Survey Techniques and Methods 6–A16.

[10] Zheng, C. and P.P. Wang, 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. 1999: US Army Corps of Engineers-Engineer Research and Development Center, Contract Report SERDP-99-1. p. 220.

[11] Kohenon, T., et al., Engineering Applications of the Self-Organizing Map. IEEE, 1996. 84(10): p. 1358-1384.

[12] Vesanto, J., et al., Self-Organizing Map in Matlab: the SOM toolbox. 2000.

[13] Mahar, P.S. and B. Datta, Optimal identification of ground-water pollution sources and parameter estimation. Journal of Water Resources Planning and Management-Asce, 2001. 127(1): p. 20-29.

[14] Freeze, R.A., A Stochastic-Conceptual Analysis of One-Dimensional Groundwater Flow in Nonuniform Homogeneous Media. Water Resources Research, 1975. 11: p. 17.

[15] Dokou, Z. and G.F. Pinder, Optimal search strategy for the definition of a DNAPL source. Journal of Hydrology, 2009. 376(3-4): p. 542-556.

[16] Borman, G.K., F.J. Molz, and O. Guven, An Evaluation of Interpolation Methodologies for Generating Three-Dimensional Hydraulic Property Distribution from Measured Data. 1995. 33: p. 12.