Hydrogeology Journal (2018) 26:923–932
https://doi.org/10.1007/s10040-017-1690-1

PAPER

Comparative study of surrogate models for groundwater contamination source identification at DNAPL-contaminated sites

Zeyu Hou [1,2] · Wenxi Lu [1,2]

Received: 26 May 2017 / Accepted: 3 November 2017 / Published online: 27 November 2017

Abstract

Knowledge of groundwater contamination sources is critical for effectively protecting groundwater resources, estimating risks, mitigating disaster, and designing remediation strategies. Many methods for groundwater contamination source identification (GCSI) have been developed in recent years, including the simulation–optimization technique. This study proposes utilizing a support vector regression (SVR) model and a kernel extreme learning machine (KELM) model to enrich the content of the surrogate model. The surrogate model was itself key in replacing the simulation model, reducing the huge computational burden of iterations in the simulation–optimization technique to solve GCSI problems, especially in GCSI problems of aquifers contaminated by dense nonaqueous phase liquids (DNAPLs). A comparative study between the Kriging, SVR, and KELM models is reported. Additionally, there is analysis of the influence of parameter optimization and the structure of the training sample dataset on the approximation accuracy of the surrogate model. It was found that the KELM model was the most accurate surrogate model, and its performance was significantly improved after parameter optimization. The approximation accuracy of the surrogate model to the simulation model did not always improve with increasing numbers of training samples. Using the appropriate number of training samples was critical for improving the performance of the surrogate model and avoiding unnecessary computational workload. It was concluded that the KELM model developed in this work could reasonably predict system responses in given operation conditions. Replacing the simulation model with a KELM model considerably reduced the computational burden of the simulation–optimization process and also maintained high computation accuracy.

Keywords Multiphase flow · Groundwater contamination source identification · Numerical modeling · Simulation–optimization · Surrogate model

Introduction

Dense nonaqueous phase liquids (DNAPLs), which have caused serious environmental and health hazards around the world (Fernandez-Garcia et al. 2012), have low solubility, high toxicity, high interfacial tension, and a high tendency to sink in water (Qin et al. 2007). There are many difficulties in DNAPL-contaminated aquifer remediation such as low contaminant removal rates, long remediation durations, and high remediation costs. Thus, selecting a reasonable and efficient remediation strategy based on information about the DNAPL contamination source in the aquifer is critical.

However, one of the characteristics of groundwater contamination is concealment, and the discovery of groundwater contamination usually lags behind the contamination event or events, which results in minimal knowledge about the groundwater contamination sources, including their number, location, and release history (Atmadja and Bagtzoglou 2001; Sun et al. 2006; Sun 2009), thus making groundwater contamination source identification (GCSI) especially important.

GCSI is accomplished by inversely solving a simulation model that describes contaminant transport in the aquifer based on limited groundwater contamination monitoring data. GCSI can be used to take effective action in protecting groundwater resources, estimating risks, mitigating disaster, and designing remediation strategies (Mirghani et al. 2012).

There have been several comprehensive reviews of GCSI (Atmadja and Bagtzoglou 2001; Michalak and Kitanidis 2004; Bagtzoglou and Atmadja 2005). Among the proposed

✉ Wenxi Lu
luwx999@163.com

[1] Key Laboratory of Groundwater Resources and Environment, Ministry of Education, Jilin University, Changchun 130021, China

[2] College of Environment and Resources, Jilin University, Changchun 130021, China

solutions, the simulation–optimization method (Ayvaz and Karahan 2008; Mirghani et al. 2009; Ayvaz 2010; Datta et al. 2011; Zhao et al. 2016) and the Bayesian method (Michalak and Kitanidis 2003; Wang and Jin 2013; Zeng et al. 2012; Zhang et al. 2015, 2016) are effective tools for solving GCSI problems. The effectiveness of the simulation–optimization method on programming and identification has been confirmed in many fields; however, running a multiphase flow numerical simulation model of DNAPL-contaminated aquifers is time consuming. The high computational burden that results from invoking the numerical simulation model repeatedly limits the applicability of GCSI simulation–optimization modeling at DNAPL-contaminated sites.

Previous studies (e.g., Mirghani et al. 2009, 2010) have mostly relied on parallelization and grid computing to decrease the computation time of the simulation model. The emerging surrogate model, which has a similar input and output relationship to the simulation model, can be computed several orders of magnitude faster than the simulation model (Queipo et al. 2005; Sreekanth and Datta 2010).

The most crucial requirement of the surrogate model is its approximation accuracy, because it greatly influences the reliability of the simulation–optimization model. Many surrogate model techniques have been applied to groundwater remediation strategy optimization problems such as polynomial regression (He et al. 2008), radial basis function artificial neural networks (RBFANN; Bagtzoglou and Hossain 2009; Luo et al. 2013), the Kriging algorithm (Hou et al. 2016), and support vector regression (SVR; Hou et al. 2015).

Asher et al. (2015) present a review of surrogate models and their application to groundwater modeling. The surrogate modeling techniques fall into three categories: data-driven, projection, and hierarchical-based approaches. The techniques mentioned before are all data-driven surrogates, which approximate a groundwater model through an empirical model that captures the input–output mapping of the original model, and were most widely used. Artificial neural networks (ANNs) are the most popular tool used as a surrogate of the numerical simulation model for GCSI problems (Singh et al. 2004; Rao 2006; Mirghani et al. 2012; Srivastava and Singh 2014, 2015); however, they suffer from instability and overfitting problems that are difficult to solve. Zhao et al. (2016) applied the Kriging model to GCSI problems and tested the accuracy, calculation time, and robustness of the Kriging model in three cases. However, the applicability of the Kriging model in GCSI of DNAPL-contaminated aquifers has not previously been reported; furthermore, there are few applications of other surrogate models in GCSI problems.

This study therefore proposes utilizing the SVR and kernel extreme learning machine (KELM) models to enrich the content of the surrogate model for solving GCSI problems, especially for DNAPL-contaminated aquifers. Additionally, the report examines the effectiveness of the proposed model with a comparative study between the Kriging, SVR, and KELM models, and finds that the disparities in applicability and approximation accuracy between these models for solving DNAPL-contaminated aquifer solute migration and transformation problems is significant. It is therefore necessary to select a best-fit surrogate model for the target problem.

In addition to the modeling method, the parameters and training sample dataset structure of the surrogate model also strongly impact its approximation accuracy to the simulation model; however, these aspects have been insufficiently investigated. Previous work has generally determined the parameters and the number of training samples empirically (Mirghani et al. 2012; Luo et al. 2013; Jiang et al. 2015; Zhao et al. 2016). As an extension of previous studies, this paper presents another two comparative studies analyzing the influence of these factors on the approximation accuracy of the surrogate model—first, there is an examination of the differences in the surrogate models with and without parameter optimization, and then it examines surrogate models built with different numbers of training samples.

Methodology

Multiphase flow numerical simulation model

Any meaningful approach to GSCI problems must obey the flow and transport principle. The simulation model is the principal part of the simulation–optimization model, in which the simulation model is set as an equality constraint (Datta et al. 2011). An overview of this process is shown in Fig. 1.

The fundamental mass conservation equation for each multiphase flow component can be written as follows (Hou et al. 2015; Jiang et al. 2015):

$$\frac{\partial\left(\phi\tilde{C}_k\rho_k\right)}{\partial t}+\vec{\nabla}\left[\sum_{l=1}^{2}\rho_k\left(C_{kl}\vec{v}_l-\phi S_l\vec{\vec{K}}_{kl}\vec{\nabla}C_{kl}\right)\right]=R_k \quad (1)$$

where k is a component index and l is a phase index including water and oil. The initial and boundary conditions were integrated with the mass conservation equation to build the mathematical model, which was solved by UTCHEM.

Kriging

Kriging was denoted as the sum of two components: the linear model and a systematic departure (Hemker et al. 2008). The basic formulation can be expressed as (Bagtzoglou et al. 1991, 1992)

$$y(\mathbf{x})=\mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}+Z(\mathbf{x})=\sum_{j=1}^{k}f_j(\mathbf{x})\beta_j+Z(\mathbf{x}) \quad (2)$$

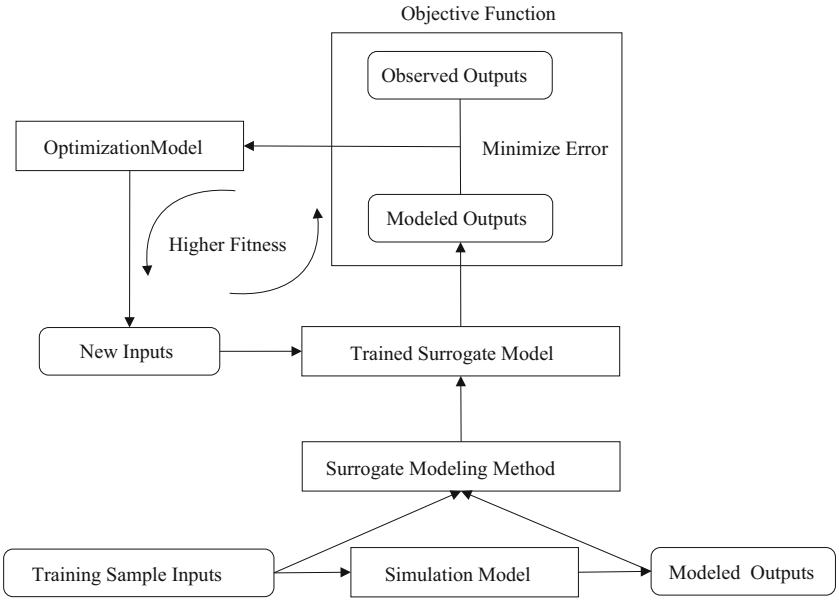

Fig. 1 Flow chart of the proposed GCSI solution process

where $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_k(\mathbf{x})]^{\mathrm{T}}$ are determinate regression functions and $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_k)^{\mathrm{T}}$ denotes the matrix of regression coefficients to be estimated from the training samples. $Z(\mathbf{x})$ is the local deviation from the regression model. A detailed introduction to the Kriging method can be found in Hou et al. (2015) and Zhao et al. (2016).

Support vector regression

SVR is a support vector machine (SVM)-based multiple regression method that balances fitting accuracy and prediction accuracy (Hu et al. 2014; Zhang et al. 2014). For training input $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x}_m]^{\mathrm{T}}$ (where each element represents an N-dimensional input vector $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,N})$, $i = 1, 2\cdots, m$) and output $\mathbf{Y} = (y_1, y_2, \cdots, y_m)^{\mathrm{T}}$, the nonlinear regression function can be expressed as:

$$f(\mathbf{x}_i) = \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle + b \tag{3}$$

where $\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle$ denotes the dot product of fitting coefficients $\mathbf{w} = (w_1, w_2, \cdots, w_N)$ and $\mathbf{x}_i$, b is the fitting error. The goal is to find a function $f(\mathbf{x}_i)$ that has at most ε deviation from the target output y_i for all training inputs; the norm of $\mathbf{w}(\|\mathbf{w}\|)$ should be as small as possible.

A kernel function is applied to project the samples from low-dimensional space to high-dimensional space:

$$k\left(\mathbf{x}, \mathbf{x}^{'}\right) = \exp\left[-\frac{\left\|\mathbf{x}\text{-}\mathbf{x}^{'}\right\|^2}{2\sigma}\right] \tag{4}$$

The regression problem can be expressed as an optimization problem:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\left(\xi_i + \xi_i^*\right) \\ &\text{subject to} \quad \begin{cases} y_i - \langle \mathbf{w}, \Phi(\mathbf{x_i}) \rangle - b \le \varepsilon + \xi_i \\ \langle \mathbf{w}, \Phi(\mathbf{x_i}) \rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \end{aligned} \tag{5}$$

where constant C determines the trade-off between the flatness and the maximum tolerable number of the samples whose deviation is larger than ε, and ξ_i and ξ_i^* are the upper and lower limits of the slack variables. The optimization problem in Eq. (4) is often solved in its Lagrange dual form (Smola and Scholkopf 2004; Hou et al. 2015):

$$\mathbf{w} = \sum_{i=1}^{m}\left(\alpha_i - \alpha_i^*\right)\Phi(\mathbf{x_i}) \ , \quad f(\mathbf{x}) = \sum_{i=1}^{m}\left(\alpha_i - \alpha_i^*\right)k(\mathbf{x_i}, \mathbf{x}) + b \tag{6}$$

where α_i and α_i^* are Lagrange multipliers. Fitting error b can be computed by exploiting the Karush-Kuhn-Tucker (KKT) conditions.

Kernel extreme learning machine (KELM)

KELM generalize extreme learning machines (ELM) by transforming their explicit activation function to an implicit mapping function (Shi et al. 2014; Chen et al. 2014). Given N training samples $(\mathbf{x}_j, t_j)$, $j = 1, \cdots, N$, the KELM is expressed as an optimization model:

$$\begin{aligned} &\min \quad \left\{\frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{C}{2}\sum_{i=1}^{N}\xi_j^2\right\} \\ &\text{subject to} \quad \mathbf{m}\left(\mathbf{x}_j\right)^{\mathrm{T}} \cdot \boldsymbol{\beta} = t_j - \xi_j \end{aligned} \tag{7}$$

where $\boldsymbol{\beta}$ denotes a vector in the feature space F, C denotes the regularization coefficient, $\mathbf{m}(\mathbf{x}_j)$ maps the input $\mathbf{x}_j$ to a vector in F, and ξ_j denotes the error (Wang and Han 2014).

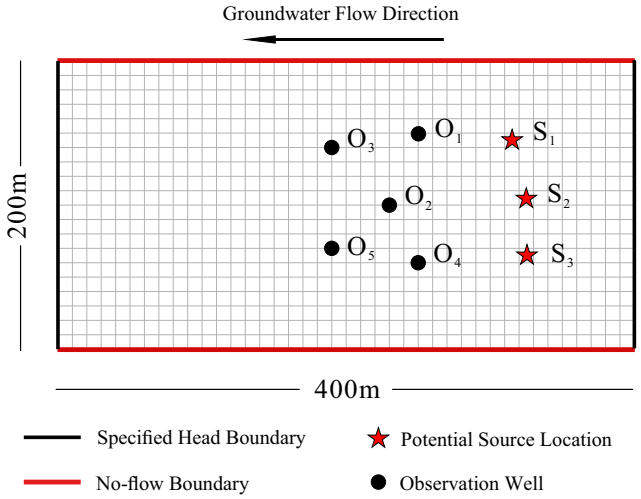

Fig. 2 Locations of potential contamination sources (S_1, S_2, S_3) and observation wells (O_1–O_5)

The optimization problem can be transformed into Lagrange dual (L_D) form

$$L_D = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{C}{2}\sum_{j=1}^{N}\xi_j^2 - \sum_{j=1}^{N}\theta_j\left(\mathbf{m}(\mathbf{x}_j)^{\mathrm{T}}\cdot\boldsymbol{\beta} - t_j + \xi_j\right) \quad (8)$$

where θ_j is the jth Lagrange multiplier. This problem can be computed by exploiting the KKT optimality conditions (Jiang et al. 2015).

The kernel matrix of the ELM can be defined as

$$\mathbf{K_{ELM}} = \mathbf{M}\mathbf{M}^{\mathrm{T}} \quad (9)$$

and

$$K_{\mathrm{ELM}(i,j)} = \mathbf{m}(\mathbf{x}_i)^{\mathrm{T}}\cdot\mathbf{m}(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j) \quad (10)$$

where $\mathbf{M}$ is the mapping matrix of training sample inputs in the feature space F.

Finally, the KELM output function can be written as

$$f(x) = \mathbf{m}(\mathbf{x})^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\left(\mathbf{M}\mathbf{M}^{\mathbf{T}} + \frac{\mathbf{I}}{C}\right)^{-1}\mathbf{T} = \begin{bmatrix} K(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ K(\mathbf{x}, \mathbf{x}_N) \end{bmatrix}^{\mathrm{T}}\left(\mathbf{K_{ELM}} + \frac{\mathbf{I}}{C}\right)^{-1}\mathbf{T} \quad (11)$$

Case study

Site overview

To analyze the practical application of different surrogate models for DNAPL-contaminated aquifer GCSI problems, a hypothetical chlorobenzene-contaminated site was set up as a case study. The site was located in the saturated zone of a 20-m-deep aquifer with a complex mixture of clay and sand deposits in which the groundwater flowed in a right-left direction. There are three potential contamination sources at the site. The goal was to simultaneously identify the actual single source, release strength, and release duration, and estimate the aquifer parameters. Five observation wells were set at the lower reaches of the groundwater gradient of the potential sources to obtain groundwater quality data (Fig. 2).

Multiphase flow numerical simulation model

A three-dimensional (3D) multiphase flow numerical model was developed in which the aquifer is homogeneous and it was assumed that the initial and boundary conditions are known. The left and right boundaries of the site were set as first-type boundary conditions, while other boundaries were no-flux boundaries. The horizontal hydraulic gradient was set to 0.0112. The simulation domain was discretized into 10 vertical layers, each of which was further discretized into 40 × 20 grid cells.

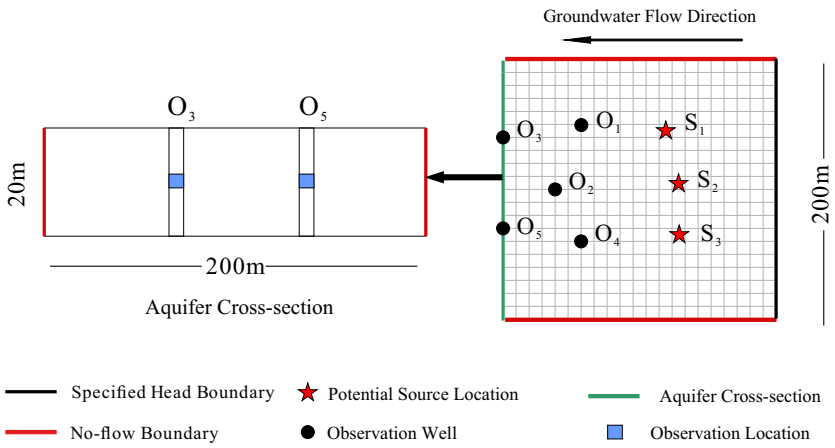

Fig. 3 Schematic diagram of chlorobenzene concentration observation location

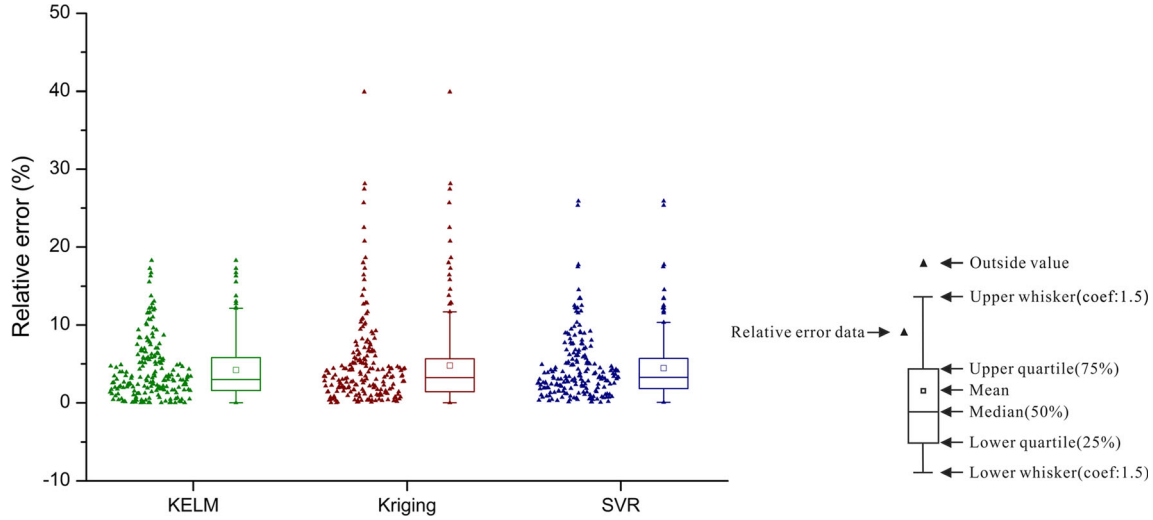

Fig. 4 Boxplot of relative errors of different surrogate models

Surrogate models of the multi-phase flow numerical simulation model

In order to identify the DNAPL source, an optimization model was established that uses the minimal deviation between actual observations and model predictions as its objective function; this model will be demonstrated in future research, as this study focuses on the surrogate model. The size of the surrogate model output should be matched to the actual groundwater-quality-observation data. There were two sets of actual groundwater-quality-observation data with an interval of 6 months between them. Each set of observation data contained five constants, i.e. the chlorobenzene concentrations at the middle of the aquifer in five observation wells (Fig. 3).

The middle aquifer was chosen as an observation object because there may be an oil phase while sampling at the bottom of an aquifer in real-world situations, and the sampling proportion of water and oil is random, leading to significant deviation between the experimental analysis results and the actual volume fraction of oil in the groundwater at the bottom of the aquifer. Thus, the output variables of the surrogate model were the chlorobenzene concentrations at the middle of the aquifer in five observation wells at two observation time points, for a total of 10 elements.

The release strengths and duration of the three potential DNAPL sources were treated as controllable input variables when building the surrogate model. In addition, calibration and verification cannot be carried out without contaminant source information; thus, the contaminant source and aquifer parameters should be identified simultaneously (Starn et al. 2015). Finally, the input vectors of the surrogate model consist of eight elements: the release durations and strengths of sources S_1, S_2, and S_3; porosity; aqueous phase dispersivity; oleic phase dispersivity; and permeability.

Four groups of training samples and 20 testing samples in feasible regions of input variables were obtained using Latin hypercube sampling (LHS; Hossain et al. 2006). Each training sample group consisted of 30 samples. The release strength and duration were uniform distribution variables in (0, 1.5 m^3 day^{-1}) and (600, 900 days), respectively. The aquifer parameters obey the normal distribution while LHS sampling and the distribution characteristics of porosity, dispersivity, and permeability were taken as N (0.3, 0.0001), N (1 m, 0.01), and N (8,500 md, 100,000), respectively. As the study case was hypothetical, the distribution characteristics of aquifer parameters were assumed. The corresponding outputs of the 140 sets of input vectors were obtained for the developed simulation model runs.

There are three factors that affect approximation accuracy: surrogate modeling method, number of training samples, and surrogate model parameters. To analyze the influence of each of these factors, three comparative studies of different surrogate models were conducted.

Table 1 Performance evaluation of different surrogate models

Performance evaluation indices	Certainty coefficient R^2	Mean relative error (%)	Max relative error (absolute value) (%)
Kriging	0.9719	5.1828	39.9035
SVR	0.9779	4.4636	25.8954
KELM	0.9793	4.2053	18.2611

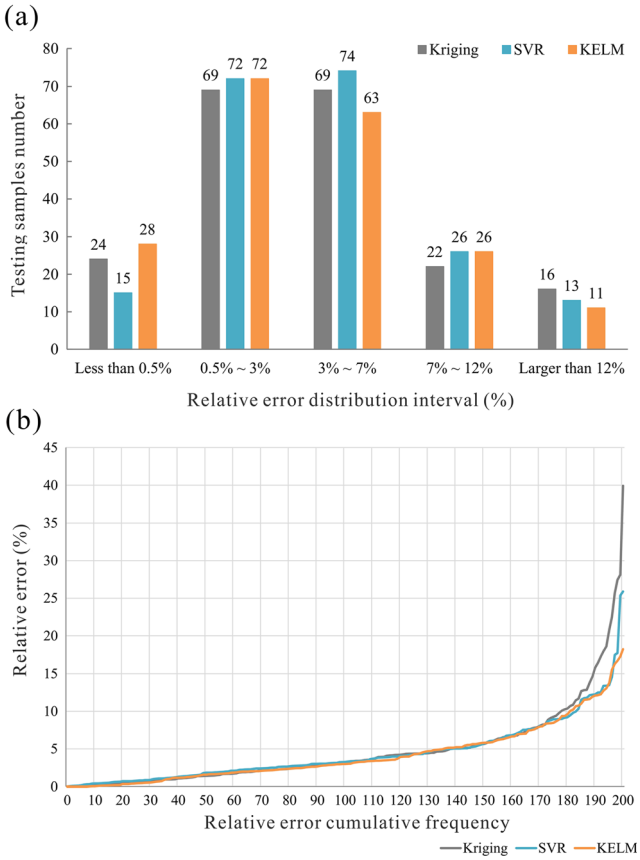

Fig. 5 Distributions of relative errors for different surrogate models. **a** Distribution of relative errors in different intervals; **b** relative error cumulative frequency curve

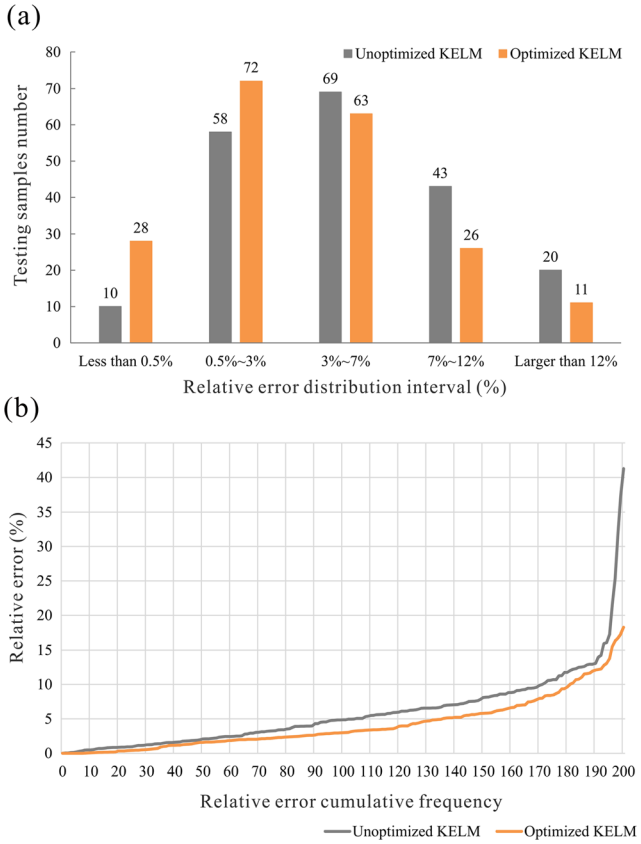

Fig. 7 Distributions of relative errors for KELM models with and without parameter optimization. **a** Distribution of relative errors in different intervals; **b** relative error cumulative frequency curve

Comparison between surrogate models built using different methods

In this experiment, the Kriging, SVR, and KELM models were built with the same training samples and the uncertain parameters of the surrogate models were optimized with a genetic algorithm (GA) to improve their approximation accuracy to the simulation model (Hou et al. 2015). The three models were then compared using test samples. The Kriging and KELM models were built in MATLAB. The Libsvm toolbox (Chang and Lin 2001) was used to train and test the SVR model (Hou et al. 2015). The comparison results showed that

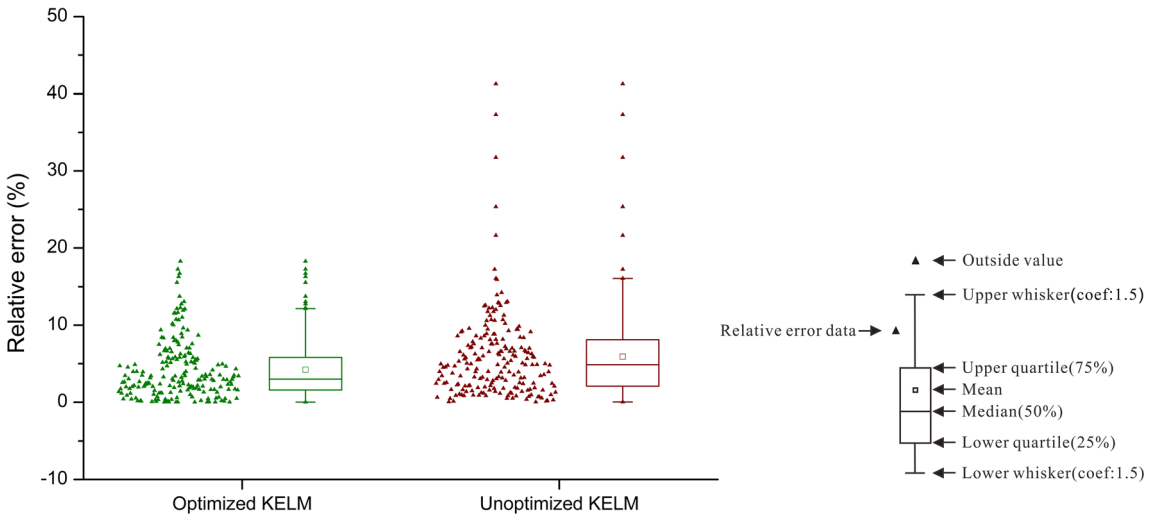

Fig. 6 Boxplot of relative errors of KELM models with and without parameter optimization

Table 2 Performance evaluation of KELM models with and without parameter optimization

Performance evaluation indices	Certainty coefficient R^2	Mean relative error (%)	Max relative error (absolute value) (%)
Unoptimized KELM	0.9615	5.9290	41.2639
Optimized KELM	0.9793	4.2053	18.2611

the KELM model performed best, so only the KELM model was chosen as the research object in the "Comparison between surrogate models with and without parameter optimization" and "Comparison between surrogate models built with different number of training samples" sections.

Comparison between surrogate models with and without parameter optimization

The KELM models with and without parameter optimization were compared using testing samples to analyze the improvement of the surrogate model after parameter optimization. The KELM model was optimized by establishing a model using the minimal sum of the relative error by threefold cross-validation with 90 training samples as its objective function. The regularization coefficient in Eq. (7) and the kernel parameters served as decision variables and the constraints were the range of parameters. A GA was used to solve the optimization model.

Comparison between surrogate models built with a different number of training samples

To analyze the influence of training sample dataset structure on the approximation accuracy of the surrogate model, three KELM models were built and compared. The number of training samples for the three surrogate models were 60, 90, and 120. The parameters of the three surrogate models were optimized by a GA. LHS was used to obtain four groups of 30 training samples; thus, the training sample datasets of three surrogate models consisted of different training sample groups.

Surrogate model performance evaluation indices

Three indices were applied to evaluate the performance of surrogate models:

1. Certainty coefficient R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{i,j} - \hat{y}_{i,j}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{i,j} - \overline{y}\right)^2} \tag{19}$$

where n is the sample number, m is the dimension of the simulation model output vector, $y_{i,j}$ is the jth element of the ith simulation model output vector, $\hat{y}_{i,j}$ is the jth element of the ith surrogate model output vector, and $\overline{y}$ is the average of the simulation model outputs. The surrogate model is better when the R^2 is closer to 1.

2. Mean relative error (MRE)

$$\mathrm{MRE} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left|y_{i,j} - \hat{y}_{i,j}\right| \Big/ y_{i,j}}{n} \tag{20}$$

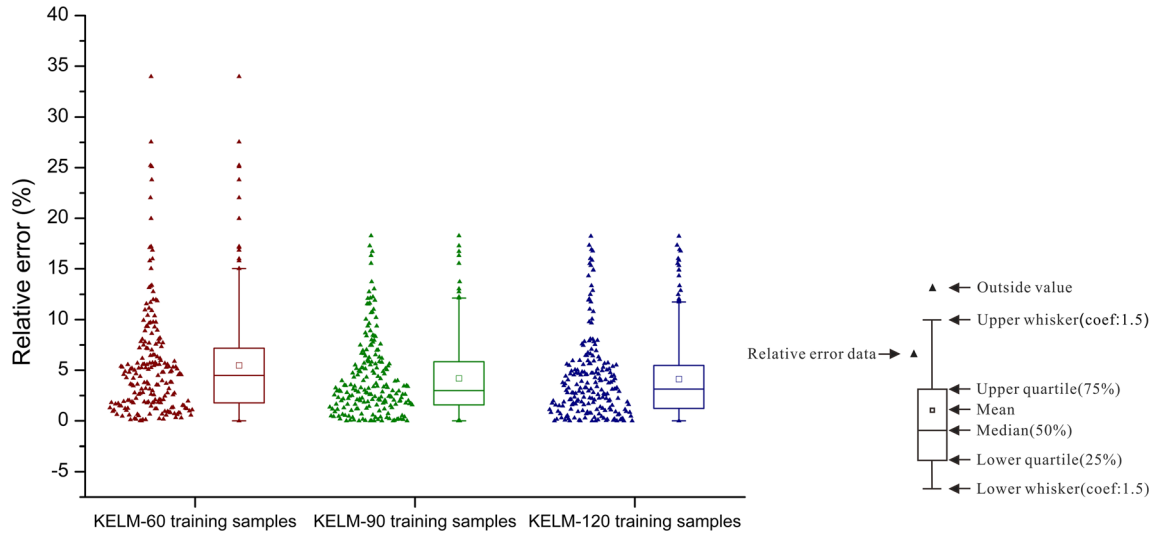

Fig. 8 Boxplot of relative errors of KELM models built with different training sample datasets

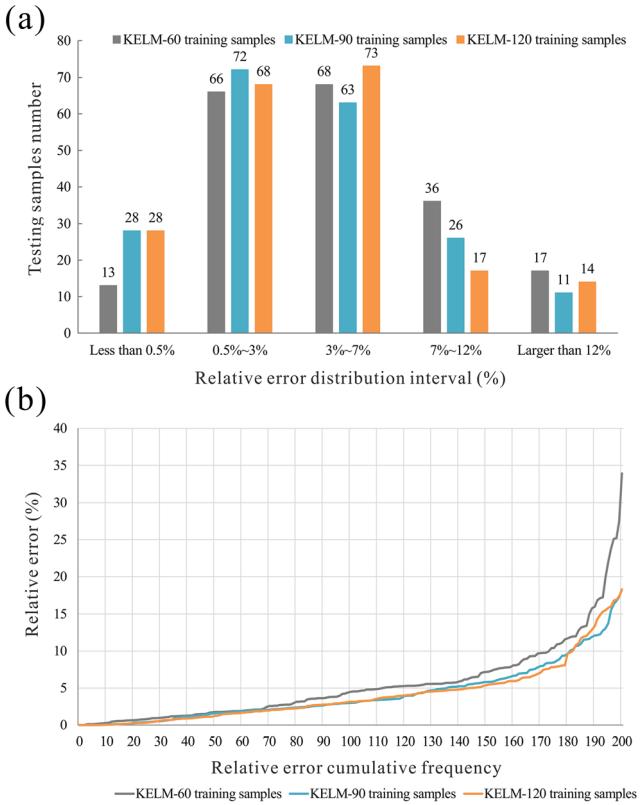

Fig. 9 Distributions of relative errors for KELM models built with different training sample datasets. **a** Distribution of relative errors in different intervals; **b** relative error cumulative frequency curve

3. Maximum relative error

$$\max \left| y_{i,j} - \hat{y}_{i,j} \right| \Big/ y_{i,j} \tag{21}$$

Results and discussion

The outputs of 20 testing samples obtained using the trained surrogate models (a total of 200 values) were compared with those obtained using the developed simulation model. Figure 4 shows boxplots of the relative error metrics corresponding to the three different surrogate models.

The transport of organic contaminants in multiphase flow is complicated and the solubility of chlorobenzene in water is particularly low, making it difficult to follow the relationship between the inputs and outputs of the simulation model. The relative errors of the three surrogate models were higher than those of the same surrogate models applied to other problems (Luo et al. 2013; Hou et al. 2015, 2016; Zhao et al. 2016).

Figure 4 clearly shows that the number of relative errors larger than 20% for the Kriging model are much larger than those of the other two models, and the max relative error for the Kriging model is 39.9035%. These findings illustrated that the Kriging model performance was unstable with respect to this problem. Three surrogate models were also evaluated using the three indices previously described (Table 1). The closer the certainty coefficient R^2 is to 1, the more accurate the surrogate model. Table 1 shows that the accuracy of the KELM and SVR models is higher than that of the Kriging model. Furthermore, the KELM model was better than the SVR model in all indices, and the max relative error for KELM model was less than 20%; thus, it is concluded that the KELM model is an acceptable method for creating a surrogate model.

Figure 5 illustrates the distribution of the relative errors of the surrogate models. The relative error values concentrated between 0.5 and 7%, and most of the relative error values were less than 12%. The KELM and SVR models were significantly superior to the Kriging model, according to the relative error cumulative frequency curves.

Figures 6 and 7 show the results corresponding to the KELM surrogate models with and without parameter optimization. The parameters of the KELM model greatly affect its approximation accuracy. After parameter optimization, all performance evaluation indices of the KELM model were significantly improved (Table 2).

Using 20 testing samples, the maximum and average relative errors of the groundwater contamination monitoring data predicted by the KELM model without parameter optimization (41.2639 and 5.9290%) were far larger than those of the optimized KELM model (18.2611 and 4.2053%). The relative error cumulative frequency curve of the optimized KELM model was located below that of the unoptimized KELM model throughout.

Figures 8 and 9 compare the results corresponding to the KELM surrogate models built with different training sample datasets. When the number of training samples increased from 60 to 90, the approximation accuracy of the KELM model improved significantly. However, the KELM model built with

Table 3 Performance evaluation of KELM models built with different training sample datasets

Performance evaluation indices	Certainty coefficient R^2	Mean relative error (%)	Max relative error (absolute value) (%)
KELM-60 training samples	0.9716	5.4856	33.9618
KELM-90 training samples	0.9793	4.2053	18.2611
KELM-120 training samples	0.9770	4.1288	18.2169

120 training samples performed no better than, or even worse than, the KELM model built with 90 training samples, as per Figs. 8 and 9 and Table 3.

The structure of the training dataset affects the approximation accuracy of the surrogate model; however, the approximation accuracy does not simply improve with increasing numbers of training samples. It is necessary to provide sufficient training samples to improve the performance of the surrogate model, while avoiding unnecessary computation.

The optimal number of training samples depends on the surrogate modeling method, the number of input variables, the number of output variables, and many other factors. Too few training samples cannot cover the input variable intervals well, while too many are unhelpful for improving approximation accuracy; thus, further research on a technique for estimating the number of training samples required for the KELM model is needed.

A conventional simulation optimization model required 20,000 runs of the simulation model. The simulation for the chlorobenzene-contaminated site required nearly 500 s of CPU time on a 3.2GHz Intel core i5 CPU and 4 GB RAM PC platform, while each run of the KELM model just takes 0.9 s. Thus, replacing the simulation model with the KELM model in the optimization process reduced the CPU time from 10,000,000 s (116 days) to 18,000 s (5 h).

Though the approximation accuracy of the surrogate model was acceptable when the optimal surrogate method and parameters were selected, the maximum relative error of the groundwater contamination monitoring data predicted by the KELM model was greater than 15%. Future studies will be needed to further improve the approximation accuracy of the surrogate model and make simulation-surrogate-optimization-based GCSI results more reliable.

Conclusions

This study demonstrates the applicability of the Kriging, SVR, and KELM models for optimal identification of unknown groundwater pollution sources by presenting performance evaluations for different surrogate models. The proposed methodology overcomes some of the severe computational limitations of the embedded simulation–optimization approach.

Three comparative studies were carried out to select the optimal surrogate model and analyze the influence of parameters and the structure of the training dataset on the approximation accuracy of the surrogate model. Several general conclusions that can be drawn from this study are summarized in the following:

1. The KELM model was the most reliable surrogate model of the Kriging, SVR, and KELM models. The KELM model reasonably predicted system responses for given operation conditions.
2. The performance of the KELM model was significantly improved through parameter optimization. Using 20 test samples, the maximum and average relative errors of the groundwater contamination monitoring data predicted by the KELM model without parameter optimization were 41.2639 and 5.9290%, whereas those of the optimized KELM model were only 18.2611 and 4.2053%.
3. The structure of the training dataset significantly affects the approximation accuracy of the surrogate model; however, additional training samples do not always lead to higher approximation accuracy. Determining and utilizing the appropriate number of training samples is critical for improving the performance of the surrogate model and avoiding unnecessary computation.

Acknowledgements This study was supported by the National Nature Science Foundation of China (Grant Nos. 41672232 and 41372237). Special gratitude is given to the journal editors for their efforts on evaluating the work, and the valuable comments of the anonymous reviewers are also greatly acknowledged.

References

Asher MJ, Croke BFW, Jakeman AJ, Peeters LJM (2015) A review of surrogate models and their application to groundwater modeling. Water Resour Res 51(8):5957–5973

Atmadja J, Bagtzoglou AC (2001) State of the art report on mathematical methods for groundwater pollution source identification. Environ Forensic 2(3):205–214

Ayvaz MT (2010) A linked simulation–optimization model for solving the unknown groundwater pollution source identification problems. J Contam Hydrol 117(1–4):46–59

Ayvaz MT, Karahan H (2008) A simulation/optimization model for the identification of unknown groundwater well locations and pumping rates. J Hydrol 357(1–2):76–92

Bagtzoglou AC, Atmadja J (2005) Mathematical methods for hydrologic inversion: the case of pollution source identification, chap. In: Environmental impact assessment of recycled wastes on surface and ground waters: engineering modeling and sustainability, vol 3. In: Kassim TA (ed) The handbook of environmental chemistry, water pollution series, vol 5, part F. Springer, Heidelberg, Germany, pp 65–96

Bagtzoglou AC, Dougherty DE, Tompson AFB (1992) Application of particle methods to reliable identification of groundwater pollution sources. Water Resour Manag 6(1):15–23

Bagtzoglou AC, Hossain F (2009) Radial basis function neural network for hydrologic inversion: an appraisal with classical and spatio-temporal geostatistical techniques in the context of site characterization. Stoch Env Res Risk A 23(7):933–945

Bagtzoglou AC, Tompson AFB, Dougherty DE (1991) Probabilistic simulation for reliable solute source identification in heterogeneous porous media, chap. In: Ganoulis J (ed) Water resources engineering risk assessment. NATO ASI Series, G 29, Springer, Heidelberg, Germany, pp 189–201

Chang, Chih-Chung, Lin, Chih-Jen (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed on December 22, 2016

Chen C, Li W, Su H, Liu K (2014) Spectral-spatial classification of hyperspectral image based on kernel extreme learning machine. Remote Sens 6(6):5795–5814

Datta B, Chakrabarty D, Dhar A (2011) Identification of unknown groundwater pollution sources using classical optimization with linked simulation. J Hydro Environ Res 5(1):25–36

Fernandez-Garcia D, Bolster D, Sanchez-Vila X, Tartakovsky DM (2012) A Bayesian approach to integrate temporal data into probabilistic risk analysis of monitored NAPL remediation. Adv Water Resour 36(SI):108–120

He L, Huang GH, Zeng GM, Lu HW (2008) An integrated simulation, inference, and optimization method for identifying groundwater remediation strategies at petroleum-contaminated aquifers in western Canada. Water Res 42(10–11):2629–2639

Hossain F, Anagnostou EN, Bagtzoglou AC (2006) On Latin hypercube sampling for efficient uncertainty estimation of satellite rainfall observations in flood prediction. Comput Geosci 32(6):776–792

Hou Z, Lu W, Chen M (2016) Surrogate-based sensitivity analysis and uncertainty analysis for DNAPL-contaminated aquifer remediation. J Water Resour Plan Manag 142(11):04016043

Hou ZY, Lu WX, Chu HB, Luo JN (2015) Selecting parameter-optimized surrogate models in DNAPL-contaminated aquifer remediation strategies. Environ Eng Sci 32(12):1016–1026

Hu JN, Hu JJ, Lin HB, Li XP, Jiang CL, Qiu XH, Li WS (2014) State-of-charge estimation for battery management system using optimized support vector machine for regression. J Power Sources 269:682–693

Jiang X, Lu WX, Hou ZY, Zhao HQ, Na J (2015) Ensemble of surrogates-based optimization for identifying an optimal surfactant-enhanced aquifer remediation strategy at heterogeneous DNAPL-contaminated sites. Comput Geosci 84(2015):37–45

Luo JN, Lu WX, Xin X, Chu HB (2013) Surrogate model application to the identification of an optimal surfactant-enhanced aquifer remediation strategy for DNAPL-contaminated sites. J Earth Sci 24(6):1023–1032

Michalak AM, Kitanidis PK (2003) A method for enforcing parameter nonnegativity in Bayesian inverse problems with an application to contaminant source identification. Water Resour Res 39(2):1033

Michalak AM, Kitanidis PK (2004) Estimation of historical groundwater contaminant distribution using the adjoint state method applied to geostatistical inverse modeling. Water Resour Res 40(8):W08302

Mirghani B, Tryby M, Ranjithan R, Karonis NT, Mahinthakumar KG (2010) Grid-enabled simulation–optimization framework for environmental characterization. J Comput Civ Eng 24(6):488–498

Mirghani BY, Mahinthakumar KG, Tryby ME (2009) A parallel evolutionary strategy based simulation–optimization approach for solving groundwater source identification problems. Adv Water Resour 32(9):1373–1385

Mirghani BY, Zechman EM, Ranjithan RS (2012) Enhanced simulation–optimization approach using surrogate modeling for solving inverse problems. Environ Forensic 13(4):348–363

Qin XS, Huang GH, Chakma A, Chen B, Zeng GM (2007) Simulation-based process optimization for surfactant-enhanced aquifer remediation at heterogeneous DNAPL-contaminated sites. Sci Total Environ 381(1–3):17–37

Queipo NV, Haftka RT, Shyy W (2005) Surrogate-based analysis and optimization. Prog Aerosp Sci 41(1):1–28

Rao SVN (2006) A computationally efficient technique for source identification problems in three-dimensional aquifer systems using neural networks and simulated annealing. Environ Forensic 7(3):233–240

Shi Y, Zhao LJ, Tang J (2014) Recognition model based feature extraction and kernel extreme learning machine for high dimensional data. Adv Mater Res 875:2020–2024

Singh RM, Datta B, Jain A (2004) Identification of unknown groundwater pollution sources using artificial neural networks. J Water Resour Plan Manag 130(6):506–514

Smola AJ, Scholkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222

Sreekanth J, Datta B (2010) Multi-objective management of saltwater intrusion in coastal aquifers using genetic programming and modular neural network based surrogate models. J Hydrol 393(3–4):245–256

Srivastava D, Singh RM (2014) Breakthrough curves characterization and identification of an unknown pollution source in groundwater system using an artificial neural network (ANN). Environ Forensic 15(2):175–189

Srivastava D, Singh RM (2015) Groundwater system modeling for simultaneous identification of pollution sources and parameters with uncertainty characterization. Water Resour Manag 29:4607–4627

Starn JJ, Bagtzoglou AC, Green CT (2015) The effects of numerical-model complexity and observation type on estimated porosity values. Hydrogeol J 23(6):1121–1128

Sun AY, Painter SL, Wittmeyer GW (2006) A constrained robust least squares approach for contaminant release history identification. Water Resour Res 42(4):263–269

Sun NZ (2009) Inverse problems in groundwater modeling. Springer, The Netherlands

Wang H, Jin X (2013) Characterization of groundwater contaminant source using Bayesian method. Stoch Env Res Risk A 27(4):867–876

Wang X, Han M (2014) Online sequential extreme learning machine with kernels for nonstationary time series prediction. Neurocomputing 145:90–97

Zeng LZ, Shi LS, Zhang DX, Wu LS (2012) A sparse grid based Bayesian method for contaminant source identification. Adv Water Resour 37(3):1–9

Zhang JJ, Li WX, Zeng LZ, Wu LS (2016) An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems. Water Resour Res 52(8):5971–5984

Zhang JJ, Zeng LZ, Chen C, Chen DJ, Wu LS (2015) Efficient Bayesian experimental design for contaminant source identification. Water Resour Res 51(1):576–598

Zhang YS, Kimberg DY, Coslett HB, Schwartz MF, Wang Z (2014) Multivariate lesion-symptom mapping using support vector regression. Hum Brain Mapp 35(12):5861–5876

Zhao Y, Lu WX, Xiao CN (2016) A Kriging surrogate model coupled in simulation–optimization approach for identifying release history of groundwater sources. J Contam Hydrol 185:51–60