Comparison of surrogate models with different methods in groundwater remediation process

Jiannan Luo and Wenxi Lu*

Key Laboratory of Groundwater Resources and Environment, Ministry of Education, College of Environment and Resources, Jilin University, Changchun 130021, China.
Corresponding author. e-mail: luwenxi@jlu.edu.cn

Surrogate modelling is an effective tool for reducing computational burden of simulation optimization. In this article, polynomial regression (PR), radial basis function artificial neural network (RBFANN), and kriging methods were compared for building surrogate models of a multiphase flow simulation model in a simplified nitrobenzene contaminated aquifer remediation problem. In the model accuracy analysis process, a 10-fold cross validation method was adopted to evaluate the approximation accuracy of the three surrogate models. The results demonstrated that: RBFANN surrogate model and kriging surrogate model had acceptable approximation accuracy, and further that kriging model's approximation accuracy was slightly higher than RBFANN model. However, the PR model demonstrated unacceptably poor approximation accuracy. Therefore, the RBFANN and kriging surrogates were selected and used in the optimization process to identify the most cost-effective remediation strategy at a nitrobenzene-contaminated site. The optimal remediation costs obtained with the two surrogate-based optimization models were similar, and had similar computational burden. These two surrogate-based optimization models are efficient tools for optimal groundwater remediation strategy identification.

1. Introduction

Groundwater contamination problem arises along with the rapid development of industry and agriculture. Since groundwater remediation is a time consuming and costly process, finding methods to increase the remediation efficiency and reduce the remediation cost gradually becomes a crucial problem. Simulation and optimization technique is an effective tool to solve this problem (Ahlfeld *et al.* 1988; Guan and Aral 1999; Liu *et al.* 2000; Schaerlaekens *et al.* 2006; Md Azamathulla *et al.* 2008). However, the enormous computational cost of running such simulations multiple times, limits the applicability of the simulation optimization techniques in a complex groundwater remediation optimization process (Qin *et al.* 2007; Razavi *et al.* 2012). One method that reduces this computational burden is replacing the numerical models with efficient surrogate models (Sreekanth and Datta 2010; Jin *et al.* 2001).

Surrogate models, also called metamodels or response surface models, are used as particular substitutes for the complex numerical models, while being computationally cheaper to evaluate (Blanning 1975; Kourakos and Mantoglou 2013). Polynomial regression (PR), artificial neural network (ANN), kriging, support vector machine, and multivariate adaptive regression spline, etc., are common methods to build surrogate models, and these surrogate models have been widely used in space approximation problems (Jin *et al.* 2001;

Keywords. Surrogate modelling; simulation optimization; groundwater remediation; polynomial regression; radial basis function artificial neural network; kriging.

Giannakoglou 2002; Jin 2005; Forrester and Keane 2009).

To improve the computation efficiency of an optimization process, surrogate models have been used to approximate the computational simulation model in groundwater simulation optimization field in recent years. Huang *et al.* (2003), Qin *et al.* (2007), He *et al.* (2008), and Fen *et al.* (2009) used PR surrogate models to improve the optimization efficiency in contaminated groundwater remediation system. Rogers *et al.* (1995), Morshed and Kaluarachchi (1998), Johnson and Rogers (2000), Arndt *et al.* (2005), Yan and Minsker (2006), Nikolos *et al.* (2008), Behzadian *et al.* (2009), Dhar and Datta (2009), Kourakos and Mantoglou (2009), Yan and Minsker (2011) and Papadopoulou *et al.* (2010) used artificial neural network surrogate models in optimal groundwater remediation strategy identification, groundwater engineering facility optimization, optimal water supply design, and sea water intrusion management problems. Hemker *et al.* (2008) used kriging method to build the surrogate model of simulation model to reduce optimization computation cost in groundwater management problem.

It is difficult to say if one of these surrogate modelling methods is generally superior to others. For any specific engineering optimization design problem, conducting a comprehensive comparison analysis of the surrogate models that are built with different methods, and selecting the proper one to be used in the optimization process is of great importance. Mirfendereski and Mousavi (2011) compared support vector machines and polynomial-based surrogate models to approximate the MODSIM river basin simulation model, and applied it in Atrak river basin water allocation problem. Shyy *et al.* (2001) compared the relative performance between polynomials and neural networks surrogate models, and applied them on aerodynamics and rocket propulsion components. Simpson *et al.* (1998) compared the polynomial-based response surface and kriging surrogates in aerodynamic design optimization of hypersonic spiked blunt bodies. However, the comparisons of different surrogate modelling methods are limited in groundwater remediation optimization field.

During the model validation and selection process, the commonly used method is dividing the data into two mutually exclusive subsets called the training set and the validation set, which is called the holdout method (Kohavi 1995). This method only uses part of the data to train the surrogate model and uses the rest of data to validate the surrogate model (Namura *et al.* 2012), which may result in overfitting of the training data, and underfitting of the other data. Cross validation is an improvement of holdout method because it uses all data for both training and validation. In groundwater optimization field, cross validation is rarely used for surrogate model accuracy estimation (Razavi *et al.* 2012).

As an extension of previous researches, this study attempts to develop an optimization process based on multi-surrogate models and cross validation method for identifying the optimal remediation strategy at a nonaqueous phase liquids (NAPLs) contaminated aquifer. This objective entails the following tasks:

- build a multiphase flow simulation model in a nitrobenzene contaminated aquifer;
- develop surrogate models of multiphase flow simulation model using PR, RBFANN, and kriging methods, and estimate the accuracy of different surrogate models with cross validation method;
- surrogate models with acceptable accuracy are then selected and used in the nonlinear optimization model for identifying the most cost effective remediation strategy.

The novelty of the paper is:

- different surrogate modelling methods were used and compared in groundwater remediation optimization field;
- cross validation method was used to estimate the accuracy of different surrogate models in groundwater optimization field.

2. Methods

2.1 *Surrogate modelling method*

Polynomial regression is the simplest approximation method to build surrogate models (Forrester and Keane 2009). The most widely used polynomial regression model is the second-order polynomial model which has the following form (Jin 2005):

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \sum_{j \geq i}^{n} \beta_{ij} x_i x_j + \cdots \qquad (1)$$

where β_0, β_i, β_{ii}, and β_{ij} are the regression coefficients, n is the number of variables, x_i and x_j are the variables. Using least square method (LSM), the regression coefficients can be solved.

RBFANN is a 3-layer feed forward neural network consisting of an input layer, a hidden layer, and an output layer (Shen *et al.* 2010).

$\mathbf{X}$ is an N dimensional input vector. The output of the neurons in the RBFANN hidden layer is assumed as:

$$q_i = \Phi\left(\|\mathbf{X} - \mathbf{c}_i\|\right) \qquad (2)$$

where $\mathbf{c}_i$ is the center associated with the ith neuron in the radial basis function hidden layer, $i = 1, 2, \ldots, H$, where H is the number of hidden units, $\|\mathbf{X} - \mathbf{c}_i\|$ is the norm of $\mathbf{X} - \mathbf{c}_i$, $\Phi(\cdot)$ is a radial basis function (Chen *et al.* 1991; Baddari *et al.* 2009). Outputs of the kth neuron in RBFANN output layer are linear combinations of the hidden layer neuron outputs as:

$$y_k = \sum_{i=1}^{H} w_{ki} q_i - \theta_k \quad (k = 1, 2, \ldots, M) \tag{3}$$

where w_{ki} is the connecting weights from the ith hidden layer neuron to the kth output layer, θ_k is the threshold value of the kth output layer neuron.

The kriging method was developed by the French mathematician Georges Matheron based on the Master's thesis of Daniel Gerhardus Krige (Matheron 1963), it was first used as a geostatistical method. Sacks *et al.* (1989) firstly introduced kriging method as a surrogate modelling method, in the paper of Sacks *et al.* (1989), kriging surrogate model was also called design and analysis of computer experiment (DACE). From that time, many researchers have used kriging method for surrogate modelling (Booker *et al.* 1998; Simpson *et al.* 2001; Ryu *et al.* 2002; Hemker *et al.* 2008; Coetzee *et al.* 2012).

The kriging model is a combination of two components (Queipo *et al.* 2005): deterministic functions and localized deviations.

$$Y(x) = \sum_{i=1}^{k} f_i(x)\,\beta_i + z(x) \tag{4}$$

where $\sum_{i=1}^{k} f_i(x)\,\beta_i$ is the term of deterministic functions, β_i are coefficients of deterministic functions, $f_i(x)$ are k known regression functions, which are usually polynomial functions. $z(x)$ is term of localized deviations with mean zero, variance σ^2, and covariance expressed as:

$$\mathrm{Cov}\left[z(x_i), z(x_j)\right] = \sigma^2 R(x_i, x_j) \tag{5}$$

where $R(x_i, x_j)$ is the correlation function between any two of the n_s samples The common types of correlation functions are linear function, exponential function, Gauss function, spline function, etc. (Ryu *et al.* 2002).

The prediction of unsampled points response $y(x)$ can be expressed as:

$$\widehat{y}(x) = f(x)^{\mathrm{T}}\beta + r^{\mathrm{T}}R^{-1}(Y - F\beta) \tag{6}$$

where Y is the vector of n_s samples response, r is the correlation vector between samples and prediction points.

$$r = [R(x, x_1), R(x, x_2), \ldots, R(x, x_{n_x})]^{\mathrm{T}}, \tag{7}$$

$$F = [f(x_1) \cdots f(x_{n_x})]^{\mathrm{T}}. \tag{8}$$

2.2 *Cross validation – an accuracy estimating method*

Cross validation is a technique for estimating the generalization errors of a predictive model. In cross validation process, all available data can be used both for validation and training, which helps avoid overfitting of the training data (Cheng and Pecht 2012). In k-fold cross validation, the data are divided into k subsets of approximately equal size. The surrogate model is built k times, each time leaving out one of the subsets as the validation data for validating the model, and using the remaining k–1 subsets for training. Total error of k times prediction is averaged to assess the approximation accuracy of surrogate models (Jiawei and Kamber 2001).

3. Case study

3.1 *Site overview*

To evaluate the advantages and disadvantages of different surrogate models of groundwater simulation model, three different surrogate models (PR model, RBFANN model, and kriging model) were applied to a test aquifer contaminated by nitrobenzene. The contaminated site is located in the second terrace of a valley alluvial plain in the lower Songhua River. The contaminated site is flat with an average altitude of 193 m. The upper part of the soil consists of an upper Pleistocene silt and silty clay with a thickness of 1–2 m, while the lower part is made up of medium sand and gravel, with a thickness of about 15 m. The main recharge sources are precipitation and runoff, while the main discharge source is runoff. Groundwater flows from northeast to southwest. The objective simulation layer is pore phreatic water in loose rock mass, the buried depth is about 4 m, and the single well yield is about 500–1000 m^3/d. The study area and initial contaminant plume are shown in figure 1.

Based on the contaminant distribution, a surfactant enhanced aquifer remediation (SEAR) with

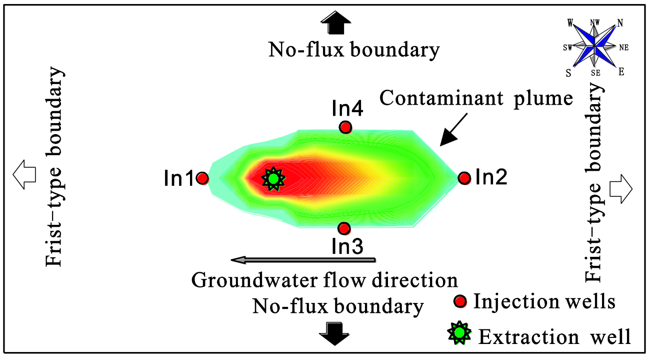

Figure 1. Contaminant conditions, and injection and extraction wells' conditions.

sodium lauryl sulfate as surfactant was designed with four injection wells and one extraction well (figure 1). 10% surfactant solution (volume fraction) was injected into the injection wells. To maintain hydraulic balance, the total extraction rates and injection rates were equal.

The optimization objective was to identify the most cost-effective strategy which can satisfy:

- more than 60% of the contaminant is removed;
- injection rate of each well is smaller than 70 m^3/d; and the remediation duration is smaller than 20 days.

3.2 *Numerical simulation model developed*

The simulation domain was generalized as a heterogeneous and anisotropic 3-D multiphase flow and transport model. A first-type boundary condition was assigned at the northeast and southwest boundaries of the site. The other boundaries were no-flux boundaries. The simulation domain was discretized into 17 vertical layers, and each layer further discretized into 35 × 19 grids. Each grid dimension was 3 m × 3 m × 1 m in the x, y, and z axes directions respectively. The physical and chemical parameters of the site are presented in table 1.

A three-dimensional mathematical model was built to evaluate the efficiency of SEAR strategies. The basis mass conservation equation for each component can be written as (Delshad *et al.* 1996):

$$\frac{\partial(\phi\tilde{C}_k\rho_k)}{\partial t}+\vec{\nabla}\left[\sum_{l=1}^{3}\rho_k(C_{kl}\vec{v}_l-\phi S_l\vec{\vec{K}}_{kl}\vec{\nabla}C_{kl})\right]=R_k \tag{9}$$

Table 1. *Physical and chemical parameters in simulation model.*

Parameter	Value
Porosity	0.34
Permeability	$5.9\times10^{-11}\mathrm{m}^2$
Longitudinal dispersivity	1 m
Transverse dispersivity	0.3 m
Hydraulic gradient	0.001
Water density	1000 kg/m^3
Nitrobenzene density	1205 kg/m^3
Surfactant density	1090 kg/m^3
Water viscosity	0.001 Pa s
Nitrobenzene viscosity	0.00168 Pa s
Nitrobenzene/water interfacial tension	0.02566 N/m
Nitrobenzene solubility in water	1.9 kg/m^3
Residual water saturation	0.24
Residual nitrobenzene saturation	0.17

where k is component index, including water ($k=1$), oil ($k=2$) and surfactant ($k=3$), l is phase index including water ($l=1$), oil ($l=2$) and microemulsion ($l=3$) phases, ϕ is porosity, $\tilde{C}_k$ is overall concentration of component k (volume fraction), ρ_k is density of component k (kg/m^3), C_{kl} is concentration of component k in phase l (volume fraction), $\vec{v}_l$ is Darcy velocity of phase l (m/s), S_l is saturation of phase l, $\vec{\vec{K}}_{kl}$ is dispersion tensor (m^2/s), R_k is total source/sink term for component k (kg/m^3 s).

The mathematical model was constructed with the above mass conservation equation and corresponding initial conditions and boundary conditions. University of Texas Chemical Compositional Simulator (UTCHEM) was used to solve the mathematical model. UTCHEM is a three-dimensional, multiphase, multicomponent finite difference numerical simulator (Delshad *et al.* 1996; Bhattarai 2006). The simulator was originally developed by Pope and Nelson (1978) to simulate the enhanced recovery of oil using surfactant and polymer processes, and then was modified to simulate the remediation process of aquifers contaminated by NAPLs (Delshad *et al.* 1996; Delshad 1997; Qin *et al.* 2007).

3.3 *Surrogate model developed*

There are many factors that influence the remediation efficiency and remediation costs. In this study, we chose the well rates and remediation duration as the input variables. Due to the assumption that total extraction rates were equal to the total injection rates and there was only one extraction well, there were five input variables, which were remediation duration, rates of injection well In1, In2, In3, and In4 (table 2). The output variable was average contaminant removal rate.

Forty input samples were collected through sampling in the feasible region of input variables of multiphase flow numerical simulation model, and the output responses were obtained through running the developed simulation model.

With these 40 input–output data, PR, RBFANN, and kriging methods were used separately to build three surrogate models of multiphase flow numerical simulation model. Ten-fold cross validation method was adopted to evaluate the approximation accuracy of the three surrogate models.

In the 10-fold cross validation process, the 40 input–output data were randomly divided into 10 subsets, with each subset containing four samples. The surrogate model was built 10 times, with each time the surrogate model had 36 training data and four validation data. For PR model, first-order, second-order, and third-order polynomials were

Table 2. *Input variables and its value range.*

Input variables	Remediation duration (d)	Rate of well In1 (m^3/d)	Rate of well In2 (m^3/d)	Rates of well In3 (m^3/d)	Rates of well In4 (m^3/d)
Range	0–20	0–70	0–70	0–70	0–70

adopted to build the relationship between average contaminant removal rate, remediation duration, and four injection rates for each fold, using least square method (LSM), a set of regression coefficients were solved, and a total of 10 different polynomial regression models were obtained.

For RBFANN model, the input layer of the network represented the remediation duration and four injection rates, a total of five neurons. The output layer of the network contained only one neuron, which represented the average contaminant removal rate. The hidden neuron number is set as 10, 20, 30, and 40. In this study, the Gauss function was used as transfer function, and the orthogonal least square method was used for network training. After the RBFANN model training, a total of 10 RBFANN models with different parameters were obtained.

For kriging model, polynomial functions of orders 0, 1, and 2 were used as regression functions, while Gauss function was used as correlation function. Through training the kriging model, a total of 10 kriging models with different parameters were also obtained.

3.4 *Optimization model developed*

To identify the optimal remediation strategy, a nonlinear optimization model was developed using the minimal remediation cost as the objective function, with the remediation duration and rates of injection well In1, In2, In3, and In4 as the decision variables. The optimization model can be represented as follows:

$$\begin{aligned}\min f\,(Q,t) &= f_{\text{installation}} + f_{\text{operation}} \\ &= C_1 m + C_2 n + C_3 t\sum_{i=1}^{m} Q_i^{\text{In}} + C_4 t \sum_{j=1}^{n} Q_j^{\text{Ex}} \quad (10\text{a})\end{aligned}$$

Subject to:

$$0 \le Q_i^{\text{In}} \le Q_M^{\text{In}} \quad (10\text{b})$$

$$0 \le Q_j^{\text{Ex}} \le Q_M^{\text{Ex}} \quad (10\text{c})$$

$$\sum_{i=1}^{m} Q_i^{\text{In}} = \sum_{j=1}^{n} Q_j^{\text{Ex}} \quad (10\text{d})$$

$$0 \le t \le t_M \quad (10\text{e})$$

$$g(Q,t) \ge g_0 \quad (10\text{f})$$

where equation (10a) is the objective function, equation 10(b–d) are the injection and extraction rate constraints, equation (10e) is the remediation duration constraint, equation (10f) is the remediation efficiency constraint. f is total cost of the remediation system ($), the first two terms of equation (10a) account for the installation cost and the last two terms account for the operation cost, C_1 and C_2 are injection wells and extraction wells installation cost coefficients ($), respectively, C_3 and C_4 are injection and extraction operation cost coefficients ($\$/m^3$) respectively, m and n are injection and extraction wells number respectively, Q_i^{In} is the rate of ith injection well (m^3/d), Q_j^{Ex} is the rate of jth extraction well (m^3/d), Q_M^{In} and Q_M^{Ex} are maximum allowable injection rate and extraction rate of wells (m^3/d), t is the remediation duration (d), t_M is the maximum allowable remediation duration (d), $g(Q,t)$ is the average contaminant removal rate, which is an output response of the surrogate model, and g_0 is the minimum allowable value of the contaminant average removal rate. The constant of the equation is in table 3.

4. Results and discussion

4.1 *Surrogate model accuracy analysis*

For each surrogate modelling method, there are 10 folds, and in each fold, the output responses of the four validation samples were predicted with the developed surrogate models. Therefore, 40 samples' output responses can be obtained with surrogate models.

In this study, absolute error (AE) and relative error (RE) were selected as the loss function to estimate the accuracy of the surrogate models. Figures 2, 3, and 4 show the boxplots of absolute and relative error of different surrogate models. The results demonstrated that: for PR model, approximation accuracy of second order polynomial is higher than that of first-order polynomial and third-order polynomial; for RBFANN model, the RBFANN with 40 hidden neurons obtained

Table 3. *Constants included in the optimization model.*

Constant	C_1 ($)	C_2 ($)	C_3 ($/m^3)	C_4 ($/m^3)	Q_M^{In} (m^3/d)	m	n	t_{M}(day)	g_0
Value	500	500	0.3	0.005	70	4	1	20	0.60

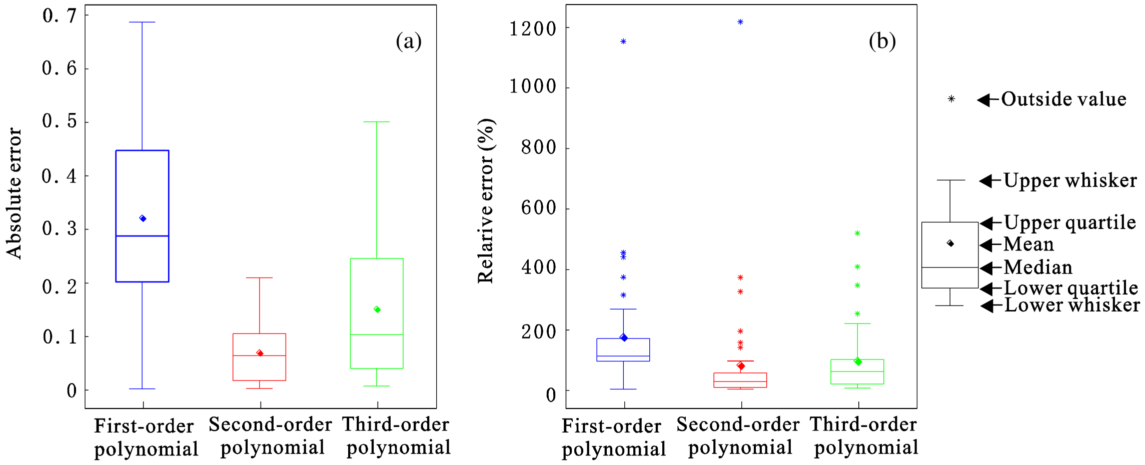

Figure 2. Boxplots of the PR models: **(a)** boxplot of absolute errors and **(b)** boxplot of relative errors.

highest approximation accuracy; for kriging model, kriging model with second order polynomial function as regression function obtained the highest approximation accuracy. Therefore, second-order polynomial model, RBFANN model with 40 hidden neurons, and kriging model with second order polynomial function as regression function are selected as the surrogates, and their parameters are in table A1, table A2, and table A3 (Appendix).

The relationship between simulation model results and surrogate model results of 40 validation samples are shown in figure 5, which shows that the accuracy of kriging and RBFANN models are greater than PR model. From the mean error (mean AE and mean RE) and maximum error (maximum AE and maximum RE) of the three surrogate models (figures 2, 3, and 4), we can conclude that the RBFANN model and kriging model had acceptable approximation accuracy, and further that the approximation accuracy of kriging model was slightly higher than that of RBFANN model. However, PR model's approximation accuracy was unacceptable (the mean relative error is 80%), this probably due to its limited fitting ability for nonlinear problem, especially for the high-order nonlinear problem. From the distribution of surrogate models we can conclude that the distribution of RBFANN model and kriging model (most of the relative error values are between 3% and 18%) are much more concentrated than that of PR model (there are many samples with relative error greater than 100%). In summary, RBFANN model and kriging model had acceptable accuracy and robustness, and can be used in the flowing optimization process.

4.2 *Optimization result analysis*

From the above observations, both the RBFANN model and kriging model were embedded in the optimization model, as the linking of the average contaminant removal rate, injection rates, and remediation duration. The genetic algorithm was adopted to solve the developed nonlinear optimization model on MATLB platform. In the genetic algorithm searching process, the surrogate model was invoked instead of the computational simulation model. The parameters of the genetic algorithm were set the same in the RBFANN surrogate based optimization model and in kriging surrogate based optimization model. Selection probability, crossover probability, and mutation probability are usually set between 0.7–1.0, 0.7–1.0, 0.01–0.05 (Simpson *et al.* 1994), and in this paper they are set as 0.9, 0.7 and 0.05; the generation number is set as 100. The population size is set as 500, and the obtained optimal remediation strategies are in table 4. The optimal remediation strategies obtained with RBFANN and kriging were evaluated using the multiphase flow numerical simulation model, and the predicted average contaminant removal rates were 0.6 and 0.6, which satisfies the contaminant removal rate constraint. The optimal solutions obtained with the two optimization models were different, but the optimal remediation costs were similar, this may be because the complex

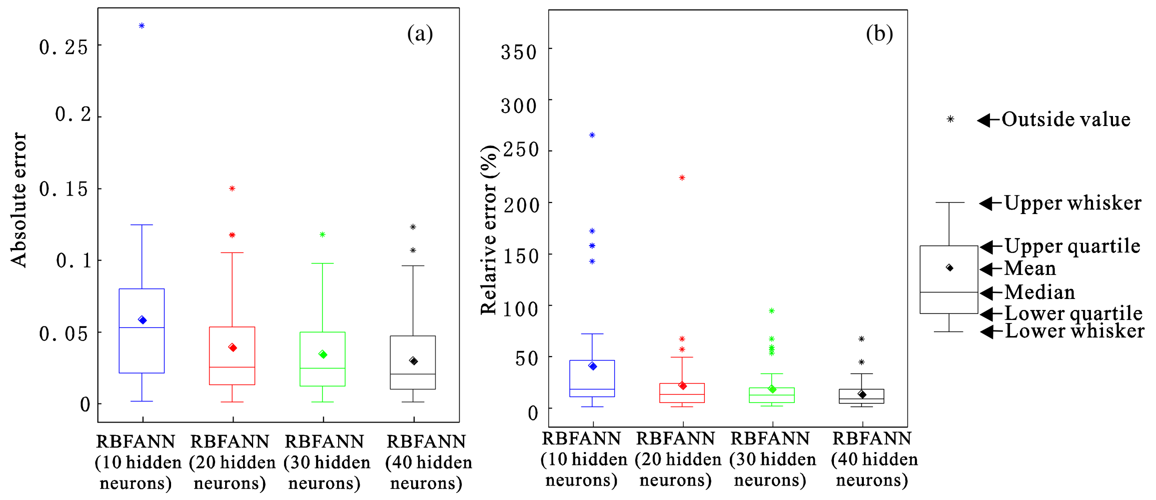

Figure 3. Boxplots of the RBFANN models: **(a)** boxplot of absolute errors and **(b)** boxplot of relative errors.

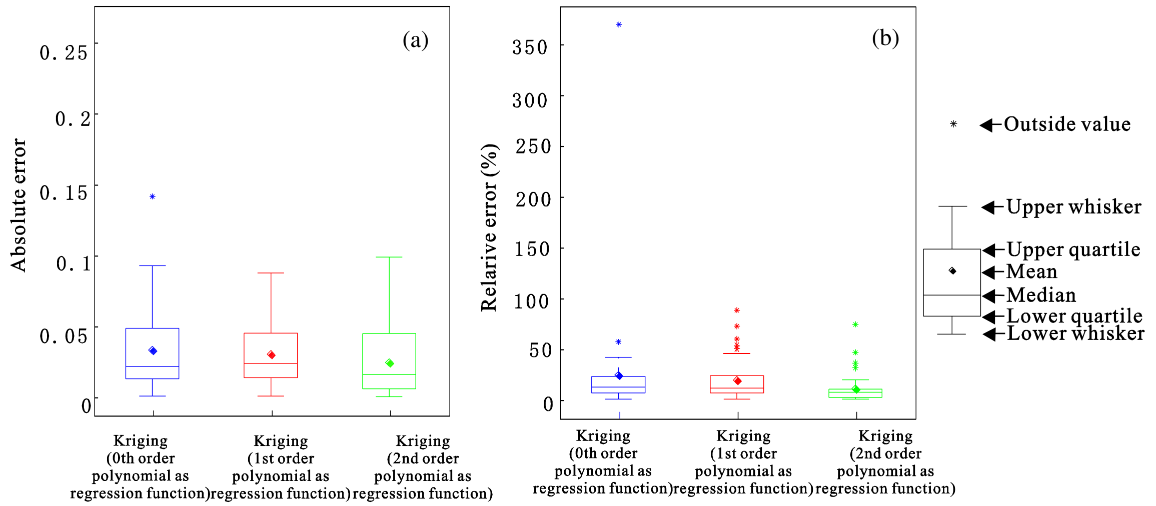

Figure 4. Boxplots of the kriging models: **(a)** boxplot of absolute errors and **(b)** boxplot of relative errors.

optimization problem had multioptimal solution. We can conclude that both the RBFANN and kriging surrogate-based optimization models obtained satisfactory solutions.

4.3 *Computational burden analysis*

Generally, there are three parts for the computational burden in the surrogate-based optimization process: repeated running of the numerical simulation model, surrogate model construction, and the optimization searching process with genetic algorithm.

The main computational burden was resulted from the repeated running of the numerical simulation model. The SEAR optimization for the nitrobenzene-contaminated site required 295 seconds of CPU time to run every simulation model on a 3.0 GHz AMD CPU and 2 GB RAM PC platform. 40 input data were sampled randomly,

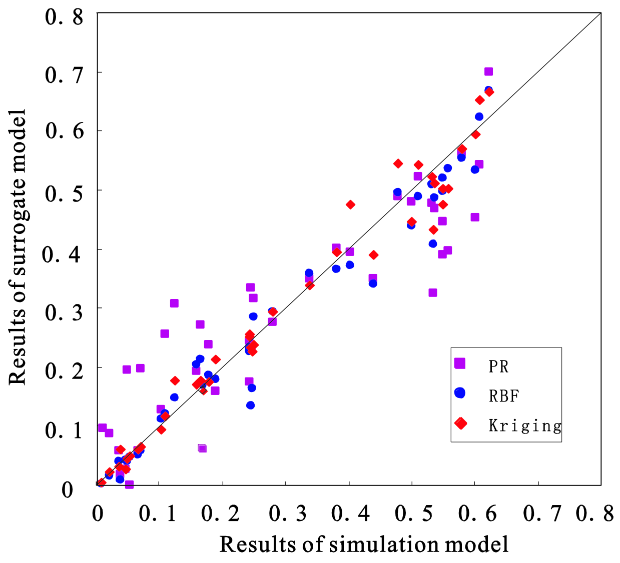

Figure 5. Simulation model results *vs.* surrogate model results.

Table 4. *Optimal remediation strategy.*

Optimal remediation duration (d)	Optimal wells rates (m^3/d)			Optimal cost ($)	Contaminant average rate	Convergence duration (s)
RBFANN surrogate-based optimization model						
17.97	Injection wells	In1	22.97	3330.79	0.60	2.61
		In2	32.51			
		In3	35.64			
		In4	60.46			
	Extraction well		151.58			
Kriging surrogate-based optimization model						
17.74	Injection wells	In1	13.43	3304.35	0.60	2.81
		In2	22.21			
		In3	45.03			
		In4	67.98			
	Extraction well		148.64			

and the output responses were obtained with the simulation model, so the simulation model needed to be run 40 times; thus, 11,800 s were required in this process with both of this two optimization models.

An average of 1.3 s were needed to train the RBFANN model for one time, while an average of 0.12 s were needed to train the kriging model. In the cross validation process, each surrogate model needed to be constructed 10 times, so the construction of RBFANN model and kriging model totally needed 13 and 1.2 s respectively.

In the optimization searching process, RBFANN surrogate-based optimization model needed 2.61 s before the genetic algorithm converged, while the kriging surrogate-based optimization needed 2.81 s.

Compared with the computational burden of repeated running of the numerical simulation model, the computational burden resulted from the surrogate model construct and the optimization searching process can be negligible. The whole process of surrogate-based optimization model solving needed nearly 3 hr (11,800 s), no matter which surrogate model was used. In GA process, 4000 evaluations were used as a termination criterion (maximum evaluation times). Therefore if the numerical model was used instead of the surrogate model, then the total CPU time would have been 1180,000 s (14 days).

5. Conclusions

In this study, three different surrogate models: polynomial regression, radial basis function artificial neural network, and kriging were used to build surrogate model at a nitrobenzene contaminated aquifer remediation problem.

Ten-fold cross validation was adopted to compare the approximation accuracy of the three surrogate models. The results showed that the radial basis function artificial neural network and kriging models had better approximation accuracy and robustness than the polynomial regression model. Therefore, the radial basis function artificial neural network and kriging-based optimization models were preferred and selected to identify the optimal remediation strategy for a nitrobenzene contaminated site. The two surrogate-based optimization models obtained similar optimal costs, with a similar computational burden. In addition, these two surrogate-based optimization models considerably reduced the computational burden compared with the conventional simulation optimization model. Therefore, we can conclude that the surrogate-based optimization models are efficient tools for optimal groundwater remediation strategy identification, and radial basis function artificial neural network method and kriging method are effective surrogate modelling methods.

Acknowledgements

This research was supported by the Nature Science Foundation of China (Nos. 41072171, 41372237) and the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20120061110058). Authors thank the editor and the anonymous reviewers for their insightful comments and suggestions.

Appendix

Table A1. *Parameters of second-order polynomial models.*

Parameters	1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
β_0	−0.129402	−0.115985	−0.105305	−0.11044	−0.145679	−0.139754	−0.134486	−0.129165	−0.144638	−0.147624
β_1	0.023804	0.011959	0.033544	0.026919	0.022834	0.026457	0.018961	0.020100	0.024650	0.020406
β_2	0.003693	0.006405	0.001935	0.002426	0.003917	0.004974	0.005205	0.003924	0.004280	0.003375
β_3	0.003435	0.001851	0.002152	0.003996	0.002386	0.002213	0.002214	0.002663	0.002556	0.003253
β_4	0.003653	0.003708	0.002383	0.006121	0.005115	0.004029	0.005699	0.004117	0.005080	0.005148
β_5	−0.003354	−0.002855	−0.001503	−0.006334	−0.003217	−0.004118	−0.003554	−0.003036	−0.002848	−0.003093
β_{11}	−0.000429	0.000085	−0.000713	−0.000416	−0.000255	−0.000453	−0.000216	−0.000134	−0.000490	−0.000181
β_{22}	−0.000055	−0.000072	−0.000047	−0.000055	−0.000061	−0.000055	−0.000074	−0.000063	−0.000084	−0.000056
β_{33}	−0.000065	−0.000023	−0.000044	−0.000065	−0.000038	−0.000038	−0.000044	−0.000044	−0.000045	−0.000039
β_{44}	−0.000035	−0.000027	−0.000028	−0.000057	−0.000047	−0.000041	−0.000054	−0.000036	−0.000053	−0.000048
β_{55}	0.000029	0.000030	0.000012	0.000050	0.000030	0.000037	0.000027	0.000052	0.000025	0.000029
β_{12}	0.000040	0.000031	0.000005	0.000061	0.000057	0.000056	0.000050	0.000058	0.000130	0.000084
β_{13}	0.000001	0.000049	0.000081	0.000011	0.000007	−0.000004	0.000065	−0.000016	0.000017	−0.000013
β_{14}	0.000138	0.000093	0.000058	0.000066	0.000071	0.000109	0.000055	0.000136	0.000078	0.000113
β_{15}	0.000192	0.000255	0.000145	0.000138	0.000167	0.000168	0.000216	0.000105	0.000163	0.000176
β_{23}	0.000056	0.000004	0.000033	0.000035	0.000031	0.000029	0.000020	0.000037	0.000036	0.000023
β_{24}	−0.000026	−0.000023	0.000034	0.000003	−0.000003	−0.000022	0.000003	0.000002	0.000006	0.000005
β_{25}	0.000017	−0.000001	0.000005	0.000029	0.000015	0.000000	0.000016	0.000009	0.000016	0.000015
β_{34}	−0.000013	−0.000010	−0.000035	−0.000043	−0.000025	−0.000021	−0.000024	−0.000030	−0.000023	−0.000040
β_{35}	−0.000013	−0.000004	0.000011	0.000040	0.000015	0.000020	0.000019	0.000021	0.000013	0.000013
β_{45}	0.000017	−0.000010	0.000007	0.000005	0.000001	0.000017	−0.000006	−0.000021	0.000000	−0.000003

Table A2. *Parameters of RBFANN models.*

Parameters	1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
Number of hidden neurons	40									
Training mean squared error goal	0.04	0.06	0.06	0.01	0.1	0.045	0.1	0.04	0.043	0.096
Spreed	1000	90	58	50	540	184	407	116	115	110

Table A3. *Parameters of kriging models.*

Parameters	1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
θ_1	6.9833	2.7224	2.1806	0.4292	2.6918	2.7602	0.4845	0.7812	0.5298	0.6500
θ_2	0.0181	0.0015	0.1106	0.1703	0.0020	0.0209	0.2255	0.1016	0.1319	0.1811
θ_3	0.0010	0.0010	0.0145	0.0024	0.0024	0.0018	0.0028	0.0171	0.0149	0.0125
θ_4	0.0163	0.0019	0.0321	0.0573	0.0026	0.0099	0.0474	0.1840	0.1456	0.1346
θ_5	0.0014	0.0018	0.1160	0.1703	0.0028	0.0078	0.1518	0.1016	0.1319	0.2208
β_1	0.0347	−0.1076	−0.0355	−0.1176	−0.0811	−0.0622	−0.0262	−0.0320	−0.0963	−0.0233
β_2	0.4837	0.2687	0.6422	0.4943	0.3350	0.5606	0.5182	0.5760	0.5432	0.5789
β_3	0.1128	0.1163	0.0876	0.0597	0.1217	0.1030	0.1465	0.1491	0.0798	0.0982
β_4	0.0756	0.0601	0.1085	0.1389	0.0495	0.0716	0.1327	0.1026	0.1287	0.1120
β_5	0.1911	0.1562	0.1933	0.2155	0.1668	0.1619	0.1921	0.1717	0.2186	0.2067
β_6	0.1800	0.1538	0.1257	0.0917	0.1610	0.1437	0.0905	0.1143	0.0943	0.0910
β_7	0.1864	0.2936	0.0715	0.0909	0.1633	0.1137	0.0413	0.0724	0.0541	0.0549
β_8	−0.0400	−0.0674	−0.0258	−0.0148	−0.0703	−0.0252	−0.0475	−0.0157	−0.0108	−0.0310
β_9	−0.0727	−0.0499	−0.0330	−0.0197	−0.0333	−0.0348	−0.0275	−0.0156	−0.0102	−0.0252
β_{10}	−0.1174	−0.0932	−0.0121	−0.0106	−0.0679	−0.0311	−0.0167	−0.0346	−0.0307	−0.0159
β_{11}	0.0466	0.0329	−0.0130	0.0022	0.0062	0.0011	−0.0085	−0.0185	0.0161	−0.0155

References

Ahlfeld D P, Mulvey J M and Pinder G F 1988 Contaminated groundwater remediation design using simulation, optimization, and sensitivity theory 2. Analysis of a field site; *Water Resour. Res.* **24** 443–452.

Arndt O, Barth T, Freisleben B and Grauer M 2005 Approximating a finite element model by neural network prediction for facility optimization in groundwater engineering; *European J. Oper. Res.* **166** 769–781.

Baddari K, Aïfa T, Djarfour N and Ferahtia J 2009 Application of a radial basis function artificial neural network to seismic data inversion; *Comput. Geosci.* **35** 2338–2344.

Behzadian K, Kapelan Z, Savic D and Ardeshir A 2009 Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks; *Environ. Modell. Softw.* **24** 530–541.

Bhattarai M 2006 *A numerical modeling study of surfactant enhanced mobilization of residual LNAPL using UTCHEM*; Southern Illinois University Carbondale, Carbondale, USA.

Blanning R W 1975 The construction and implementation of metamodels; *Simulation* **24** 177–184.

Booker A J, Frank P D, Dennis J Jr, Moore D W and Serafini D B 1998 Managing surrogate objectives to optimize a helicopter rotor design – further experiments; 8th AIAA/ISSMO symposium on multidisciplinary analysis and optimization, St. Louis.

Chen S, Cowan C F N and Grant P M 1991 Orthogonal least squares learning algorithm for radial basis function networks; *IEEE Trans Neural Networks*, pp. 302–309.

Cheng S and Pecht M 2012 Using cross-validation for model parameter selection of sequential probability ratio test; *Expert Syst. Appl.* **39** 8467–8473.

Coetzee W, Coetzer R L and Rawatlal R 2012 Response surface strategies in constructing statistical bubble flow models for the development of a novel bubble column simulation approach; *Comput. Chem. Eng.* **36** 22–34.

Delshad M 1997 UTCHEM Version 6.1, Technical documentation; Center for Petroleum and Geosystems Engineering.

Delshad M, Pope G and Sepehrnoori K 1996 A compositional simulator for modeling surfactant enhanced aquifer remediation, 1 formulation; *J. Contam. Hydrol.* **23** 303–327.

Dhar A and Datta B 2009 Saltwater intrusion management of coastal aquifers. I: Linked simulation-optimization; *J. Hydrol. Eng.* **14** 1263–1272.

Fen C S, Chan C and Cheng H C 2009 Assessing a response surface-based optimization approach for soil vapor extraction system design; *J. Water. Resour. Plan. Manag.* **135** 198–207.

Forrester A I J and Keane A J 2009 Recent advances in surrogate-based optimization; *Prog. Aerosp. Sci.* **45** 50–79.

Giannakoglou K C 2002 Design of optimal aerodynamic shapes using stochastic optimization methods and computational intelligence; *Prog. Aerosp. Sci.* **38** 43–76.

Guan J and Aral M 1999 Optimal remediation with well locations and pumping rates selected as continuous decision variables; *J. Hydrol.* **221** 20–42.

He L, Huang G H, Zeng G M and Lu H W 2008 An integrated simulation, inference, and optimization method for identifying groundwater remediation strategies at petroleum-contaminated aquifers in western Canada; *Water Res.* **42** 2629–2639.

Hemker T, Fowler K R, Farthing M W and von Stryk O 2008 A mixed-integer simulation-based optimization approach with surrogate functions in water resources management; *Optim. Eng.* **9** 341–360.

Huang Y, Li J, Huang G, Chakma A and Qin X 2003 Integrated simulation-optimization approach for real-time dynamic modeling and process control of surfactant-enhanced remediation at petroleum-contaminated sites; *Pract. Period Hazard Toxic Radioact. Waste Manag. (ASCE)* **7** 95–105.

Jiawei H and Kamber M 2001 *Data mining: Concepts and techniques*; Morgan Kaufmann, San Francisco, 744p.

Jin Y 2005 A comprehensive survey of fitness approximation in evolutionary computation; *Soft. Comput.* **9** 3–12.

Jin R, Chen W and Simpson T W 2001 Comparative studies of metamodelling techniques under multiple modelling criteria; *Struct. Multidiscip. O.* **23** 1–13.

Johnson V M and Rogers L L 2000 Accuracy of neural network approximators in simulation-optimization; *J. Water Resour. Plan Manag.* **126** 48–65.

Kohavi R 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection; In: Proc. Fourteenth International Joint Conference on Artificial Intelligence (San Francisco: Morgan Kaufmann), pp. 1137–1143.

Kourakos G and Mantoglou A 2009 Pumping optimization of coastal aquifers based on evolutionary algorithms and surrogate modular neural network models; *Adv. Water Resour.* **32** 507–521.

Kourakos G and Mantoglou A 2013 Development of a multi-objective optimization algorithm using surrogate models for coastal aquifer management; *J. Hydrol.* **479** 13–23.

Liu W H, Medina M A Jr, Thomann W, Piver W T and Jacobs T L 2000 Optimization of intermittent pumping schedules for aquifer remediation using a genetic algorithm; *J. Am. Leather Chem. As.* **36** 1335–1348.

Matheron G 1963 Principles of geostatistics; *Econ. Geol.* **58** 1246–1266.

Md Azamathulla H, Wu F-C, Ghani A A, Narulkar S M, Zakaria N A and Chang C K 2008 Comparison between genetic algorithm and linear programming approach for real time operation; *J. Hydrol. Environ. Res.* **2** 172–181.

Mirfendereski G and Mousavi S J 2011 Comparison of support vector machines and response surface models in meta-modeling applied in basin-scale optimum water allocation; 19th International Congress on Modelling and Simulation, Perth.

Morshed J and Kaluarachchi J J 1998 Application of artificial neural network and genetic algorithm in flow and transport simulations; *Adv. Water Resour.* **22** 145–158.

Namura N, Shimoyama K, Jeong S and Obayashi S 2012 Kriging/RBF-hybrid response surface methodology for highly nonlinear functions; *J. Comput. Sci. Tech.* **6** 81–96.

Nikolos I K, Stergiadi M, Papadopoulou M P and Karatzas G P 2008 Artificial neural networks as an alternative approach to groundwater numerical modelling and environmental design; *Hydrol. Process.* **22** 3337–3348.

Papadopoulou M P, Nikolos I K and Karatzas G P 2010 Computational benefits using artificial intelligent methodologies for the solution of an environmental design problem – saltwater intrusion; *Water Sci. Technol.* **62** 1479–1490.

Pope G and Nelson R 1978 A chemical flooding compositional simulator; *SPEJ* **18** 339–354.

Qin X S, Huang G H, Chakma A, Chen B and Zeng G M 2007 Simulation-based process optimization for surfactant-enhanced aquifer remediation at heterogeneous DNAPL-contaminated sites; *Sci. Total Environ.* **381** 17–37.

Queipo N V, Haftka R T, Shyy W, Goel T, Vaidyanathan R and Kevin Tucker P 2005 Surrogate-based analysis and optimization; *Prog. Aerosp. Sci.* **41** 1–28.

Razavi S, Tolson B A and Burn D H 2012 Review of surrogate modeling in water resources; *Water Resour. Res.* **48**, doi: 10.1029/2011WR011527.

Rogers L L, Dowla F U and Johnson V M 1995 Optimal field-scale groundwater remediation using neural networks and the genetic algorithm; *Environ. Sci. Technol.* **29** 1145–1155.

Ryu J S, Kim M S, Cha K J, Lee T H and Choi D H 2002 Kriging interpolation methods in geostatistics and DACE model; *J. Mech. Sci. Tech.* **16** 619–632.

Sacks J, Welch W J, Mitchell T J and Wynn H P 1989 Design and analysis of computer experiments; *Stat. Sci.* **4** 409–423.

Schaerlaekens J, Mertens J, Van Linden J, Vermeiren G, Carmeliet J and Feyen J 2006 A multi-objective optimization framework for surfactant-enhanced remediation of DNAPL contaminations; *J. Contam. Hydrol.* **86** 176–194.

Shen W, Guo X, Wu C and Wu D 2010 Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm; *Knowl.-Based Syst.* **3** 378–385.

Shyy W, Papila N, Vaidyanathan R and Tucker K 2001 Global design optimization for aerodynamics and rocket propulsion components; *Prog. Aerosp. Sci.* **37** 59–118.

Simpson A R, Dandy G C and Murphy L J 1994 Genetic algorithms compared to other techniques for pipe optimization; *J. Water Res. Pl-ASCE* **120** 423–443.

Simpson T W, Mauery T M, Korte J J and Mistree F 1998 Comparison of response surface and kriging models for multidisciplinary design optimization; 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, St. Louis, pp. 1–11.

Simpson T W, Mauery T M, Korte J J and Mistree F 2001 Kriging models for global approximation in simulation-based multidisciplinary design optimization; *AIAA J.* **39** 2233–2242.

Sreekanth J and 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** 245–256.

Yan S and Minsker B 2006 Optimal groundwater remediation design using an adaptive neural network genetic algorithm; *Water Resour. Res.* **42** 1145–1155.

Yan S and Minsker B 2011 Applying dynamic surrogate models in noisy genetic algorithms to optimize groundwater remediation designs; *J. Water Resour. Plan. Manag.* **137** 284–292.

MS received 18 July 2012; revised 6 May 2014; accepted 16 May 2014