Optimization design based on ensemble surrogate models for DNAPLs-contaminated groundwater remediation

Haibo Chu and Wenxi Lu

ABSTRACT

The optimization design of surfactant-enhanced aquifer remediation for dense non-aqueous phase liquids (DNAPLs)-contaminated groundwater is proposed through integrating simulation and optimization models. Many studies have demonstrated that the surrogate model is an effective tool for building a bridge between simulation and optimization models. In this paper, the simulation model was first established to simulate a surfactant-enhanced aquifer remediation process, and on this basis, the ensemble surrogate models were constructed by applying the polynomial regression model, radial basis function neurons network and Kriging surrogate models. Second, the best surrogate model was selected in terms of the three performance indices. Lastly, a non-linear programming optimization model was constructed with the target of minimizing the DNAPLs-contaminated aquifer remediation cost. Meanwhile, the best surrogate model was embedded into the optimization model as a constrained condition, and it was used to reflect the non-linear complex relationship between injection and extraction rates with DNAPLs removal rates instead of simulation models. The results showed that the ensemble surrogate models improved the performance of the single surrogate models. Moreover, ensemble surrogate models improved the computational efficiency, and the optimal strategies have been proved to be an effective guide for contaminant remediation processes.

Key words | DNAPLs, ensemble surrogate model, groundwater remediation, optimization

Haibo Chu
Wenxi Lu (corresponding author)
Key Laboratory of Groundwater Resources and Environment, Ministry of Education,
Jilin University,
Changchun 130021,
China
and
College of Environment and Resources,
Jilin University,
Changchun 130021,
China
E-mail: *luwenxi@jlu.edu.cn*

INTRODUCTION

Dense non-aqueous phase liquids (DNAPLs) are typically immiscible with water and denser than water, and they are widely used in industrial and manufacturing operations (Liang & Falta 2008). DNAPLs may act as long-term sources of groundwater contamination and lead to a great risk for drinking water resources. Because of low solubility, high interfacial tension and the sinking tendency of DNAPLs, as well as the complex subsurface condition of remediation sites, it is difficult to remediate groundwater contaminated by DNAPLs (Falta *et al.* 2005). Traditional remediation methods included cosolvent flooding, steam flooding, air sparging, and soil vapor extraction. More recently, surfactant-enhanced aquifer remediation (SEAR), which was developed on the basis of pump-and-treat techniques, has been widely considered as one of the most promising techniques to remediate DNAPL contaminations (Schaerlaekens *et al.* 2006). It is necessary to optimize the SEAR remediation design because the remediation is very expensive.

An optimization model was constructed on the basis of a simulation model, and the simulation model was used to forecast the fate of contaminants in the subsurface under various conditions and describe the response relationship between the inputs and outputs of the groundwater system. It is necessary to embed the simulation model into an optimization model. The traditional methods, such as the embedding method, the response matrix method and the state transition equation method, have limitations in practical applications.

In recent years, surrogate models have also been proved effective in embedding simulation models into optimization models. The surrogate models mainly included the polynomial regression model, Kriging, artificial neural networks (ANNs), and support vector regression, and were

doi: 10.2166/aqua.2015.089

widely used in various areas. Cooper *et al.* (1998) presented a simulation-optimization model for supporting the light non-aqueous phase liquid recovery process decisions, and the regression model was used as a surrogate model. Shourian *et al.* (2008) built an optimization model for water allocation planning at basin scale, and an ANN model was trained as a surrogate model to approximate the simulation model. Fu *et al.* (2010) also proposed ANN as a surrogate for the simulation of the urban wastewater system. Simpson *et al.* (2001) used Kriging models as alternatives for constructing global approximations in a real aerospace engineering application. Yun *et al.* (2009) proposed a multi-objective optimization method based on support vector regression as a surrogate model. Some research efforts have also been made for the surrogate models of the SEAR remediation operations. Huang *et al.* (2003) and Qin *et al.* (2007) presented integrated simulation-optimization systems for supporting decisions of SEAR. The surrogate models were both dual-response surface models. He *et al.* (2008) developed a simulation-based fuzzy chance-constrained programming model, and the surrogate model was built with ANN, which was used to reflect the relationship inputs and outputs instead of the simulation model. The surrogates used in SEAR remediation were discussed in detail by Qin *et al.* (2009). Hence, the ensemble surrogate models were proposed to improve the performance of single surrogate models. However, few efforts have been made to apply the ensemble surrogate model in this area.

The simulation-optimization process was carried out through integrating the simulation, surrogate model, and optimization methods, which was solved by simulated an annealing algorithm (Ramesh & Slobodan 2002). The flow chart of a simulation-optimization model is shown in Figure 1. The simulation model had been first established to simulate a surfactant-SEAR process by UTCHEM. Second, the Latin hypercube sampling (LHS) method and the multiphase flow simulation models were used for collecting input-output samples. Steps 1 and 2 were the research basis; parameter identification and calibration for the simulation model had been performed, so the simulation model behaved well enough to forecast the fate of contaminants in the subsurface under various conditions. The process was not described in detail because the main research concentrated on the surrogate models. In this paper, the polynomial regression model, RBFNN (radial basis function neurons network), and Kriging were built as the surrogate model, and the ensemble surrogate models were constructed with various combinations of single surrogate models, and the best surrogate model was selected. Finally, a non-linear programming optimization model was built to minimize the

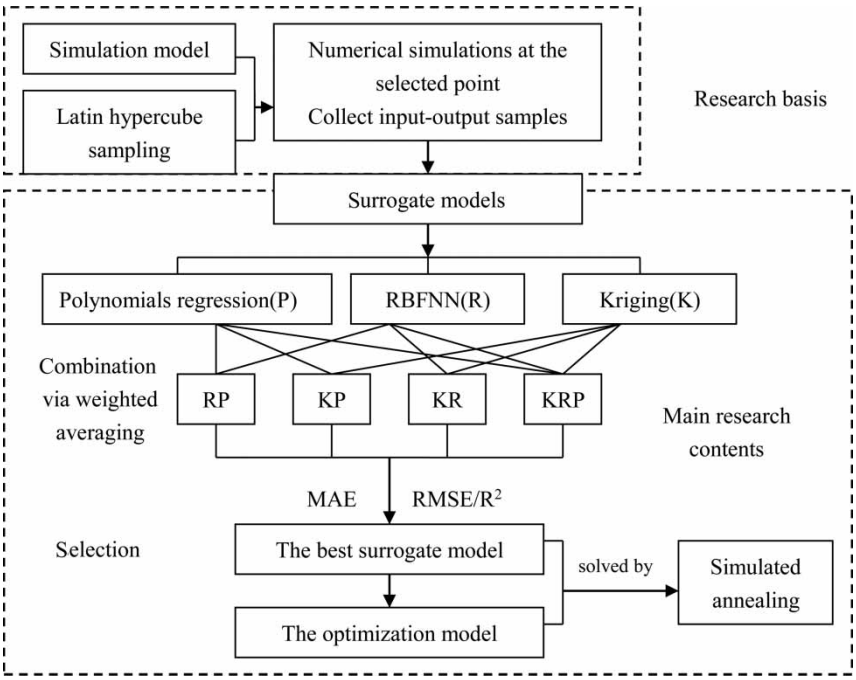

Figure 1 | Flow chart of optimization framework.

contaminated DNAPLs aquifer remediation cost, and so the optimal remediation strategy would be obtained for decision-makers with satisfying the remediation requirement.

METHOD

Polynomial regression model

The polynomial regression model is a multivariate statistical method, and it is widely applied in describing the functional relation between response variables and control variables. A second-order polynomial approximation without cross terms can be written as

$$f(x) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j + \sum_{j=1}^{n} \beta_{jj} x_j^2 + \varepsilon \quad (1)$$

where β_0, β_j, β_{jj} are the regression coefficients, ε is a normal random variable, $j = 1, 2, \ldots, n$. Equation (1) can be expressed in matrix form as

$$f(X) = \beta X + \varepsilon \quad (2)$$

$$f = [y_1, y_2, \ldots, y_n]^T$$

$$X = \left(1, x_j \ldots, x_n, \ldots x_j^2 \ldots x_n^2\right)^T$$

The parameters β are estimated by least squares, which are unbiased and have minimum variance (Zhou *et al.* 2005; Chang & Hsu 2006).

$$\beta = (X^T X)^{-1} X^T f \quad (3)$$

Radial basis function neurons network

RBFNN, which was proposed by Moody & Darken (1989), has been widely used for classification and function approximation. It is demonstrated that the network has local approximation ability and can quickly predict the system response relationship directly from input-output data sets. Basically, RBFNN consists of three layers: the input layer, the hidden layer, and the output layer (Samanta & Bandopadhyay 2009; Herrera *et al.* 2011; Luo & Lu 2014). A single-output RBF neural network with m hidden layer neurons can be expressed as

$$y = \sum_{i=1}^{m} \omega_i R_i(x) + \delta \quad (4)$$

where $y =$ the output neurons in the output layer, $\omega_i =$ the weight connection between the ith hidden and the output neurons, $\delta =$ the threshold value of the output neurons (Chattopadhyay 2007).

The function most commonly used in the hidden layer is the Gaussian function:

$$R_i(x) = e[-\|x - c_i\|^2 / 2\sigma^2] \quad i = 1, 2, \ldots m \quad (5)$$

where $x = n$-dimensional input vector, $c_i =$ center of the ith radial basis function; $\sigma =$ spread of the radial basis function, $m =$ the number of the hidden neurons, $\|x - c_i\| =$ the radial distance between x and the RBF center.

The learning process of the RBF network is divided into two phases: (1) the input layer-hidden layer, where the K-means clustering center method is usually used to allocate the centers and the spread of the Gaussian functions in the input space; and (2) the hidden layer-output layer, where the weights connecting the hidden and output neurons are estimated by the least mean square method (Han *et al.* 2011). The performance of an RBFNN mainly depends on the number of neurons in the hidden layer. In this paper, the optimal number of neurons in the hidden layer is determined by a trial-and-error process.

Kriging

The Kriging method was first used in the 1960s by Krige, and then developed by Matheron (Kleijnen 2009). In recent years, Kriging models have attracted much attention as a special technique and been widely used in many areas.

In this method, it is assumed that data have been collected at n points denoted by $x = [x_1, x_2, \ldots, x_n]$ and the associated response is denoted by $y = [y_1, y_2, \ldots, y_n]$; Y can be expressed as

$$Y(x) = f(x) + Z(x) \quad (6)$$

where $f(x)$ represents the regression function model, and $Z(x)$ is a model of a Gaussian and stationary stochastic

process with mean of 0 and variance of σ^2, which can be written as

$$Y(x) = \sum_{j=1}^{n} \beta_j f_j(x) + Z(x) \tag{7}$$

$$f(x) = [f_1(x), f_2(x), \ldots, f_n(x)]^T \quad \beta = [\beta_1, \beta_2, \ldots, \beta_n]^T$$

where β_j denotes the matrix of regression coefficients. In the stochastic process, the errors for the samples are related and the correlation is linked to the distance between the corresponding samples, and usually is expressed as

$$R(\theta, x_i, x_j) = \exp\left(-\theta \sum_{l=1}^{d} \|x_i - x_j\|^2\right) \tag{8}$$

where R represents the $n \times n$ correlation function matrix between x_i and x_j, which can be formed by the Gaussian correlation function with a single correlation parameter θ. The covariance of $Z(x)$ is expressed as

$$\mathrm{cov}[Z(x_i), Z(x_j)] = \sigma^2 R(\theta, x_i, x_j) \tag{9}$$

The unknown correlation parameters β, θ, and σ^2 can be estimated by maximizing the log-likelihood function (Forrester & Keane 2009; Huang *et al.* 2011):

$$\ln(\beta, \sigma^2, \theta) = -\frac{n}{2}\ln(\sigma^2) + \frac{1}{2}\ln(|R|) - \frac{((y - f\beta)^T R^{-1}(y - f\beta))}{2\sigma^2} \tag{10}$$

by differentiating the log-likelihood function with respect to β, θ, and σ^2, respectively, and letting them be equal to zero (Lophaven *et al.* 2002; Fu *et al.* 2009).

The function value Y at new points x can be approximately predicted by

$$Y(x) = f^T(x)\beta + r^T(x)R^{-1}(y - F\beta) \tag{11}$$

and

$$r^T(x) = [R(\theta, x, x_1), R(\theta, x, x_2), \ldots, R(\theta, x, x_n)]^T \tag{12}$$

The ensemble surrogate models

Three types of surrogate models are considered in this study: the polynomial regression model (P), Kriging (K), and radial basis function neural network (R). The ensemble models were constructed with various combinations of single surrogate models

$$F(x) = \sum_{i=1}^{n} w_i f_i(x) \quad \sum_{i=1}^{n} w_i = 1 \tag{13}$$

where $F(x)$ was the ensemble surrogate model, $f_i(x)$ was the ith surrogate model, w_i was the weight coefficient of the ith surrogate models, n was the number of surrogate models. If the ensemble models were constructed by Kriging (K) and radial basis neural network (R), the ensemble models (KR) can be expressed as

$$\mathrm{KR} = w_1 * K + w_2 * R \tag{14}$$

So the other ensemble models can also be expressed as

$$\mathrm{KP} = w_1 * K + w_2 * P \tag{15}$$

$$\mathrm{RP} = w_1 * R + w_2 * P \tag{16}$$

$$\mathrm{KRP} = w_1 * K + w_2 * R + w_3 * P \tag{17}$$

The weight coefficients were calculated according to PRESS (predicted residual sum of squares) weighted average surrogate, as follows:

$$w_i = (E_i + 0.05\bar{E})^{-1} / \sum_{i=1}^{M} (E_i + 0.05\bar{E})^{-1} \tag{18}$$

$$\bar{E} = \frac{1}{n}\sum_{i=1}^{n} E_i \tag{19}$$

where E_i was the PRESS error of the ith surrogate, n was the number of surrogate models (Goel *et al.* 2007).

Excel 2003 and Matlab were used to calculate the parameters of the surrogates developed in the present study. The root mean squared error (RMSE), the maximum absolute error (MAE) and coefficient of determination (R^2) were used in order to assess the effectiveness of each model and its ability to approximate the simulation model (May *et al.* 2008). RMSE, MAE, and R^2 were employed to evaluate the accuracy of surrogate models. The lower RMSE and

MAE were, the nearer the R^2 value to 1, the more approximated to the simulation model the surrogate model was.

CASE STUDY

To illustrate the application of the ensemble surrogate model in the groundwater remediation process, an illustrative PCE (petrachloroethylene)-contaminated aquifer was selected as the case study.

The study area was 882 m^2, 49 m in the x direction and 18 m in the y direction. The thickness of aquifer in porous layers is 24 m, and the aquifer is confined and heterogeneous anisotropic. Two soil types, clay and sand, occupied the simulation domain. Initial conditions were prescribed such that the groundwater flowed from left to right, with a hydraulic gradient of 0.004706, and the west and east boundaries were set as first-type boundaries, south and north were set as no-flow boundaries. The other main parameters are listed in Table 1.

Table 1 | Physical and chemical parameters in the research domain

Parameter	Value	Parameter	Value
Porosity	0.32	Hydraulic gradient	0.004706
Longitudinal dispersivity	3 m	Transverse dispersivity	0.3 m
PCE solubility in water	240 mg/L	PCE/water interfacial tension	44.67 dyn/cm
Water density	0.998 g/cm^3	PCE density	1.623 g/cm^3
Water viscosity	1.00 cp	PCE viscosity	0.89 cp
Residual water saturation	0.24	Residual PCE saturation	0.17

According to the contaminant distribution shown in Figure 2, a SEAR system with three extraction wells and three injection wells at a constant flow rate of 80 m^3/d and a 4% surfactant solution was designed. Numbers 1–3 are injection wells and numbers 4–6 are extraction wells. The simulations were conducted using a numerical simulator UTCHEM.

Surrogate model

In all the surrogate models, the injection and extraction rates were explanatory variables, and DNAPLs removal rates were considered as system response variables. The injection and extraction rates were set as the inputs into the simulation model, and UTCHEM software was called, and finally DNAPLs removal rates were calculated as the system response. The sampling was performed with the LHS technique. LHS provides a random sampling but ensures a stratified sample within the full range of each dimension of the sample space (Hora & Helton 2003; Davey 2008). For this study, a total of 160 samples were obtained by running the simulation models. They were divided into training and testing sets. The training set with 140 samples was used for constructing the surrogate models, and the remaining 20 samples were used for testing the surrogate models, as shown in Table 2.

The polynomial regression model, RBFNN, and Kriging were used to build the surrogate models, and the ensemble

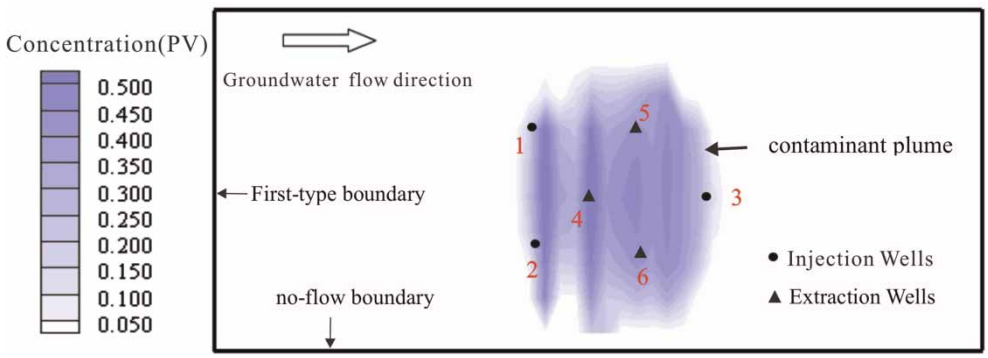

Figure 2 | The well locations of the study area.

Table 2 | The parts of the testing samples

	Inputs (extraction and injection rate) m³/d						Outputs (DNAPLs removal rates)
	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Year (%)
1	7.97	37.66	75.43	56.73	46.26	18.08	57.02
2	38.00	59.96	23.90	46.83	38.49	36.53	71.17
3	39.89	27.91	4.84	58.42	8.01	6.21	57.29
4	5.38	62.59	1.59	56.10	0.66	12.79	46.45
5	12.69	11.13	58.73	40.21	7.86	34.47	52.05
6	35.87	38.98	62.22	48.70	52.72	35.64	73.74
7	28.78	47.27	51.61	12.49	37.70	77.46	70.19
8	16.60	49.15	22.97	16.33	13.99	58.40	56.66
9	75.86	76.27	11.41	57.08	27.39	79.08	80.33
10	45.91	28.12	30.74	54.09	13.68	36.99	69.36

surrogate models were constructed according to single surrogate models, then the best surrogate model was selected according to the three performance indices.

Optimization model

In the optimization model, minimizing the operating system costs was considered as the objective function, and operating cost was simplified as the function with extraction and injection rates. Therefore, the objective function can be transformed to minimize the total extraction and injection rates. The total extraction and injection rates are the decision variables, with the main constraints being: (1) extraction and injection rates meet both the lower and upper limit; (2) the total extraction rates equal the injection rates, which does not exert impact on groundwater flow; (3) the final surrogate model, which reflects the relationship between extraction and injection rate and DNAPLs removal rates; and (4) the remediation aims to determine the most cost-effective strategy for removing more than 80% of the DNAPLs contamination.

In the study, the objective is to minimize the total cost of the remediation while meeting the remediation effectiveness.

$$\min Z = Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6 \tag{20}$$

$$\begin{cases} 0 \le Q_i \le 80 \ (i = 1,\ 2,\ 3,\ 4,\ 5,\ 6) \\ Q_{\text{in}} = Q_{\text{ex}} \\ y = f(Q_i) \ (i = 1,\ 2,\ 3,\ 4,\ 5,\ 6) \\ y \ge 0.80 \end{cases} \tag{21}$$

where y denotes DNAPLs removal rate, Q_{in}, Q_{ex} denote the total injection rates and the total extraction rates, and $y = f(Q_i)$ is the best surrogate model.

RESULTS AND DISCUSSION

Surrogate models

The parameters of the polynomial regression model were estimated with the training samples by the least squares method. The final polynomial regression model was expressed as

$$\begin{aligned} y = &-0.005x_1^2 - 0.0013x_2^2 - 0.0011x_3^2 - 0.0021x_4^2 \\ &- 0.0019x_5^2 + 0.3965x_1 - 0.1322x_2 - 0.1565x_3 + 0.6764x_4 \\ &+ 0.6027x_5 + 0.529x_6 + 13.5039 \end{aligned}$$

The RBFNN surrogate model was trained. The *newrb* function in MATLAB was called. The optimal number of the neurons in the hidden layer was identified using a trial and error procedure by changing the number of hidden neurons from 1 to 20. The effect of number of hidden neurons on the RMSE of the testing samples is shown in Figure 3. It can be seen that the RMSE reached the minimum when the number of hidden neurons was 12, so the numbers of hidden neurons were set as 12. Hence, a three-layer ANN model was built with six neurons in the input layer, 12 neurons in the hidden layer and one neuron in the output layer. The spread was 1.9, and the target error of training is 0.003; as shown in Figure 4, the error performance is 0.0028, which is less than the target error of training.

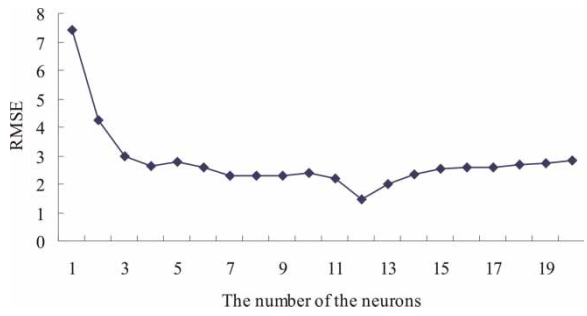

Figure 3 | Root mean square error (RMSE) for various numbers of hidden neurons.

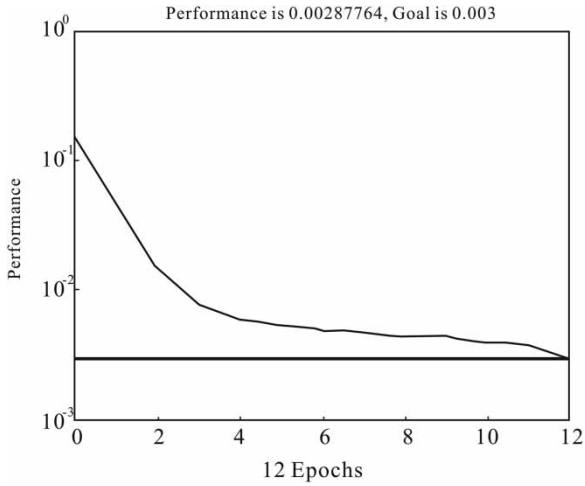

Figure 4 | Variation in error function during training of the RBFNN.

The Kriging model was completed with the computing program, dacefit and predictor function in a MATLAB toolbox. It was expressed as [dmodel, perf] = dacefit(S, Y, regr, corr, theta0). S and Y are, respectively, the input and output data sets, second order polynomial and Gaussian function were selected as regression models and correlation models, and theta0 was the initial value of the parameter θ. The Kriging model parameters corresponding to the initial sample are listed in Table 3.

Figure 5 illustrates a comparison among predicted and simulated DNAPLs removal rates for testing by three surrogate models. The results showed that the surrogate models were all accurate in forecasting the DNAPLs removal rate, which implies that they are capable of being alternatives of the simulation model. However, the Kriging model approximates best the simulation model, and the polynomial response surface model performs worst. The RMSE for the Kriging and polynomial regression model is 1.30 and 2.30, respectively, the MAE 2.37 and 3.94, respectively, and the R^2 0.99 and 0.97, respectively.

The ensemble surrogate models were constructed with various combinations of single surrogate models. The weight coefficients for all the surrogate models are shown in Table 4 (K, Kriging model; R, radial basis function neural network; P, polynomial regression model). For example, in the first column of Table 4, it means that the weighted average surrogate models can be expressed as

$$\mathrm{KR} = 0.542{*}K + 0.458{*}R$$

The results show that the better the single surrogate model performs, the bigger weight coefficients are. On the other hand, the weight coefficients for the single surrogate models that perform poorly are relatively smaller.

Figure 6 shows the modeling results of the different ensemble surrogate models. It is obvious that the KRP model fits the predicted and simulated DNAPL removal rate best. It can also be seen in Figure 7 that the minimum and maximum value of R^2 for all the surrogates (single and ensemble) is 0.975 by the polynomial regression model and 0.993 by the KRP model, and the better fit between observed and predicated values, and the bigger R^2 value would be, until it reached 1. The minimum and maximum value of MAE is 2.12 by the KRP model and 3.94 by the polynomial regression model, minimum and maximum value of RMSE is 1.15 by the KRP model and 2.30 by the polynomial response surface model; MAE and RMSE demonstrate a similar trend. The lower the RMSE and MAE, the more accurate the prediction is. The KRP model has the lower RMSE/MAE values and higher R^2 compared with the other models, which implies that KRP has a superior performance in reflecting the non-linear relationships between explanatory and response variables.

Figure 7 indicates that all three single surrogate models can be used for constructing a linkage between extraction and injection rates and the DNAPLs removal rates of the real groundwater system. Polynomials could capture the global trend over the entire input space, which is relatively easy to construct and has the simplest type of parameters, but the accuracy is not satisfactory. The RBF model is appropriate to reproduce the non-linear problem, and the accuracy is satisfactory, but there is a risk of overfitting. The Kriging model is the combination of a regression model and a localized deviation model based on spatial correlation of samples. It could interpolate at the sample points and is more accurate for non-linear problems. Considering the overall characteristics of all single surrogate models,

Table 3 | Kriging model parameters corresponding to the initial sample

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	σ^2
0.1242	0.1130	0.1655	0.1879	0.1000	0.1809	3.8592

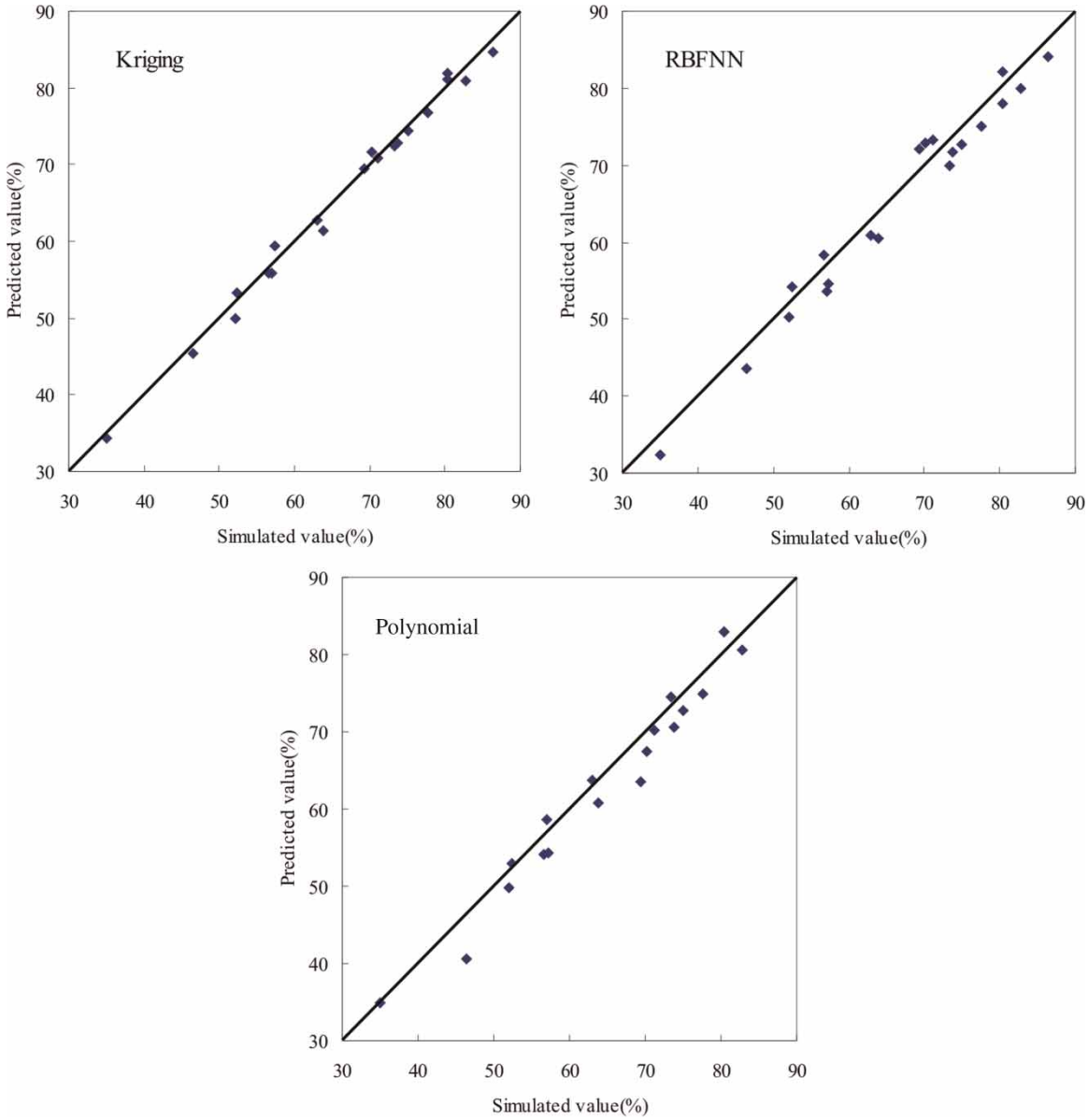

Figure 5 | Predicted and simulated DNAPLs removal rate (single surrogate model).

the ensemble of surrogates not only improved the performance of single surrogate models, but also performed more stably than the single surrogate models. In this paper, the ensemble surrogate model, which was composed of the polynomial regression model, the radial basis function neurons network, and the Kriging (KRP model) performed the best among all the surrogates, and the polynomial regression model performed the worst among all the surrogates. However, the ensemble needed many parameters for all the single surrogate models, and it took a long training time to get the parameters. Hence, the ensemble of surrogates method regarding optimizing parameters should be further studied.

Table 4 | The weight coefficients for different surrogate models

	KR	KP	RP	KRP
K	0.542	0.745	–	0.455
R	0.458	–	0.713	0.386
P	–	0.255	0.287	0.159

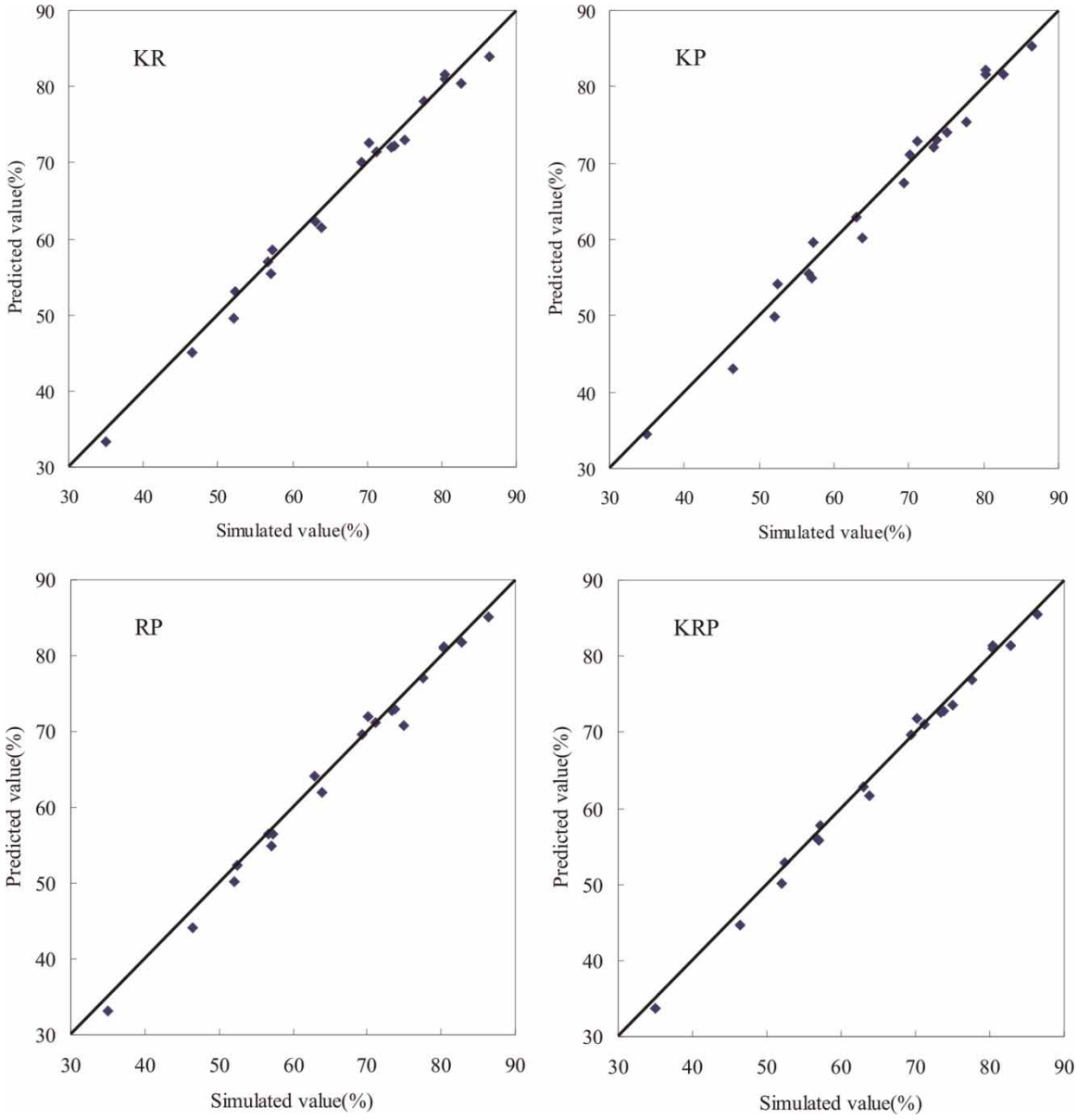

Figure 6 | Predicted and simulated DNAPLs removal rate (ensemble surrogate models).

The optimization results

The simulated annealing method was used to solve the optimization model. The optimal pumping rates Q_1–Q_6 were 73.89, 29.13, 25.83, 58.5, 15.31 and 54.97 m^3/d, respectively. The remediation effectiveness was 80.06%, and the total extraction rate and injection rate were 257.69 m^3/d, respectively. If the total extraction rate and injection rate remain constant, that is, the remediation cost remains constant, the extraction rate and injection rate are randomly generated, the optimal pumping rates Q_1–Q_6 were 58.89, 39.13, 30.83, 43.57, 20.31, 20.31 and 64.97 m^3/d, respectively, and the remediation effectiveness was 77.79%. It was obvious that the optimization model based on the ensemble surrogate model effectively improved DNAPLs removal rate even with the same cost.

In the optimization model, the remediation levels can be changed according to the requirements. If the optimal strategy

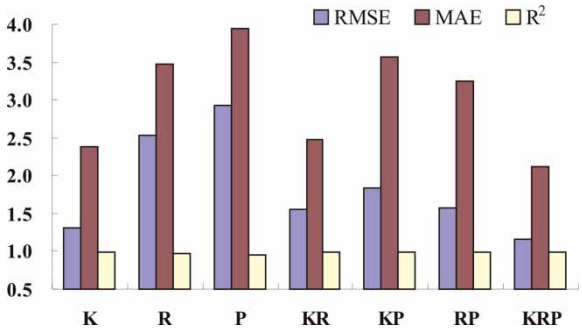

Figure 7 | RMSE, MAE, and R^2 for different surrogate models.

needs to be implemented in the water supply region, the remediation levels should be as high as possible, because it is important to protect drinking water resources. If the optimal strategy needs to be implemented in remote areas, where there are few people, the remediation levels can be relatively low. The variations in the total extraction and injection rates for the different remediation levels (DNAPLs removal rates) are shown in Figure 8. As can be seen in Figure 8, the remediation levels varied from 50 to 90%. As the remediation levels improved, which means the values of DNAPLs removal rates became larger, the total extraction and injection rates also increased. It is significant for decision-makers to obtain the optimal strategies under different situations.

The surrogate model was an important part of the optimization process as constraints, and it could capture better the response relationship between the extraction (injection) rates and DNAPLs removal rates. If the optimal solution was obtained under 2,000 times to call the simulation model or surrogate models in the optimization process, it would take 30 minutes to assess the solution feasibility by calling the simulation model in the optimization process, so it needed 52 days to get the optimal solution; however, it took only 2 minutes by calling the surrogate model. The results showed that the surrogate model reduced the huge computational burden and remediation cost and it could approximate the best solution quickly.

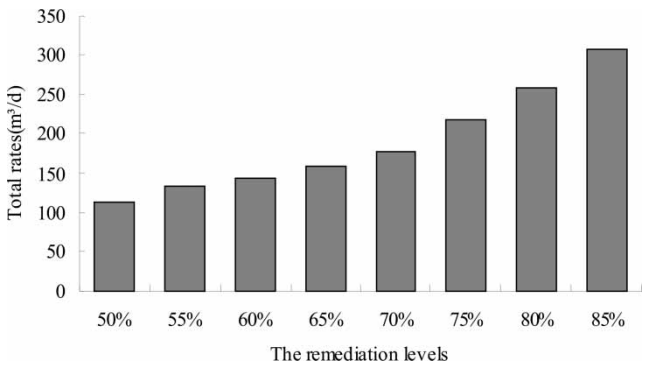

Figure 8 | Total extraction and injection rates for different remediation levels.

CONCLUSIONS

A numerical scheme was developed on the integrated simulation and optimization model for exploring groundwater remediation with DNAPLs contamination. The framework included the UTCHEM model that simulated a three-dimensional, multicomponent, multiphase, compositional surfactant flooding process. This was followed by the ensemble surrogate model to reflect the non-linear complex relationship between injection and extraction rates with DNAPLs removal rates instead of the simulation model. Finally, an optimization model to optimize the operation cost with the ensemble surrogate model was proposed. The results show that the ensemble surrogate models not only improved the performance of single surrogate models, but also performed more stably than the single surrogate models. However, the ensemble surrogate models needed many parameters for all the single surrogate models; further study is needed to provide the parameters more quickly and accurately. Moreover, ensemble surrogate models could reduce the computational burden and save a lot of time. The optimal strategies have proved themselves as an effective guide for decision-makers in the DNAPLs contaminants remediation process.

There are a large number of uncertainty factors during the optimization process based on the surrogate model, which inevitably have an effect on the reliability of the remediation design. Future studies need to be undertaken to investigate the uncertainty source and type, consider how the uncertainty is introduced into the optimization model, and how to combine the optimization design and its reliability.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Funds of China (no. 41072171 and 41372237) and Key Laboratory of Groundwater Resources and Environment, Ministry of Education, Jilin University, Changchun, China.

REFERENCES

Chang, A. C. & Hsu, J. P. 2006 A polynomial regression model for the response of various accelerating techniques on maize wine maturation. *Food Chem.* **94** (4), 603–607.

Chattopadhyay, S. 2007 Prediction of mean monthly total ozone time series – application of radial basis function network. *Int. J. Remote Sens.* **28** (18), 4037–4046.

Cooper, G. S., Peralta, R. C. & Kaluarachchi, J. J. 1998 Optimizing separate phase light hydrocarbon recovery from contaminated unconfined aquifers. *Adv. Water Resour.* **21** (5), 339–350.

Davey, K. R. 2008 Latin hypercube sampling and pattern search in magnetic field optimization problems. *IEEE Trans. Magnetics* **44** (6), 974–977.

Falta, R. W., Rao, P. S. & Basu, N. 2005 DNAPL Source zones I. Analytical modeling of source strength functions and plume response. *J. Contam. Hydrol.* **78**, 259–280.

Forrester, A. I. & Keane, A. J. 2009 Recent advances in surrogate-based optimization. *Prog. Aerosp. Sci.* **45** (1), 50–79.

Fu, G., Khu, S. & Butler, D. 2009 Use of surrogate modelling for multiobjective optimisation of urban wastewater systems. *Water Sci. Technol.* **60** (6), 1641–1647.

Fu, G., Makropoulos, C. & Butler, D. 2010 Simulation of urban wastewater systems using artificial neural networks: embedding urban areas in integrated catchment modelling. *J. Hydroinform.* **12** (2), 140–149.

Goel, T., Haftka, R. T., Shyy, W. & Queipo, N. V. 2007 Ensemble of surrogates. *Struct. Multidiscip. Optim.* **33** (3), 199–216.

Han, H. G., Chen, Q. L. & Qiao, J. F. 2011 An efficient self-organizing RBF neural network for water quality prediction. *Neural Networks* **24**, 717–725.

He, L., Huang, G. H. & Lu, H. W. 2008 A simulation-based fuzzy chance-constrained programming model for optimal groundwater remediation under uncertainty. *Adv. Water Resour.* **31**, 1622–1635.

Herrera, L. J., Pomares, H., Rojas, I., Guillén, A., Rubio, G. & Urquiza, J. 2011 Global and local modelling in RBF networks. *Neurocomputing* **74**, 2594–2602.

Hora, S. C. & Helton, J. C. 2003 A distribution-free test for the relationship between model input and output when using Latin hypercube sampling. *Reliab. Eng. Syst. Safe.* **79** (3), 333–339.

Huang, Y. F., Li, J. B., Huang, G. H., Chakma, A. & Qin, X. S. 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 Manage.* **7** (2), 95–105.

Huang, Z. J., Wang, C. G., Chen, J. & Tian, H. 2011 Optimal design of aeroengine turbine disc based on kriging surrogate models. *Comput. Struct.* **89**, 27–37.

Kleijnen, J. P. 2009 Kriging metamodeling in simulation: a review. *Eur. J. Oper. Res.* **192** (3), 707–716.

Liang, H. L. & Falta, R. W. 2008 Modeling field-scale cosolvent flooding for DNAPL source zone remediation. *J. Contam. Hydrol.* **96**, 1–16.

Lophaven, S., Nielsen, H. & Sondergaard, J. 2002 *Dace-a matlab Kriging toolbox, version 2.0. Tech Rep Informatics and mathematical modelling report IMM-REP -2002–12.* Technical University of Denmark, Lyngby, Denmark.

Luo, J. & Lu, W. 2014 Sobol' sensitivity analysis of NAPL-contaminated aquifer remediation process based on multiple surrogates. *Comput. Geosci.* **67**, 110–116.

May, R. J., Dandy, G. C., Maier, H. R. & Nixon, J. B. 2008 Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. *Environ. Modell. Softw.* **23**, 1289–1299.

Moody, J. & Darken, C. 1989 Fast learning in networks of locally-tuned processing units. *Neural Comput.* **1**, 281–294.

Qin, X. S., Huang, G. H., Chakma, A., Chen, B. & Zeng, G. M. 2007 Simulation-based process optimization for surfactant-enhanced aquifer remediation at heterogeneous DNAPL-contaminated sites. *Sci. Total Environ.* **381** (1–3), 17–37.

Qin, X. S., Huang, G. H. & He, L. 2009 Simulation and optimization technologies for petroleum waste management and remediation process control. *J. Environ. Manage.* **90** (1), 54–76.

Ramesh, S. V. T. & Slobodan, P. S. 2002 Optimal operation of reservoir systems using simulated annealing. *Water Resour. Manage.* **16**, 401–428.

Samanta, B. & Bandopadhyay, S. 2009 Construction of a radial basis function network using an evolutionary algorithm for grade estimation in a placer gold deposit. *Comput. Geosci.* **35** (8), 1592–1602.

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

Shourian, M., Mousavi, S., Menhaj, M. & Jabbari, E. 2008 Neural-network-based simulation-optimization model for water allocation planning at basin scale. *J. Hydroinform.* **10** (4), 331–343.

Simpson, T. W., Mauery, T. M., Korte, J. J. & Mistree, F. 2001 Kriging models for global approximation in simulation-based multidisciplinary design optimization. *AIAA J.* **39** (12), 2233–2241.

Yun, Y., Yoon, M. & Nakayama, H. 2009 Multi-objective optimization based on meta-modeling by using support vector regression. *Optim. Eng.* **10** (2), 167–181.

Zhou, Z., Ong, Y. S., Nguyen, M. H. & Lim, D. 2005 A study on polynomial regression and Gaussian process global surrogate model in hierarchical surrogate- assisted evolutionary algorithm. *Evolut. Comput.* **3**, 2832–2839.

First received 21 June 2014; accepted in revised form 22 January 2015. Available online 25 March 2015