A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen

Xue Li, Jian Sha and Zhong-liang Wang

ABSTRACT

Dissolved oxygen (DO) is an important indicator reflecting the healthy state of aquatic ecosystems. The balance between oxygen supply and consuming in the water body is significantly influenced by physical and chemical parameters. This study aimed to evaluate and compare the performance of multiple linear regression (MLR), back propagation neural network (BPNN), and support vector machine (SVM) for the prediction of DO concentration based on multiple water quality parameters. The data set included 969 samples collected from rivers in China and the 16 predicted variables involved physical factors, nutrients, organic substances, and metal ions, which would affect the DO concentrations directly or indirectly by influencing the water–air exchange, the growth of water plants, and the lives of aquatic animals. The models optimized by particle swarm optimization (PSO) algorithm were calibrated and tested, with nearly 80% and 20% data, respectively. The results showed that the PSO-BPNN and PSO-SVM had better predicted performances than linear regression methods. All of the evaluated criteria, including coefficient of determination, mean squared error, and absolute relative errors suggested that the PSO-SVM model was superior to the MLR and PSO-BPNN for DO prediction in the rivers of China with limited knowledge of other information.

Key words | back propagation neural network, multiple linear regression, particle swarm optimization algorithm, support vector machine, water quality parameters

Xue Li
Jian Sha
Zhong-liang Wang (corresponding author)
Tianjin Key Laboratory of Water Resources and Environment,
Tianjin Normal University,
Tianjin 300387,
China
E-mail: *wangzhongliang@vip.skleg.cn*

INTRODUCTION

Water quality monitoring always involves a series of physical, chemical, and biological parameters (Collins *et al.* 2012; Cao *et al.* 2016). It is a difficult problem to find out the complex relationships among the large number of variables (Carlyle & Hill 2001; Bonansea *et al.* 2015). Among the multiple physical and chemical parameters, the concentration of dissolved oxygen (DO) is an important parameter affecting the healthy functioning of aquatic species and an integrated indicator reflecting the state of aquatic ecosystems (Ficklin *et al.* 2013). The sources of DO mainly include re-aeration from the atmosphere and photosynthetic oxygen production, while the consumption processes include the oxidation of carbonaceous and nitrogenous material, respiration of aquatic animals and plants, as well as sediment oxygen demand (Kuo *et al.* 2007; Klose *et al.* 2012). During the momentary balance between oxygen supply and metabolic consumption, the concentration of DO is known to be influenced directly by many biological processes, such as respiration, photosynthesis, and decomposition (Kannel *et al.* 2007). Although biological processes directly influence DO, other physical and chemical parameters could control and limit the effects of these biological processes to some extent (Rounds 2002; Salami Shahid & Ehteshami 2016). For example, nitrogen and phosphorus have an effect on the growth of algae and plants, while metal ions are

doi: 10.2166/nh.2016.149

important to the lives of animals and bacteria in the aquatic environment (Li *et al.* 2015; Surinaidu 2016). As an integrated indicator, many inaccessible parameters during complex biological processes are an important basis to construct deterministic models. It is highly desirable to determine a DO model for rivers which could quantify and predict DO concentrations accurately, only based on physico-chemical parameters, for water resources managers.

Several models generally grouped as deterministic models and statistical models have been developed for the analysis of DO (Cox 2003). Most physically based models have complex structures and require many types of input data which are not easily accessible, making it a very costly and time-consuming modeling process (Stefan & Fang 1994; Sear *et al.* 2014). Besides, the demands of explicit understanding of a series of physical processes, a degree of expertise and experience with the models raise high demands for the researchers. The data-driven models without any information on the physical, chemical, and biological reaction processes are very useful for DO prediction in rivers. Many different methods, such as multiple linear regression (MLR), partial least squares regression, various types of artificial neural networks (ANNs), genetic algorithms (GAs), and support vector machines (SVMs) have been developed and applied widely in recent years (Modaresi & Araghinejad 2014; Were *et al.* 2015; Isunju & Kemp 2016). Especially the ANN and SVM, which offer advantages over conventional non-linear models for the handling of complex relationships between input and output variables, have been successfully used in various water resources problems (Hosseini & Mahjouri 2014), including the modeling and forecasting of DO concentrations (Liu *et al.* 2013; Wen *et al.* 2013). However, it is still not clear which method has the best performance on the prediction of DO concentrations in rivers based on other physico-chemical parameters.

In this study, the MLR, ANN, and SVM were applied to forecast DO concentrations in the rivers of China. The results of the three models were compared to each other based on various statistical evaluation measures. The aim of this study was to discuss and evaluate the performance of three data-driven models and choose the best one on the prediction of DO concentration influenced by other water quality parameters.

MATERIALS AND METHODS

Data sets

In this study, three models with different structures were designed to predict DO concentrations based on multiple water quality parameters. To achieve this objective, a data set including 600 monitoring sites distributed on nearly all the main streams and chief tributaries in China from 2009 to 2010 was obtained from national environmental agencies. The monitoring sites with missing data were not taken into account and 969 records with 21 parameters were selected as the initial data set. The parameters included DO, water temperature (TEMP), pH values, potassium permanganate index (COD_{Mn}), biochemical oxygen demand (BOD), ammonia nitrogen (NH_3-N), total nitrogen (TN), total phosphorus (TP), petroleum (PE), volatile phenol (VP), chemical oxygen demand (COD_{Cr}), mercury (Hg), lead (Pb), copper (Cu), fluoride (F), zinc (Zn), arsenic (As), cadmium (Cd), hexavalent chrome (Cr), cyanide (Cyn), and anionic surfactant (LAS). In order to reduce the large numbers of predictors and select the most effective variables, the Spearman correlation analysis was used to evaluate the degree of association between DO and other parameters, shown in Table 1. According to the results of correlation analysis, significant ($P < 0.001$) correlations were observed between DO and most of the water quality parameters except for Pb, Cu, Cd, and Cr. Although some values of correlation coefficient were relatively small, weak linear relationships were indicated. However, the statistically significant correlations showed appropriate and significant associations between these variables, maybe non-linear relationships. There were 16 parameters finally selected as input data and the basic statistics of these measured water quality parameters are summarized in Table 2. The predicted variables were concerned with physical factors, nutrients, organic substances, and metal ions, which affected the DO concentrations directly or indirectly by influencing the water–air exchange, the growth of water plants, and

Table 1 | Spearman correlation coefficients between DO and other water quality parameters

Parameters	Correlation coefficient	Sig. (2-tailed)	Parameters	Correlation coefficient	Sig. (2-tailed)
TEMP (°C)	−0.31**	<0.01	TN (mg/L)	−0.52**	<0.01
pH	0.40**	<0.01	TP (mg/L)	−0.55**	<0.01
COD_{Mn} (mg/L)	−0.50**	<0.01	Cu (mg/L)	−0.01	0.81
BOD (mg/L)	−0.55**	<0.01	F (mg/L)	−0.39**	<0.01
NH_3-N (mg/L)	−0.65**	<0.01	Zn (mg/L)	−0.22**	<0.01
PE (mg/L)	−0.39**	<0.01	As (mg/L)	−0.17**	<0.01
VP (mg/L)	−0.37**	<0.01	Cd (mg/L)	0	0.99
Hg (mg/L)	−0.13**	<0.01	Cr (mg/L)	−0.04	0.22
Pb (mg/L)	−0.002	0.95	Cyn (mg/L)	−0.20**	<0.01
COD_{Cr} (mg/L)	−0.48**	<0.01	LAS (mg/L)	−0.45**	<0.01

**Correlation is significant at the 0.01 level (2-tailed).

Table 2 | The basic statistics of 16 measured water quality parameters in rivers of China

Parameters	Unit	Minimum value	Mean value	Maximum value	Standard deviation
DO	mg/L	0.10	7.13	14.92	1.86
TEMP	°C	2.00	16.74	30.93	3.79
pH	dimensionless	6.76	7.72	8.96	0.36
COD_{Mn}	mg/L	0.68	5.28	52.99	5.15
BOD	mg/L	0.42	4.36	98.35	7.50
NH_3-N	mg/L	0.02	1.92	37.10	4.58
PE	mg/L	0.01	0.07	1.39	0.12
VP	mg/L	0.0002	0.003	0.11	0.008
Hg	mg/L	0	0.00003	0.001	0.00005
COD_{Cr}	mg/L	1.65	21.80	258.42	22.29
TN	mg/L	0.03	4.27	54.72	6.41
TP	mg/L	0.01	0.24	4.80	0.45
F	mg/L	0.01	0.54	7.40	0.47
Zn	mg/L	0.0005	0.03	0.44	0.04
As	mg/L	0.00003	0.003	0.08	0.006
Cyn	mg/L	0.0005	0.003	0.07	0.004
LAS	mg/L	0.01	0.09	2.88	0.19

the lives of aquatic animals, etc. The data set included 969 samples randomly split into 769 samples as the training set and 200 samples as the testing set, which was nearly 80% and 20% of the whole data set. The test set was used to evaluate the effects of the calibrated models. The raw data of both training set and testing set were standardized between 0.1 and 0.9 before analysis to eliminate the effects of various dimensions and maintain the same or similar importance.

SVM for regression

SVM is a recently developed supervised machine learning method for classification and prediction (Meyer *et al.* 2003).

It is based on non-linear statistical theory to transform input space into a higher dimensional feature space for the purpose of separating the data patterns (Baylar *et al.* 2009). The goal of the SVM is to find an optimal hyperplane, which could differentiate the data in different classes by the maximum gap (Smola & Schölkopf 2004). The SVM is one of the best algorithms used for binary classification, and also has been extended to solve non-linear regression problems (He *et al.* 2014). Considering a set of training data (x_1, y_1), (x_2, y_2), ···, (x_i, y_i), ···, (x_n, y_n), where x_i is the input vector containing m features, y_i is the observed output value related to x_i, and n represents the number of samples in the data set. The regression function of SVM is constructed as follows:

$$f(x)=\langle w,x\rangle+b \tag{1}$$

where w is a vector of weights in a feature space with the same dimension of x, b is the bias term, and $\langle *, *\rangle$ denotes the inner product. The regression problem can be expressed as a process to minimize the following regularized risk function with ϵ-insensitivity loss function:

$$\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i*) \tag{2}$$

$$\text{subject to:}\begin{cases} y_i-\langle w,x_i\rangle-b\le\varepsilon+\xi_i\\ \langle w,x_i\rangle+b-y_i\le\varepsilon+\xi_i*\\ \xi_i,\ \xi_i*\ \ge 0\end{cases}$$

where i = 1, 2, ···, n, ξ_i and ξ_i* are the two slack variables to form the distance from actual values to the corresponding boundary values of ε, C is a constant that determines the penalty for the prediction error higher than ε. This optimization problem is often transformed into a quadratic programming problem by using Lagrangian multipliers, and the form of the solution can be given by:

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i*)K(x,x_i)+b \tag{3}$$

$$\text{subject to:}\begin{cases} \sum_{i=1}^{n}(\alpha_i-\alpha_i*)=0\\ 0\le\alpha_i\le C\\ 0\le\alpha_i*\le C\end{cases}$$

where α_i and α_i^* are Lagrange multipliers. One of the basic ideas in SVM is to map the data set x_i into a higher dimensional feature space by the function ϕ. $K(\cdot)$ is the kernel function and defined as an inner product of the points $\phi(x_i)$ and $\phi(x_j)$ as follows:

$$K(x_i,x_j)=\varnothing(x_i)\cdot\varnothing(x_j) \tag{4}$$

The most popular kernel functions used in the literature are:

linear kernel: $K(x_i,x_j)=x_i^Tx_j$
polynomial kernel: $K(x_i,x_j)=(\gamma x_i^Tx_j+r)^d,\ \gamma>0$
radial basis kernel (RBF): $K(x_i, x_j)=exp\left(-\gamma\|x_i-x_j\|^2\right),\ \gamma>0$
sigmoid kernel: $K(x_i,x_j)=tanh(\gamma x_i^Tx_j+r),\ \gamma>0$

Here, γ, r, and d are kernel parameters. The performance of SVM for regression depends on a set of parameters: C, the kernel type and corresponding kernel parameters (Min & Lee 2005). In this paper, as the structure of predictors was not accurately recognizable, the four types of kernel functions were all prepared to be tried in the SVM model and the most appropriate one would be selected based on the results. The penalty parameter C and kernel function's parameters are determined by particle swarm optimization (PSO) algorithm with cross-validation, as described below.

ANNs

ANNs have been applied in many fields as useful tools to recognize patterns from complex, non-linear data sets. There are various types of neural networks aimed at different kinds of problems (Lek & Guégan 1999). The multi-layer feed-forward neural networks with the back propagation algorithm (BPNN) is one of the most popular ANNs for data prediction (Goh 1995; Mandal *et al.* 2007). The typical BPNN includes three types of neuron layers: an input layer, a hidden layer, and an output layer. The information flows unidirectionally from input layer to output layer, through the hidden layer (Pradhan & Lee 2010). As a supervised learning algorithm, the number of neurons in the input layer is equal to the number of input variables (20 water quality parameters in this study), and the neuron in the output layer is the related concentration of DO for each case. The differences between predicted outputs and observed data are defined as error values, which would be propagated backwards through the network. During the training process, the weights between the nodes in response to the errors are adjusted until the overall error value reduces below the pre-determined threshold. The number of neurons of the hidden layer (NNH) has significant influence on the result of BPNN. If the number is too small, the fault tolerance and generalization capability of the net would be bad. However, the problem of over-fitting will exist when the hidden layer has a large number of neurons. In this study, the upper and lower bounds of NNH were calculated based on first the empirical formula:

$$\begin{cases} N < M - 1 \\ N_1 = \sqrt{M+J} + a \\ N_2 = log_2 M \end{cases} \tag{5}$$

where N was NNH, M was the number of neurons of the input layer (16), J was the number of neurons of the output layer (1), a was a constant from 1 to 10. The value range of NNH was defined as 4–14 according to the values of N, N_1, and N_2 in this study. A loop was designed to be nested into the optimized process and the training processes replicated several times, starting with 4 neurons and then increasing the number up to 14. The sum of the absolute values of errors (SAE) were monitored and the number with the minimum SAE chosen. The choices of initial values of the network connection weights and NNH are very important to the convergent behavior of the BPNN, and PSO algorithm with cross-validation is chosen as the optimized method to improve the model performance, as described below.

PSO algorithms

The PSO was developed by Eberhart and Kennedy in 1995 as a population-based stochastic optimization technique motivated by the social behavior of bird flocking (Eberhart & Shi 2001). The PSO algorithm is similar to the GA, which is also an optimization tool with broad applications. They are both initialized with a population of random solutions and evaluated by the fitness values through the updating of generations. However, the PSO has no complicated evolutionary operators, such as crossover and mutation, which is used in the GA. Each particle in the PSO has its individual memory and knowledge used to find the best solution, while the GA has no memory and previous knowledge will be destroyed with the change of population. In most cases, the convergence rates of the PSO are faster than the GA. With a strong ability to find the most optimistic result, the PSO is chosen as the optimization algorithm in this study. In PSO, a set of potential solutions are 'bird flock' with a number of individuals (particles). Each particle is given a random velocity and position at first, and then its individual memory and knowledge gained by the swarm are used as a whole to find the best solution in multidimensional space. A fitness function is defined to calculate fitness values for all the particles during their processes of movement (Chau 2006; Fei et al. 2009). The best position of each particle in the problem space is obtained by keeping track of its previous best position, which is stored as p_{best}. Then the new position that the particle should fly to is calculated through p_{best} and the best position of its neighbors (g_{best}). The dimension of the searching space is defined as D, the total number of particles is n, and the position of ith particle is represented as vector $X_i = (x_{i1}, x_{i2}, \cdots, x_{iD})$. The velocity of ith particle is defined as vector-$V_i = (v_{i1}, v_{i2}, \cdots, v_{iD})$ and the updating processes could be described as:

$$v_{id}(k+1) = \omega v_{id}(k) + c_1 rand \times [p_{bestid} - x_{id}(k)] + c_2 rand \times [g_{bestid} - x_{id}(k)] \tag{6}$$

then

$$x_{id}(k+1) = x_{id}(k) + v_{id}(k+1) \tag{7}$$

where k and $k+1$ represent the iteration count, c_1 and c_2 are the acceleration coefficients with positive values, *rand* is a random number between 0 and 1, ω is the inertia weight representing the degree of the current velocity of the particle influenced by its previous one. The PSO is a flexible algorithm to be combined with SVM and BPNN for better model performances (Zhang *et al.* 2007; Lin *et al.* 2008). The fitness functions of the *i*th particle in both models are expressed in terms of an output error as follows:

$$f(x_i) = \sum_{k=1}^{S} t_k - p_k(x_i) \tag{8}$$

where f is the fitness value, S is the number of training samples, t_k is the target output (observed values), p_k is the predicted output based on x_i. In the PSO-BPNN, x_i indicates the connection weight matrixes between the input layer and hidden layer, as well as between the hidden layer and output layer. In the PSO-SVM, x_i indicates the penalty parameter and kernel function's parameter. The main goal of the optimization is to search the best parameters that produce the most accurate predictions for different models.

Leave-one-out cross-validation

Cross-validation is a popular statistical method to evaluate and compare learning algorithms by dividing data into two segments, in which one part is used to train a model and the other is used to validate the same model (Bengio & Grandvalet 2004). The basic and most used form of cross-validation is *k*-fold cross-validation. In *k*-fold cross-validation, the training data are randomly partitioned into *k* mutually exclusive subsets with approximately equal size. A different subset is held-out for validation while the remaining $k-1$ subsets are used for model training at each time. This process is repeated *k* times, and the estimated parameters and accuracy are derived by averaging the runs (Diamantidis *et al.* 2000). The goal of cross-validation is to improve the generalization ability by defining several independent data sets to test the model, in order to limit overfitting problem, especially when the size of the training data are small or the number of parameters in the model is large (Prechelt 1998). However, it is always difficult to determine the number of *k* for common *k-fold* cross-validation, as the results may have considerable bias (Kohavi 1995). However, when *k* is significantly increased to a much larger number, the condition improves. The most extreme form of *k-fold* cross-validation, when *k* is given by the number of training patterns, is known as *leave-one-out* cross-validation, which has been shown to provide an 'almost' unbiased estimate of the true generalization ability of the model (Cawley & Talbot 2004). Although the key disadvantage of *leave-one-out* cross-validation is the high computational cost, it was overcome and combined into MLR, BPNN, and SVM, respectively.

In this study, the data set with 969 samples were divided into a training subset and testing subset at first. The training subset with 769 samples was used to obtain the best parameters of MLR, PSO-BPNN, and PSO-SVM through *leave-one-out* cross-validation method. The testing subset, including 200 samples, were randomly extracted from the whole data pool to be used for the comparison of predictive capacities among different models. All of the calibration and the following predicting work were performed by programming codes in the MATLAB R2013b. A library for support vector machines developed by Lin *et al.* (Chang & Lin 2011) was used to design the PSO-SVM model.

Models' performance criteria

In this study, the performances of MLR, PSO-BPNN, and PSO-SVM were examined on the prediction of DO concentrations and assessed by two standard statistical performance evaluation criteria, coefficient of determination (R^2) and mean squared error (MSE). The degree of correlation between the observed and predicted values is defined as R^2 and described as follows:

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(DO_{io} - \overline{DO_o}\right)\left(DO_{ip} - \overline{DO_p}\right)\right]^2}{\sum_{i=1}^{n}\left(DO_{io} - \overline{DO_o}\right)^2 \sum_{i=1}^{n}\left(DO_{ip} - \overline{DO_p}\right)^2} \tag{9}$$

The MSE is an estimator measuring the difference between observed data and predicted data, which can be calculated as:

$$MSE = \frac{\sum_{i=1}^{n}\left(DO_{io} - DO_{ip}\right)^2}{n} \tag{10}$$

where n is the number of input samples, DO_{io} and DO_{ip} are observed and predicted DO concentrations of sample i, respectively, $\overline{DO_o}$ and $\overline{DO_p}$ are the mean values of observed and predicted DO concentration. The best fit between observed and predicted values was $R^2 = 1$, MSE = 0.

RESULTS AND DISCUSSION

In order to get the effective performance evaluation of the MLR, PSO-BPNN, and PSO-SVM, the best and stable model structures had been obtained by repeated attempts and training. The NNH of PSO-BPNN was decided as 5 according to the result of SAE, shown in Figure 1. The results obtained from PSO-SVM with different types of kernel function are summarized in Table 3. During both training and testing processes, the RBF kernel showed the best performance. Thus, the RBF kernel was chosen as the most appropriate kernel function in the PSO-SVM model.

The results of the three models with best fit structures during training and testing periods are summarized in Table 4. The performance of the MLR was not as good as the other two non-linear methods with unacceptable R^2 values in both training and testing periods. As the concentration of DO was a complex outcome influenced by a series of factors, it was not able to describe the combining effect as a linear relationship. Differently from the ANN and SVM, which could not reveal the functional relationships between the target and predictor variables and always referred to as 'black box' approaches, the MLR was able to define the coefficient of each parameter; however, the results would be meaningful only if they could satisfy a series of statistical criteria. The importance of input variables could be defined by *leave-one-out* cross-validation method for the ANN and SVM, which was not the objective of this study. The PSO-BPNN gave higher R^2 values than the MLR in both processes, while the difference between MSEs was very small. The PSO-SVM had the best results for both indices among the three models with higher R^2 and lower MSE. The line charts obtained by using the MLR, PSO-BPNN, and PSO-SVM during the testing period are

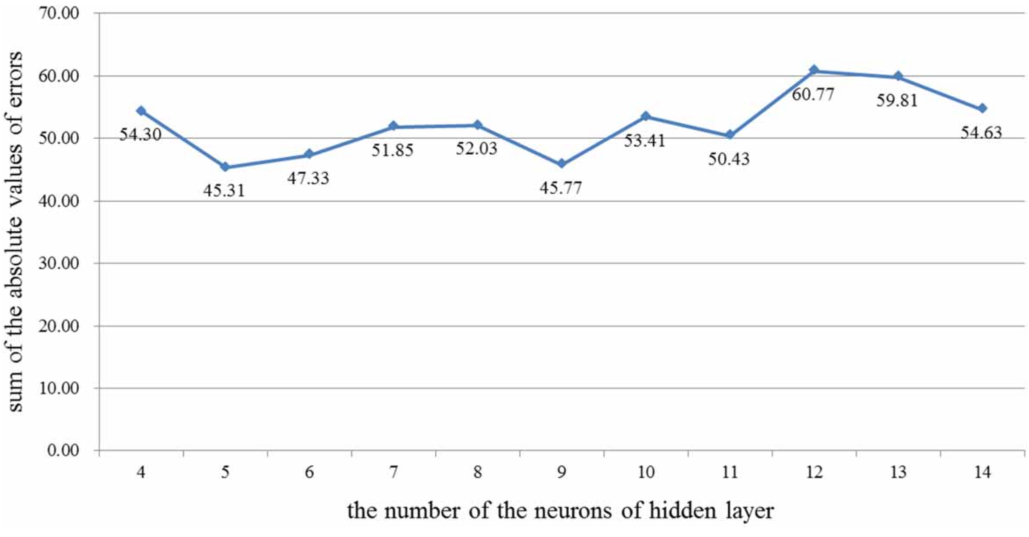

Figure 1 | The sum of the absolute values of errors with different number of neurons of the hidden layer in PSO-BPNN.

Table 3 | The performance statistics of PSO-SVM using different types of kernel functions during training and testing periods

	Training				**Testing**			
	Linear	**Polynomial**	**RBF**	**Sigmoid**	**Linear**	**Polynomial**	**RBF**	**Sigmoid**
R^2	0.50	0.73	0.76	0.14	0.55	0.43	0.74	0.13
MSE	1.56	1.67	1.64	3.56	1.74	2.20	1.62	3.16

Table 4 | The performance statistics of MLR, PSO-BPNN, and PSO-SVM during training and testing periods

	Training			Testing		
	MLR	PSO-BPNN	PSO-SVM	MLR	PSO-BPNN	PSO-SVM
R^2	0.20	0.69	0.76	0.22	0.63	0.74
MSE	1.68	1.65	1.64	1.93	1.80	1.62

shown in Figures 2–4. It can be seen that the estimated DO concentrations of the PSO-SVM were closer to the corresponding observed values than the other two models.

The absolute relative errors in the testing period were calculated for the three models. The results showed that the PSO-SVM had 59.0% estimates lower than 10% relative error, while the percent of the PSO-BPNN and the MLR

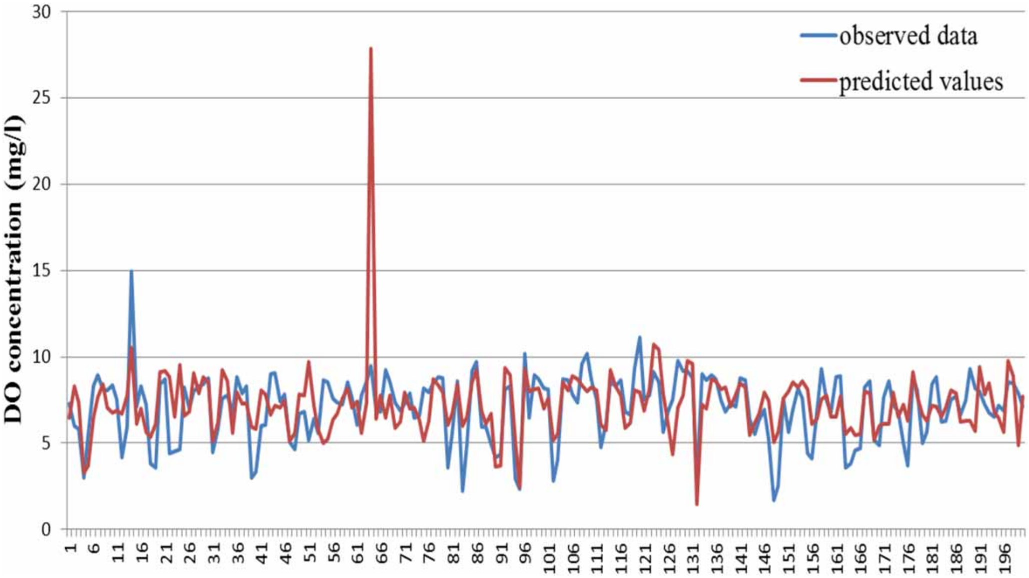

Figure 2 | The observed and predicted DO concentrations by MLR in the testing period.

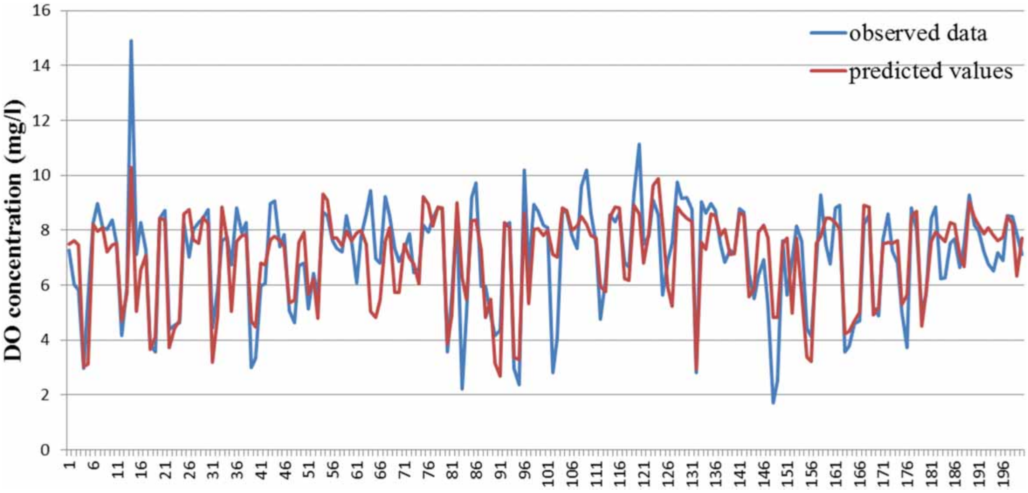

Figure 3 | The observed and predicted DO concentrations by PSO-BPNN in the testing period.

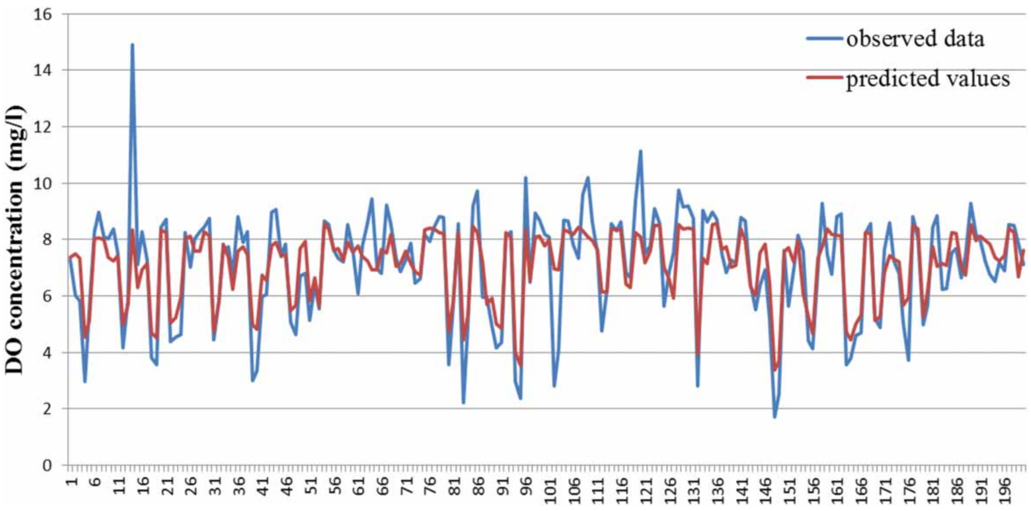

Figure 4 | The observed and predicted DO concentrations by PSO-SVM in the testing period.

were 52.5% and 39%, respectively (Table 5). There were five, three, and two predicted cases with relative error higher than 100% for the MLR, the PSO-BPNN, and the PSO-SVM, respectively (Figure 5). The diamond markers indicate the MLR, and samples 39, 149, 83, 148, and 64 had the largest predicted errors. The rectangle markers indicted the PSO-BPNN, and samples 102, 148, and 83 had the largest predicted errors. The PSO-SVM was represented by triangle markers, where the largest errors were for samples 83 and 102, which were also included in the other two models. The samples with larger errors were similar among the three models, all of which had lower DO concentrations than the others except for sample 64. Sample 64 was collected from the Liao River, Northeast China. The water temperature in this basin was lower than other basins in China and was the reason for the larger error in the MLR model. The other samples with largest predicted errors were all seriously polluted, while the TN concentrations were all higher than 20 mg/L, TP concentrations were all higher than 1 mg/L, and DO concentrations were all lower than 3 mg/L. These river reaches were faced with the process of severe eutrophication. The values of most parameters in these five samples were much higher than the mean value, which indicated that all the three models were not good at predicting outlier samples. The MLR had a higher mean relative error than the other two methods, while the lowest and highest relative errors were all estimated by the PSO-BPNN. The large differences between the estimated accuracies of the PSO-BPNN indicated a relatively instable predicted capacity for diverse cases of ANNs. The PSO-SVM model showed better performance than the other models from the relative error viewpoint, which indicated it was the most effective model in terms of predicting DO concentration accurately during the testing process.

Table 5 | The statistics of relative errors (%) for MLR, PSO-BPNN, and PSO-SVM in the testing period

	Min value	Mean value	Max value	Percent lower than 10
MLR	0.48	22.16	172.36	39%
PSO-BPNN	0.01	15.00	183.65	52.5%
PSO-SVM	0.02	13.00	147.36	59.0%

However, the variation in the relative error curve of the PSO-SVM model was also remarkable, where there was a huge difference between the minimum and the maximum value. In order to understand the distributions of absolute relative errors more clearly, the relationship between observed data and absolute relative errors was studied for the PSO-SVM model (Figure 6). There were 11.5% absolute relative errors larger than 25%. The number of concerned samples was 23, in which six samples were distributed in Tai Lake, six samples were in Huai River, three samples were in Huang River, and the other four samples were distributed in Chao Lake, Dianchi, Liao River, and inland rivers, respectively. The DO concentrations of these samples

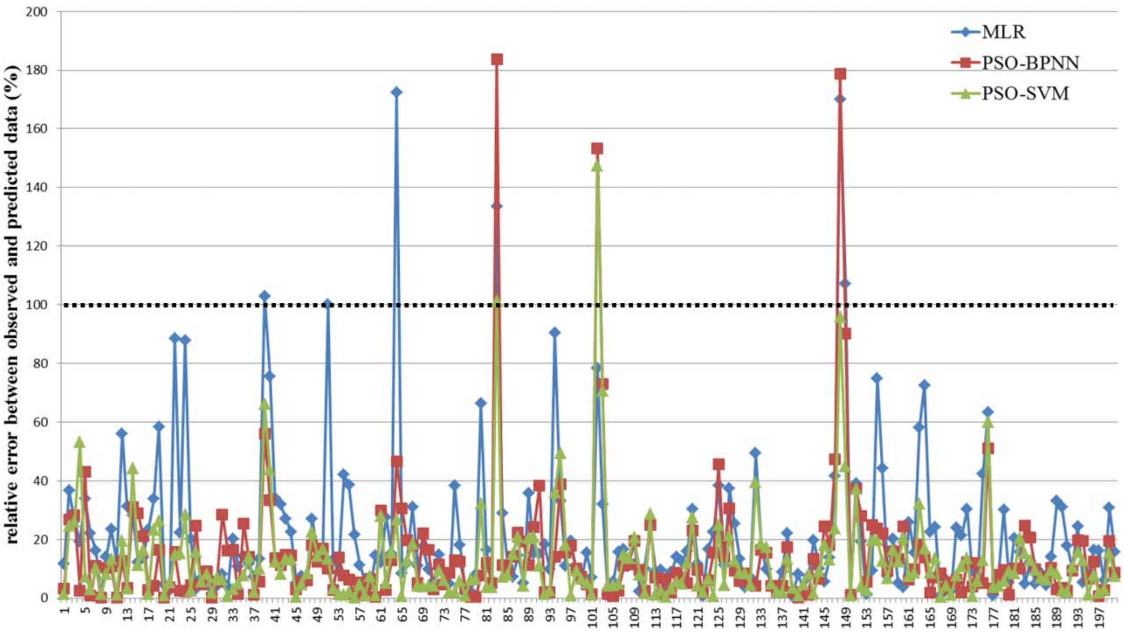

Figure 5 | The absolute relative errors between observed and predicted values in the testing period.

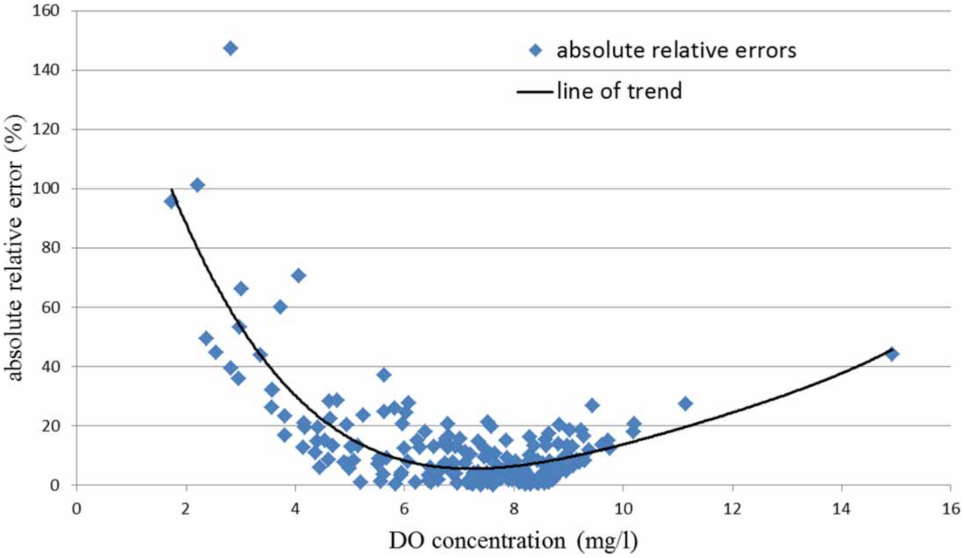

Figure 6 | The relationship between observed DO concentrations and absolute relative errors of PSO-SVM.

all deviated from the mean value of the testing data set (7.07 mg/L). Among them, 87.0% corresponding observed DO concentrations were lower than 6 mg/L, while the others were higher than 8 mg/L. This indicated that the PSO-SVM model overestimated the lowest values and underestimated the highest values. According to the distributions of absolute relative errors in Figure 4, the other two models had the same characteristics. The data-driven models showed good prediction accuracy for data around average values but were unable to maintain their accuracy for extreme values. Regardless of errors at extreme value points, the results suggested that the PSO-SVM model was superior to the MLR and PSO-BPNN in the prediction of DO concentration.

CONCLUSION

In this study, the BPNN and SVM optimized by PSO as well as the MLR models were developed to predict DO concentration based on water quality parameters. To achieve this objective, 969 samples collected during 2009 and 2010 from rivers in China were selected as input data. The training process was mainly used to calibrate the parameters in the models. In order to avoid over-fitting problem and biased estimation, the *leave-one-out* cross-validation was applied for the three models during the training process. The results of the testing period reflected the prediction and generalization capabilities, while 20% input data were separated from the whole to test the models. The statistical criteria obtained from the three models with best fit structures during both training and testing periods showed that the PSO-BPNN and PSO-SVM had better predicted performances than linear regression methods, which suggested non-linear relationships between DO concentration and water quality parameters.

The absolute relative errors calculated for the testing results showed that the PSO-SVM had minimum mean error, while the MLR had the maximum one. The PSO-BPNN with more violent error fluctuation indicated less stable predicted capacity than the PSO-SVM, but the distribution of absolute relative errors was not smooth even for the PSO-SVM. The analysis between observed values and absolute relative errors showed that the model had better predictive capacities when the data were close to average values. The errors at the points with extreme values were much larger, indicating inaccurate predictions of the data-driven models. The samples with larger predictive errors were similar for the three models, and had the lowest observed DO concentrations. Overall, the PSO-SVM model was superior to the MLR and PSO-BPNN in the DO prediction based on multiple physical and chemical parameters. It can be considered to predict DO levels with limited knowledge of other information for the rivers in China.

However, improvement will be necessary in future research. For example, the prediction accuracy, especially for the extreme values, needs to be improved by combining other model parameters. The factors influencing DO concentration of the rivers distributed in various regions may be different. The selection of suitable predictors would lead to more efficient models and accurate results.

ACKNOWLEDGEMENTS

The Chinese Academy for Environmental Planning is acknowledged for the support of water quality data. This work was supported by Tianjin Normal University Doctor Foundation (52XB1517), the innovation team training plan of the Tianjin Education Committee (TD12-5037) and the National Natural Science Foundation of China (No. 41372373). Comments and suggestions from the anonymous reviewers and the editor are greatly appreciated.

REFERENCES

Baylar, A., Hanbay, D. & Batan, M. 2009 Application of least square support vector machines in the prediction of aeration performance of plunging overfall jets from weirs. *Expert Systems with Applications* **36** (4), 8368–8374.

Bengio, Y. & Grandvalet, Y. 2004 No unbiased estimator of the variance of k-fold cross-validation. *The Journal of Machine Learning Research* **5**, 1089–1105.

Bonansea, M., Bazán, R., Ledesma, C., Rodriguez, C. & Pinotti, L. 2015 Monitoring of regional lake water clarity using Landsat imagery. *Hydrology Research* **46** (5), 661–670.

Cao, J., Chu, Z., Du, Y., Hou, Z. & Wang, S. 2016 Phytoplankton dynamics and their relationship with environmental variables of Lake Poyang. *Hydrology Research* **47** (S1), 249–260.

Carlyle, G. & Hill, A. 2001 Groundwater phosphate dynamics in a river riparian zone: effects of hydrologic flowpaths, lithology and redox chemistry. *Journal of Hydrology* **247** (3), 151–168.

Cawley, G. C. & Talbot, N. L. 2004 Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks* **17** (10), 1467–1475.

Chang, C.-C. & Lin, C.-J. 2011 LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2** (3), 27.

Chau, K. 2006 Particle swarm optimization training algorithm for ANNs in stage prediction of Shing Mun River. *Journal of Hydrology* **329** (3), 363–367.

Collins, A., Ohandja, D.-G., Hoare, D. & Voulvoulis, N. 2012 Implementing the Water Framework Directive: a transition from established monitoring networks in England and Wales. *Environmental Science & Policy* **17**, 49–61.

Cox, B. 2003 A review of currently available in-stream water-quality models and their applicability for simulating dissolved oxygen in lowland rivers. *Science of the Total Environment* **314**, 335–377.

Diamantidis, N., Karlis, D. & Giakoumakis, E. A. 2000 Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence* **116** (1), 1–16.

Eberhart, R. C. & Shi, Y. 2001 Particle swarm optimization: developments, applications and resources. In: *Proceedings of* *the 2001 Congress on Evolutionary Computation*, Vol. 1, Seoul, Korea, pp. 81–86.

Fei, S.-W., Wang, M.-J., Miao, Y.-b., Tu, J. & Liu, C.-l. 2009 Particle swarm optimization-based support vector machine for forecasting dissolved gases content in power transformer oil. *Energy Conversion and Management* **50** (6), 1604–1609.

Ficklin, D. L., Stewart, I. T. & Maurer, E. P. 2013 Effects of climate change on stream temperature, dissolved oxygen, and sediment concentration in the Sierra Nevada in California. *Water Resources Research* **49** (5), 2765–2782.

Goh, A. 1995 Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering* **9** (3), 143–151.

He, Z., Wen, X., Liu, H. & Du, J. 2014 A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *Journal of Hydrology* **509**, 379–386.

Hosseini, S. M. & Mahjouri, N. 2014 Developing a fuzzy neural network-based support vector regression (FNN-SVR) for regionalizing nitrate concentration in groundwater. *Environmental Monitoring and Assessment* **186** (6), 3685–3699.

Isunju, J. B. & Kemp, J. 2016 Spatiotemporal analysis of encroachment on wetlands: a case of Nakivubo wetland in Kampala, Uganda. *Environmental Monitoring and Assessment* **188** (4), 1–17.

Kannel, P. R., Lee, S., Lee, Y.-S., Kanel, S. R. & Khan, S. P. 2007 Application of water quality indices and dissolved oxygen as indicators for river water classification and urban impact assessment. *Environmental Monitoring and Assessment* **132** (1–3), 93–110.

Klose, K., Cooper, S. D., Leydecker, A. D. & Kreitler, J. 2012 Relationships among catchment land use and concentrations of nutrients, algae, and dissolved oxygen in a southern California river. *Freshwater Science* **31** (3), 908–927.

Kohavi, R. 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai'95 Proceedings of the 14th Joint Conference on Artificial Intelligence*, Vol. 2, pp. 1137–1143.

Kuo, J.-T., Hsieh, M.-H., Lung, W.-S. & She, N. 2007 Using artificial neural network for reservoir eutrophication prediction. *Ecological Modelling* **200** (1), 171–177.

Lek, S. & Guégan, J.-F. 1999 Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling* **120** (2), 65–73.

Li, B.-D., Zhang, X.-H., Xu, C.-Y., Zhang, H. & Song, J.-X. 2015 Water balance between surface water and groundwater in the withdrawal process: a case study of the Osceola watershed. *Hydrology Research* **46** (6), 943–953.

Lin, S.-W., Ying, K.-C., Chen, S.-C. & Lee, Z.-J. 2008 Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications* **35** (4), 1817–1824.

Liu, S., Xu, L., Li, D., Li, Q., Jiang, Y., Tai, H. & Zeng, L. 2013 Prediction of dissolved oxygen content in river crab culture based on least squares support vector regression optimized by improved particle swarm optimization. *Computers and Electronics in Agriculture* **95**, 82–91.

Mandal, D., Pal, S. K. & Saha, P. 2007 Modeling of electrical discharge machining process using back propagation neural network and multi-objective optimization using non-dominating sorting genetic algorithm-II. *Journal of Materials Processing Technology* **186** (1), 154–162.

Meyer, D., Leisch, F. & Hornik, K. 2003 The support vector machine under test. *Neurocomputing* **55** (1), 169–186.

Min, J. H. & Lee, Y.-C. 2005 Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications* **28** (4), 603–614.

Modaresi, F. & Araghinejad, S. 2014 A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. *Water Resources Management* **28** (12), 4095–4111.

Pradhan, B. & Lee, S. 2010 Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia. *Landslides* **7** (1), 13–30.

Prechelt, L. 1998 Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks* **11** (4), 761–767.

Rounds, S. A. 2002 Development of a neural network model for dissolved oxygen in the Tualatin River, Oregon. In: *Proceedings of the Second Federal Interagency Hydrologic Modeling Conference*, Las Vegas, Nevada, USA.

Salami Shahid, E. & Ehteshami, M. 2016 Application of artificial neural networks to estimating DO and salinity in San Joaquin River basin. *Desalination and Water Treatment* **57** (11), 4888–4897.

Sear, D. A., Pattison, I., Collins, A. L., Newson, M. D., Jones, J., Naden, P. & Carling, P. A. 2014 Factors controlling the temporal variability in dissolved oxygen regime of salmon spawning gravels. *Hydrological Processes* **28** (1), 86–103.

Smola, A. J. & Schölkopf, B. 2004 A tutorial on support vector regression. *Statistics and Computing* **14** (3), 199–222.

Stefan, H. G. & Fang, X. 1994 Dissolved oxygen model for regional lake analysis. *Ecological Modelling* **71** (1), 37–68.

Surinaidu, L. 2016 Role of hydrogeochemical process in increasing groundwater salinity in the central Godavari delta. *Hydrology Research* **47** (2), 373–389.

Wen, X., Fang, J., Diao, M. & Zhang, C. 2013 Artificial neural network modeling of dissolved oxygen in the Heihe River, Northwestern China. *Environmental Monitoring and Assessment* **185** (5), 4361–4371.

Were, K., Bui, D. T., Dick, Ø. B. & Singh, B. R. 2015 A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators* **52**, 394–403.

Zhang, J.-R., Zhang, J., Lok, T.-M. & Lyu, M. R. 2007 A hybrid particle swarm optimization–back-propagation algorithm for feedforward neural network training. *Applied Mathematics and Computation* **185** (2), 1026–1037.

First received 16 May 2016; accepted in revised form 17 August 2016. Available online 6 October 2016