Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-018-2147-3

RESEARCH ARTICLE

Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake

Xue Li[1] · Jian Sha[1] · Zhong-Liang Wang[1]

Received: 24 February 2018 / Accepted: 25 April 2018

Abstract
As a representative index of the algal bloom, the concentration of chlorophyll-a (Chl-a) is a key parameter of concern for environmental managers. The relationships between environmental variables and Chl-a are complex and difficult to establish. Two machine learning methods, including support vector machine for regression (SVR) and random forest (RF), were used in this study to predict Chl-a concentration based on multiple variables. To improve the model accuracy and reduce the input number, two feature selection methods, including minimum redundancy and maximum relevance method (mRMR) and RF, were integrated with regression models. The results showed that the RF model had a higher predictive ability than the SVR model. Furthermore, the less computational time cost and unnecessary prior data transformation also indicated a better applicability of the RF model. The comparison between ensemble models of mRMR-RF and RF-RF showed that the RF-RF yielded a better performance with fewer variables. Seven variables selected from the candidate predictors could interpret most information, and their potential implications to Chl-a were discussed based on the level of importance. Overall, the RF-RF ensemble model can be considered as a useful approach to determine the significant stressors and achieve satisfactory prediction of Chl-a concentration.

Keywords Feature selection · Random forest · Minimum redundancy and maximum relevance · Support vector machine

Introduction

Algal blooms have occurred in many lakes around the world in recent years, leading to the death of aquatic life and potential deterioration of water quality in freshwater environments (Paerl and Paul 2012; Park et al. 2015). As a representative index of algal biomass, the concentration of chlorophyll-a (Chl-a) has always been a key parameter concerned by environmental managers. Due to the complex relationships between multiple factors and algal community, it is difficult to directly predict Chl-a concentration in practice (Lee et al. 2003). The process-based mathematical models, which focus on the study of the ecological response of phytoplankton to environmental factors and underlying physical laws, have been subjected to the uncertainty of kinetic rate coefficients for different algal species (Lee and Lee 1995). To improve the process-based model accuracy, a coupling of process-based and data-driven approaches through so-called data assimilation was proposed to capture the dynamic variation of water quality (Wang et al. 2016). With the improvement of computing technology, machine learning (ML) methods are the most popular data-driven approaches in recent years and have been successfully developed to model nonlinear and complex relationships (Babovic 2005; Babovic et al. 2001; Li et al. 2017b, d). In this study, support vector machine for regression (SVR) and random forest (RF) was selected among current ML models and compared as regression models to predict Chl-a concentration in a shallow lake.

The SVR is a widely used method that has been applied to various water resource problems, and it has proven to be a robust technique in establishing nonlinear relationships (He et al. 2014; Modaresi and

Responsible editor: Philippe Garrigues

✉ Zhong-Liang Wang
wangzhongliang@vip.skleg.cn

[1] Tianjin Key Laboratory of Water Resources and Environment, Tianjin Normal University, Tianjin 300387, China

Published online: 05 May 2018

Araghinejad 2014; Yu et al. 2004). The RF as a new method of data mining has several advantages compared to classical ML techniques, such as the resistance to over fitting problem and small number of user-defined parameters (Were et al. 2015). However, the performances of these ML methods are often limited, due to the unnecessary input of variables. The input feature selection is a required step in modeling non-linear and complex systems (Hejazi and Cai 2009). It could lead to shorter training times, simplification of models, avoidance of dimension curse, and enhanced generalization. There are many approaches to select features for ML techniques, and among them, mutual information-based approaches are an important feature selection paradigm in data mining (Li et al. 2017a). Peng et al. (2005)) proposed a minimum redundancy and maximum relevance framework (mRMR) based on mutual information for feature selection, and it has been widely used in areas including economics, industry, engineering, and ecological systems (Bao-Gang and Yong 2008; Galelli and Castelletti 2013; Tsanas et al. 2012). The RF method, with the measures of variable importance for each candidate predictor, could also be used for feature selection (Genuer et al. 2010). Based on model aggregation ideas, the quantification of the variable importance allows the RF to rank the explanatory variables (Breiman 2001). The mRMR and RF were both proven to be effective feature selection methods in previous studies; we were interested in which method was better for the prediction of Chl-a concentration, and whether combining multiple methods has any effect on the prediction performance.

In this study, a detailed assessment for two feature selection methods and two regression methods was performed to evaluate the predictabilities of diverse ensemble models. The general strategy of this study was to generate two feature sequences by mRMR and RF at first, and then evaluate input sub-sets using a stepwise ascending method through the comparison of regression models' performances. The main objectives of this study were as follows: (1) to predict Chl-a concentration using the SVR and RF methods and compare their results to suggest the one with better performance and (2) to examine the effectiveness of mRMR and RF and select the optimal input sub-set among candidate predictors for Chl-a prediction.

Material and methods

Study area and dataset

The Baiyangdian Lake, with an area of 366 km^2, is located in the central of the North China Plain (38°43′–39°02′N, 115°38′–116°07′E) (Fig. 1). Under the influence of semiarid monsoon climate, the average air temperature is 12.1 °C, and the mean annual precipitation is approximately 550 mm, with nearly 70% occurring from July to September (Yuan et al. 2017). There are six main rivers flowing into the Baiyangdian Lake, some of which have become seasonal or persistently dried out since the last decade (Yang et al. 2016). The alteration of these rivers leads to a decrease in both volume and area of the lake. The average annual inflow to Baiyangdian Lake was 1.94 billion m^3 in the 1950s, and it was 0.065 billion m^3 in the 2000s (Yang et al. 2017). Moreover, with the economic development and population increase in recent years, a large amount of domestic sewage and pollutants directly discharge into the lake. The decreased water area, combined with elevated water pollutants, is destroying the service functions (such as regulating the water cycle, supporting biodiversity, and maintaining the ecological balance) of the lake and leading to the deleterious proliferation of planktonic algae.

The data set used in this study was collected from eight monitoring sites distributed in the lake (Fig. 1) during the period from May 2006 to December 2010. The local environmental monitoring center collected water samples monthly and 13 water quality parameters were analyzed according to the national standards for surface waters in China (GB3838-2002) in local laboratories. The parameters included chlorophyll-a (Chl-a), water temperature (Temp), pH, clarity (SD), dissolved oxygen (DO), potassium permanganate index (COD_{Mn}), biochemical oxygen demand (BOD), ammonia nitrogen (NH_3-N), petroleum, total nitrogen (TN), total phosphorus (TP), and chemical oxygen demand (COD_{Cr}). The mean values and standard deviations (STD) of the 13 parameters are summarized in Table 1. The Pearson correlation coefficients were calculated to assess linear relationships between Chl-a and the other variables and also shown in Table 1. All of the correlation coefficients were lower than 0.5, which indicated that there was no strong linear correlation between Chl-a and any other variable. Due to the lack of data in some individual months and variables, the final dataset included 391 records and the records with missing data were not taken into account.

Minimal redundancy and maximal relevance

The minimum redundancy and maximum relevance (mRMR) is a method based on mutual information (MI) to rank features based on both their relevance to the target and the redundancy among features (Gao et al. 2013). As a quantified method, MI could estimate the information one random variable explains

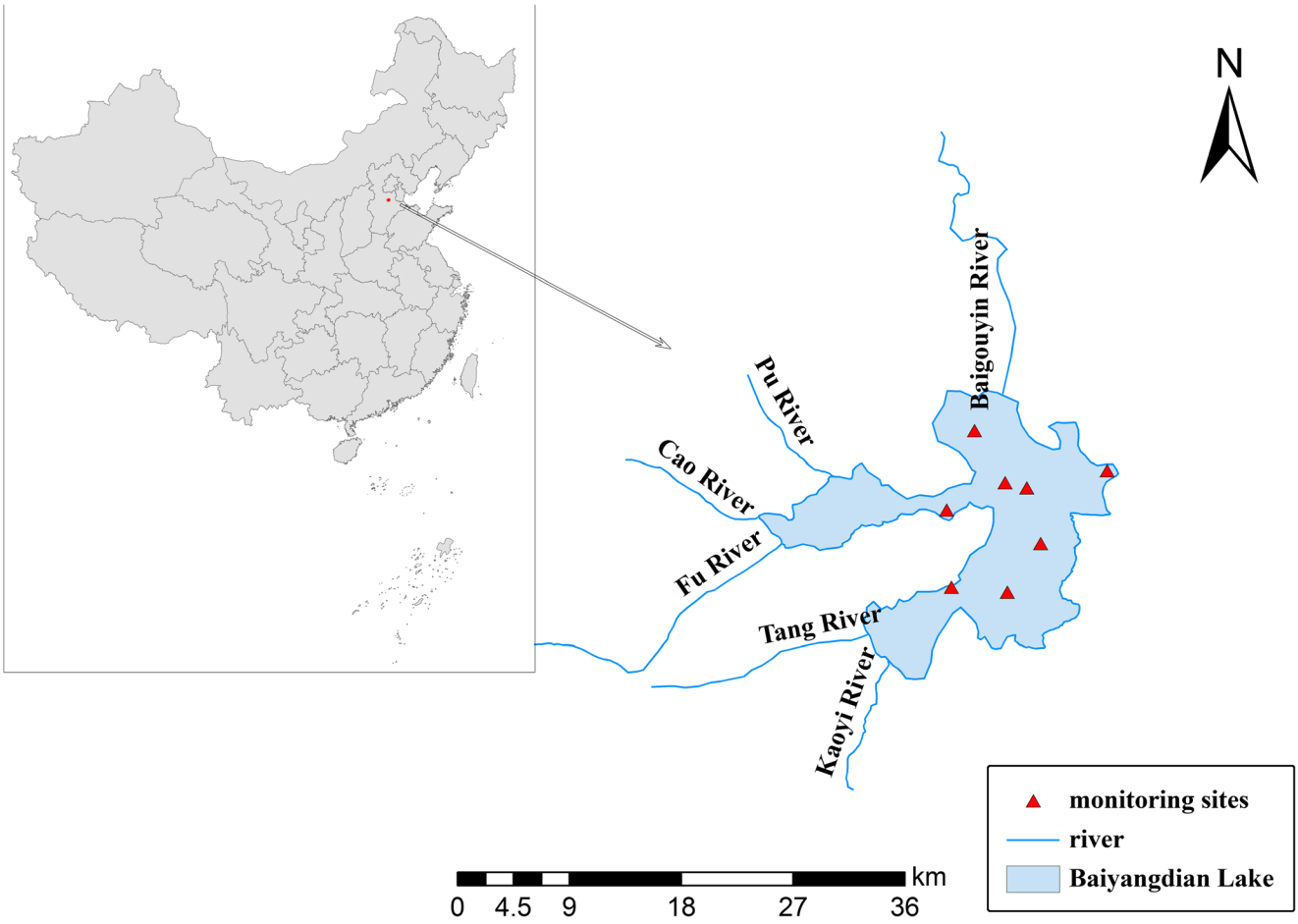

Fig. 1 Location and monitoring sites in the study area

about the other (Babovic and Keijzer 2000; Hejazi and Cai 2009). Given two random discrete variables x and y, the MI equation was defined as follows:

$$I(x,y) = \iint p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy \quad (1)$$

where $p(x,y)$ is the joint probability density function of x and y and $p(x)$ and $p(y)$ are the marginal probability density functions of x and y, respectively. The whole input set is defined as S with m features (S_m), based on MI, the relevance D of S_m for the target y is defined by the average value of all mutual information values between the individual feature x_i and the target y as follows:

$$D(S_m,y) = \frac{1}{m} \sum_{x_i \in S_m} I(x_i;y) \quad (2)$$

The redundancy R of all features in the set S_m is the average value of mutual information values between the feature x_i and x_j:

$$R(S_m) = \frac{1}{m^2} \sum_{x_i, x_j \in S_m} I(x_i;x_j) \quad (3)$$

The mRMR criterion is a combination of the two equations mentioned above and defined as $\max_{S_m} [D(S_m,y) - R(S_m)]$. The incremental search method proposed by Peng et al. (2005)) is used to find the near-optimal features in this study. J_{n-1} denotes the selected feature set containing $n-1$ features, and the nth feature is aimed to be selected out from the remaining $\{S_m - J_{n-1}\}$ set. The optimization problem can be expressed as follows:

$$\max_{x_j \in S_m - J_{n-1}} \left[I(x_j;y) - \frac{1}{n-1} \sum_{x_i \in J_{n-1}} I(x_j;x_i) \right] \quad (4)$$

We used the incremental selection method to select n sequential features from S_m, which lead to n sequential feature sets as $S_1 \subset S_2 \subset \cdots \subset S_n \subset \cdots \subset S_m$ ($n = 1,2,\ldots,m$). Each feature subset was applied to predict Chl-a, and the one with first peak of prediction performance would be determined.

It is often difficult to compute the integral in the continuous space based on a limited number of samples (Peng et al. 2005); therefore, the continuous variables are suggested to be discrete at first. In this study, the self-organizing map (SOM) method was used for data discretization (Li et al.

Table 1 The basic statistics of water quality parameters in Baiyangdian Lake

Parameters	Unit	Mean	Max	Min	STD	Correlation coefficient
Temp	°C	17.79	33.9	0.1	9.39	0.14[a]
pH	Dimensionless	8.12	9.24	7.19	0.30	0.15[a]
SD	Cm	57.16	200	10	38.06	− 0.39[a]
DO	mg/l	8.57	22.1	0.62	3.44	− 0.02
COD_{Mn}	mg/l	9.07	25.8	4	2.45	0.35[a]
BOD	mg/l	3.55	25.1	1	2.91	0.44[a]
NH_3-N	mg/l	2.80	56.8	0.128	7.42	0.08
Petroleum	mg/l	0.02	0.31	0.005	0.04	0.23[a]
TN	mg/l	4.36	59.7	0.16	8.72	0.12[b]
TP	mg/l	0.21	3.95	0.005	0.49	0.34[a]
COD_{Cr}	mg/l	29.24	64	6	8.68	0.34[a]
Fluoride	mg/l	0.76	0.99	0.3	0.13	− 0.09
Chl-a	μg/l	17.04	174	0.5	25.08	1

Note: Correlation coefficient is the results of Pearson correlation coefficients between Chl-a and other variables

[a] Correlation is significant at the 0.01 level (two-tailed)

[b] Correlation is significant at the 0.05 level (two-tailed)

2018). The SOM, as a pattern recognition technique, incorporates unsupervised learning and a type of neural network (Kohonen 1998), and it is often applied to extract the inner relationship inherent in data sets to group similar data into the same or a nearby class. All of the variables were preprocessed into three categories using the SOM Toolbox in MATLAB. A Toolbox for MI developed by Peng et al. (2005)) was used to generate the mRMR feature sequence.

Random forest

The random forest (RF) method is an ensemble learning technique developed by Breiman (Breiman 2001) and has been successfully used in dealing with various prediction problems (Ellis et al. 2014; Huang et al. 2016). It is a machine-learning algorithm that combines a large set of decision trees to improve the prediction performance of classification and regression trees (CART) method (Mutanga et al. 2012). Each decision tree of RF is grown by using a randomly selected bootstrap sample from the original data set (Acharjee et al. 2011), and the final outcome of RF is the average result of all the trees (Rahmati et al. 2016). Compared to other regression methods, the number of parameters needed to be defined in the RF is very few. There are only two necessary parameters, including the number of variables used in each tree-building process (m_{try}) and the number of trees built in the forest (n_{tree}). In this study, m_{try} was calculated based on empirical formulas at first:

$$\begin{cases} m_{try} = [log_2(M) + 1] \\ m_{try} = \sqrt{M} \\ m_{try} = \dfrac{M}{3} \end{cases} \quad (5)$$

where M denoted the number of features in the original data set (12). The value of m_{try} was defined as 3 or 4 in this study. The n_{tree} has significant influence on the result of RF. The insufficient number of trees would result in poor forecasting performance, while the excessive number of trees may lead to complicated predictors (Huang et al. 2016). In this study, the initial value of n_{tree} was defined as 1, and a series of values with increment of 1 until 500 would be tried to determine the most appropriate one. Generally, a part of the overall sample is left out for validation, i.e., the out-of-bag (OOB) predictions. The OOB error is often thought as an unbiased estimation of the generalization error and calculated as follows:

$$OOB_{error} = \frac{1}{n}\sum_{i=1}^{n}\left(y_{oi} - y_{pi}\right)^2 \quad (6)$$

where n denotes the number of OOB samples, and y_{oi} and y_{pi} are observed and predicted values of sample i, respectively. The main goal of the selection of n_{tree} and m_{try} is to search the best parameters that could minimize the OOB error.

In the RF framework, the importance of a given input variable is often evaluated by the increase in mean of the error of a tree when the variable is replaced by random noise. After ranking the input variables, it is significant to select the fewest number of inputs that offer the best predictive power and the most interpretation of the final model (Mutanga et al. 2012). In this study, the incremental selection method, which was used in the mRMR method, was also applied to determine the optimal sub-feature set of Chl-a prediction for the RF. A Toolbox in MATLAB developed by Jaiantilal (2009)), which was on the basis of

Andy Liaw et al. C code (Liaw and Wiener 2002), was used to perform the RF feature selection and regression.

Support vector machine for regression

Support vector machine developed by Cortes and Vapnik (1995) is a popular supervised machine learning method for classification and regression. The main goal of support vector machine for regression (SVR) is to construct a hyperplane that could minimize the sum of the distances from the data points to the hyperplane (Trafalis and Ince 2000). More details of the mathematical formulations of SVR have been described in previous studies (Li et al. 2017d; Smola and Schölkopf 2004). The radius basis function (RBF) with high effectiveness and great speed was selected as the kernel function in this study (Vojinovic et al. 2003). During the training process, the two most influenced parameters, including the penalty parameter C and kernel function's parameter c, were optimized by the particle swarm optimization algorithm (PSO). PSO is a stochastic global optimization method that simulates the behavior of birds searching for food (Zhang et al. 2015). Each particle in the PSO is given a random initial velocity and position, and the velocity and position would be updated until the termination condition is satisfied. In the PSO-SVR, the optimal solution with the minimum value of error between observed and predicted data will be obtained, which could escape the model from getting trapped into local optima. The library for support vector machines (LIBSVM) developed by Chang and Lin (2011)) was used to develop the PSO-SVR model.

Evaluation criteria

Suppose that a series of feature subsets $S_1^1 \subset S_2^1 \subset \cdots \subset S_n^1 \subset \cdots \subset S_m$ were generated based on the feature selection method mRMR, and similar, $S_1^2 \subset S_1^2 \subset \cdots \subset S_n^2 \subset \cdots \subset S_m$ were the subsets generated from the RF method. To determine which one of the feature subsets was superior for Chl-a prediction, and simultaneously, to evaluate the predictive capacities of RF and SVR, the performances of models were assessed by three statistical evaluation criteria, including the coefficient of efficiency (CE), ratio of the root mean square error to the standard deviation of the observations (RMSE/STD), and mean absolute error (MAE). These criteria were defined as follows:

$$CE = 1 - \frac{\sum_{i=1}^{n} \left(Chl_{io} - Chl_{ip}\right)^2}{\sum_{i=1}^{n} \left(Chl_{io} - \overline{Chl_o}\right)^2} \tag{7}$$

$$RMSE/STD = \sqrt{\frac{\sum_{i=1}^{n} \left(Chl_{io} - Chl_{ip}\right)^2}{n}} \Big/ \sqrt{\frac{\sum_{i=1}^{n} \left(Chl_{io} - \overline{Chl_o}\right)^2}{n}} \tag{8}$$

$$MAE = \frac{\sum_{i=1}^{n} \left|Chl_{ip} - Chl_{io}\right|}{n} \tag{9}$$

where n was the number of samples, Chl_{io} and Chl_{ip} were the observed and predicted Chl-a concentrations of sample i, respectively. $\overline{Chl_o}$ was the mean value of observed data. When the predicted data were exactly equal to the observed data, the values of CE, RMSE/STD, and MAE were 1, 0, and 0.

Results and discussion

The observed dataset was separated into a training subset (80%) and a testing subset (20%). The RF and SVR models were calibrated with the training data and validated with the testing data, which were not used during the training period, i.e., out-of-sample forecasts. The raw data were standardized between − 1 and 1 before analysis to eliminate the effects of various dimensions for the SVR model, while the RF model did not need prior data transformation. The mRMR method was applied to the whole dataset, of which all variables were discretized into three classes based on the SOM method in advance. As an important issue existing in the data-driven models, over-fitting problem always affected the performances of models. Cross-validation was a popular statistical method to improve the generalization ability and limit over-fitting problem by defining several independent datasets to test the model, especially when the size of the training data is small. In this study, a *10-fold* cross-validation was applied to the training subset to obtain robust and stable parameters of the RF and the SVR models.

To determine the most appropriate values of m_{try} and n_{tree} for RF, the OOB (i.e., the testing subset) errors from

Table 2 The ranking results of twelve input variables by mRMR and RF methods

Parameters	Sequence number	
	mRMR	RF
Temp	6	10
pH	4	11
SD	3	2
DO	9	6
COD_{Mn}	5	4
BOD	1	1
NH_3-N	12	9
petroleum	8	12
TN	11	7
TP	7	3
COD_{Cr}	10	5
Fluoride	2	8

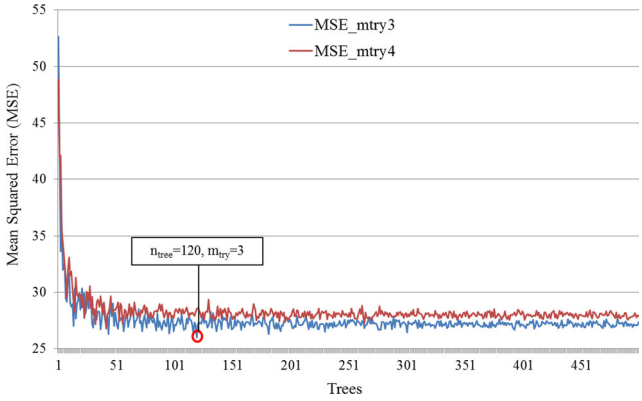

Fig. 2 Optimization values of m_{try} and n_{tree} based on OOB error (MSE) in RF model

n_{tree} 1 to n_{tree} 500 were compared between m_{try} 3 and m_{try} 4. As seen in Fig. 2, the OOB errors sharply reduced with the increase of n_{tree} from 1 to 10 and gradually tended to be steady when more trees were added to the RF. Based on the analysis of Fig. 2, the OOB errors were lower at m_{try} 3 than m_{try} 4, and the lowest OOB error was equal to 26.00, when m_{try} and n_{tree} obtained 3 and 120, respectively.

The original input set for Chl-a prediction included 12 variables. To assess the feature importance, the mRMR and RF models were applied separately to rank these variables. Based on the sequential forward selection method, the ranking results of variables are shown in Table 2. According to the Spearman correlation analysis, there were no strong correlations between Chl-a and other parameters in Baiyangdian Lake. This result was noticeably different from many previous studies, in which nutrients were significantly correlated with Chl-a concentrations and could develop linear models to link nutrient stressors to the Chl-a response (Conrad Lamon and Qian 2008; Freeman et al. 2009). The relationships between Chl-a and multiple physicochemical parameters in the study lake were nonlinear and complicated, which could be fit well by the machine learning models (Li et al. 2017c). Considering that the variables ranked by mRMR and RF were in a descending order, the variables in the front of the ranking list contained more effective information than the back ones. Although the mRMR feature sequence (mRMRFS) was not the same with the RF feature sequence (RFFS), BOD was given the first rank in both

Table 3 The performance statistics of SVR during training and testing periods for mRMR feature sequence and RF feature sequence (the optimal feature subsets were marked in italics)

Feature selection method	Number of input variables	Training period			Testing period		
		CE	RMSE/STD	MAE	CE	RMSE/STD	MAE
mRMR	1	0.36	0.80	10.11	0.33	2.12	7.57
	2	0.38	0.79	9.42	0.23	2.28	7.47
	3	0.23	0.88	12.09	0.37	2.06	7.26
	4	0.24	0.88	12.07	0.38	2.04	7.24
	5	0.31	0.84	11.17	0.23	2.28	7.72
	6	0.36	0.80	10.10	0.36	2.07	7.64
	7	0.40	0.78	9.85	0.38	2.05	7.58
	8	0.46	0.74	9.53	0.38	2.05	7.58
	9	0.52	0.69	7.21	0.44	1.94	7.62
	10	0.62	0.62	6.86	0.51	1.82	6.56
	11	0.76	0.49	4.57	0.63	1.58	6.46
	12	*0.82*	*0.42*	*3.92*	*0.74*	*1.33*	*4.03*
RF	1	0.36	0.80	10.11	0.33	2.12	7.57
	2	0.23	0.88	12.09	0.37	2.05	7.24
	3	0.30	0.84	11.35	0.16	2.38	8.00
	4	0.31	0.83	11.09	0.26	2.23	7.48
	5	0.36	0.80	9.99	0.27	2.22	7.90
	6	0.39	0.78	9.61	0.29	2.18	7.73
	7	0.68	0.57	6.15	0.43	1.96	7.48
	8	0.68	0.57	6.14	0.43	1.96	7.50
	9	0.73	0.53	5.58	0.51	1.81	7.02
	10	0.76	0.50	4.57	0.54	1.77	7.46
	11	0.76	0.49	4.55	0.66	1.50	6.11
	12	*0.82*	*0.42*	*3.92*	*0.74*	*1.33*	*4.03*

models. BOD was a widely applied indicator to quantify the consumption of oxygen in the water column from the decay of organic matter (Sullivan et al. 2010). In addition, the parameter SD was also in the front rank of both sequences. The light conditions in the lakes influenced the growth of plankton community, and excessive algae growth along with high Chl-a concentration would lead to reduced water clarity (Li et al. 2017c).

In order to assess the performances of SVR and RF, the statistical measures of these two models during training and testing periods for mRMRFS and RFFS are summarized in Table 3 and Table 4, respectively. When the input data was the same, the results showed that the performance of RF was always better than SVR. To compare the performances of SVR and RF more visually, the RMSE/STD and MAE values obtained from both mRMRFS and RFFS are presented in Fig. 3 and Fig. 4. The RMSE/STD values emphasized the penalization of large errors, while the MAE values gave the same weight to all errors and reflected the average deviation of predictions (Chai and Draxler 2014). As shown in Fig. 3a and Fig. 4a, the RMSE/STD values of SVR models (SVR-RMSE/STD) were larger than RF models (RF-RMSE/STD) during both training and testing periods, indicating that the error fluctuation range was larger, and the prediction accuracy of extreme data was poorer in SVR models. During the training period, the MAE values of SVR models (SVR-MAE) were larger than RF models (RF-MAE) at first; subsequently, SVR-MAE reduced rapidly and lower than RF-MAE since input 11 of mRMRFS and input 9 of RFFS. It indicated that the addition of input variables could improve the performance of models, especially the SVR models. However, SVR-MAEs during the testing period were larger than RF-MAEs for both feature sequences, which showed a better generalization capacity of the RF models. Moreover, the RF models did not need prior elimination of outliers or data transformation, which was easier to use than SVR models. Simultaneously, the computing time of RF models were much shorter than SVR models. Therefore, the comparison of feature selection methods and the determination of optimal input variables were performed based on RF models.

By using RF as the predictor, the feature selection methods mRMR and RF were compared based on the results of Chl-a prediction with different numbers of input variables. The parameters were sequentially added to an empty candidate set

Table 4 The performance statistics of RF during training and testing periods for mRMR feature sequence and RF feature sequence (the optimal feature subsets were marked in italics)

Feature selection method	Number of input variables	Training period			Testing period		
		CE	RMSE/STD	MAE	CE	RMSE/STD	MAE
mRMR	1	0.43	0.76	12.21	0.30	2.17	7.89
	2	0.71	0.54	8.20	0.65	1.53	4.78
	3	0.78	0.47	6.99	0.62	1.60	5.07
	4	0.81	0.44	6.55	0.66	1.51	4.64
	5	0.82	0.42	6.59	0.78	1.22	3.85
	6	0.82	0.42	6.41	0.79	1.20	3.87
	7	0.83	0.41	6.19	0.76	1.27	4.08
	8	0.84	0.40	6.11	0.79	1.20	3.42
	9	0.84	0.40	6.16	0.81	1.14	3.36
	10	*0.85*	*0.39*	*6.07*	*0.82*	*1.09*	*3.32*
	11	0.84	0.40	6.10	0.81	1.12	3.52
	12	0.85	0.38	5.47	0.82	1.22	3.34
RF	1	0.43	0.76	12.21	0.30	2.17	7.89
	2	0.67	0.58	9.06	0.32	2.14	6.68
	3	0.80	0.45	7.06	0.74	1.32	4.67
	4	0.81	0.43	6.63	0.77	1.24	4.18
	5	0.82	0.43	6.65	0.78	1.23	4.21
	6	0.82	0.43	6.59	0.78	1.21	4.02
	7	*0.85*	*0.39*	*5.78*	*0.80*	*1.21*	*3.66*
	8	0.84	0.40	6.04	0.78	1.26	3.95
	9	0.83	0.41	5.88	0.79	1.25	3.67
	10	0.84	0.40	5.75	0.80	1.21	3.76
	11	0.85	0.38	5.54	0.81	1.21	3.70
	12	0.85	0.38	5.47	0.82	1.22	3.34

until all of the 12 variables were taken into account. The objectives of variable selection were to determine the most important variables related to the response variable for interpretation purpose, as well as the number of variables sufficient for a satisfied prediction was the smallest (Genuer et al. 2010). In order to determine the optimal feature subset, the first maximum value of CE and first minimum values of RMSE/STD and MAE were searching among inputs 1–12. As shown in Table 4, the first peaks appeared at number 10 for mRMRFS (mRMRFS_10) and number 7 for RFFS (RFFS_7). This result indicated that the later addition of variables could not improve the performance of RF and may have caused adverse effects. The comparison between mRMRFS_10 and RFFS_7 showed that the performances were similar during the training period, while the result of mRMRFS_10 was a little superior to the RFFS_7 during the testing period. By examining the distribution of absolute errors between observed data and predicted data during the testing period (Fig. 5), mRMRFS_10 provided no significant advantages over RFFS_7 and the absolute error fluctuation of mRMRFS_10 was larger than RFFS_7. With the fewer number of input variables and similar estimation accuracy, the RFFS_7 was chosen as the optimal input subset for Chl-a prediction in Baiyangdian Lake.

The parameters of optimal input subset included BOD, SD, TP, COD_{Mn}, COD_{Cr}, DO, and TN. Except for the first two variables (BOD and SD) mentioned above, the third important variable was TP, which was consistent with many previous studies (Downing and McCauley 1992; Morgan et al. 2006). As significant factors for algae growth, TN and TP were always discussed together. In this study, TN contributed weaker to model outputs than TP. The average concentrations of TN and TP were 4.36 and 0.21 mg/l in Baiyangdian Lake, both of which exceeded the criterion of grade V ([TN] > 2.0 mg/l, [TP] > 0.2 mg/l in lakes) in the Chinese System of Environmental Quality Standards for Surface Water. The excess input of nutrients would lead to the deleterious proliferation of planktonic alga and cause lake eutrophication (Chen et al. 2003). The indicators of the organic matter in the water, including COD_{Mn} and COD_{Cr}, were the next two important variables. Together with BOD, these three parameters indicated that organic substances had great influence on the Chl-a concentration in Baiyangdian Lake. Even without toxicity,

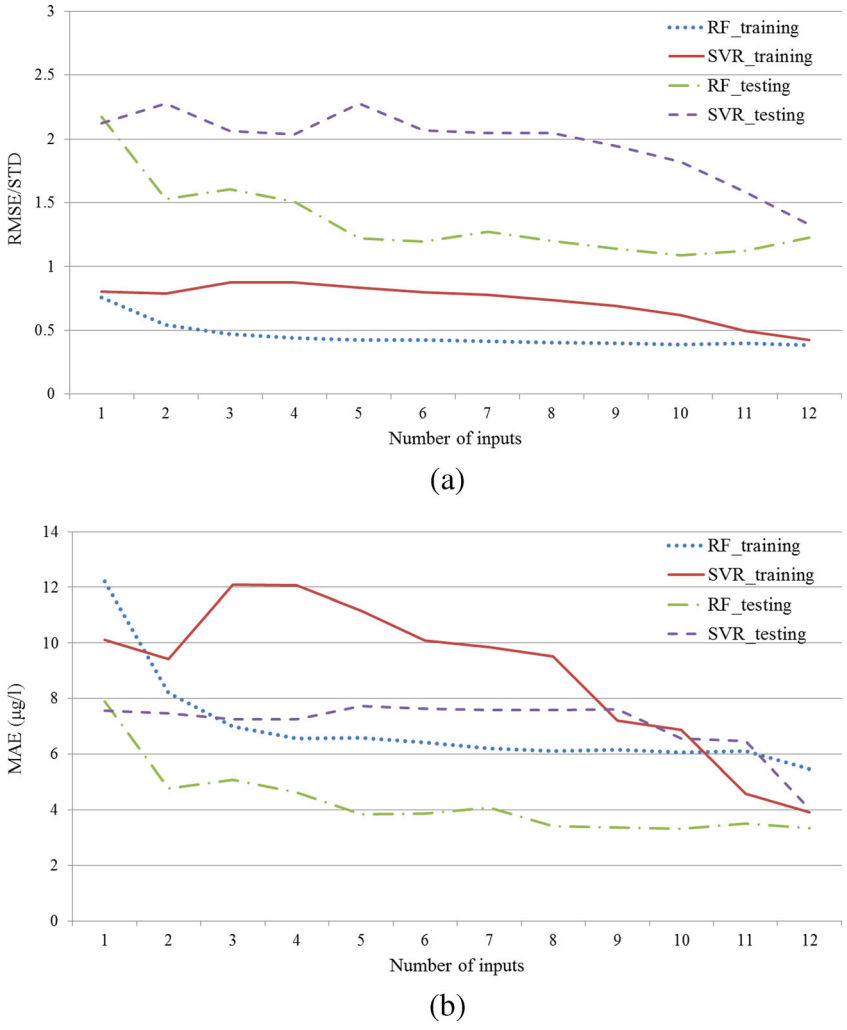

Fig. 3 RMSE/STD and MAE values for mRMR feature sequence during training and testing periods: **a** RMSE/STD and **b** MAE

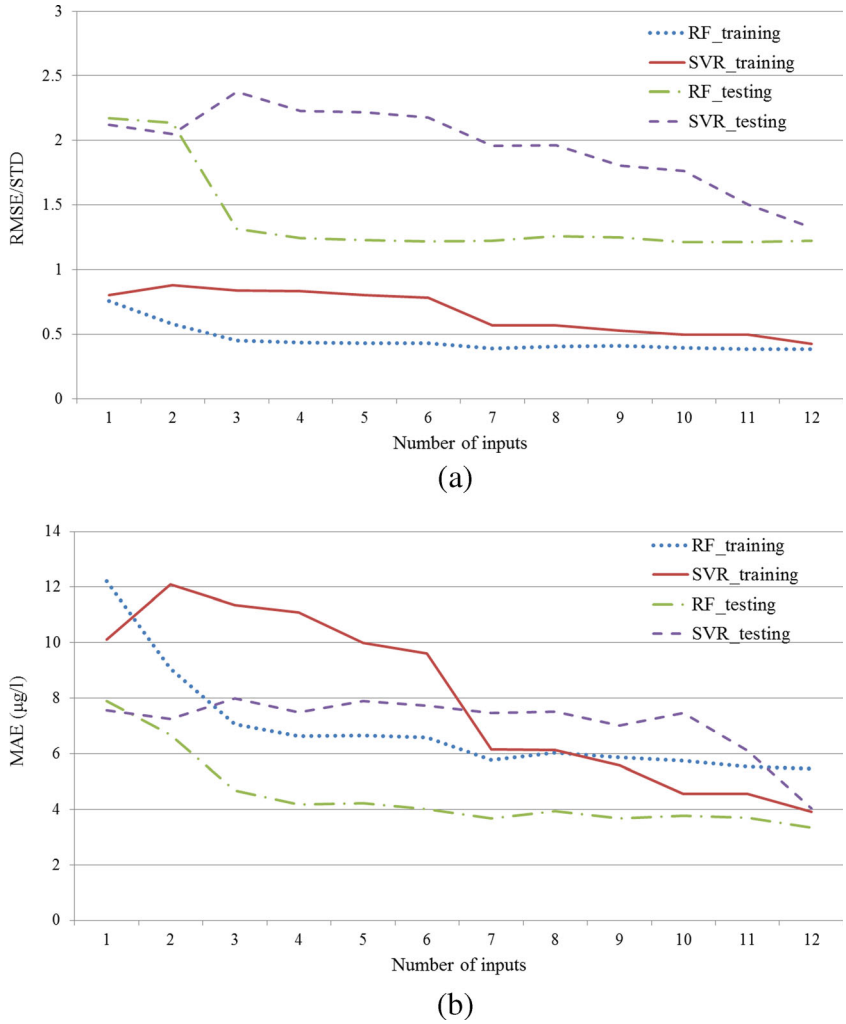

Fig. 4 RMSE/STD and MAE values for RF feature sequence during training and testing periods: **a** RMSE/STD and **b** MAE

organic matters were one of the causes of water pollution because of the consumption of DO in the water (Li et al. 2017c). Adequate DO was essential for the survival of aquatic organisms, and large fluctuations in DO concentrations often had significant influence on the productivity in freshwater ecosystems (Morgan et al. 2006). As mentioned by Convertino et al. (2014)), the importance rank of variables could help the managers and modelers to understand the model structure better and identify the most influential input variables on the model output.

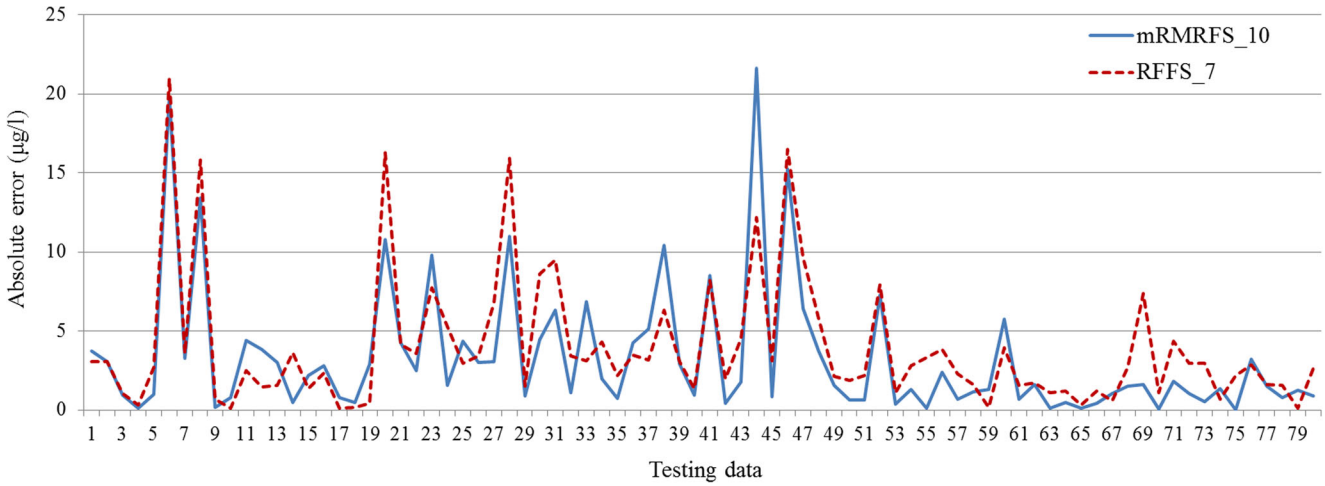

Fig. 5 Absolute errors between observed data and predicted data during testing period for mRMRFS_10 and RFFS_7

Conclusion

As a challenging step in modeling and identifying the most relevant inputs to predict the Chl-a concentration in a shallow lake, the feature selection process could lead to improved model accuracy and low model complexity. In this study, two popular feature selection algorithms, mRMR and RF, were applied to pinpoint the set of inputs that contained the most amount of information. The twelve input variables were ranked by mRMR and RF, respectively. Based on the two feature sequences, the performances of two regression models, SVR and RF, were evaluated and compared. The results of RMSE/STD and MAE showed that the RF models had better prediction accuracies during both training and testing periods. Moreover, the less computational time cost and unnecessary prior data transformation also indicated a superior performance of the RF model. Then, the optimal input subset for Chl-a prediction in this study was selected by using the RF model as a predictor. The forward selection method was applied on both mRMR and RF feature sequences, and the RFFS reached the first result peak with fewer input variables than the mRMRFS. With the similar prediction accuracy, the RFFS with seven variables were selected as the optimal input subset. It was no surprise that the RF performed better than SVR, as it was not strongly dependent upon training parameters and immune to the problems of over fitting. However, it was interesting to find out that the RF had better performance of feature selection than mRMR models.

The RF-RF ensemble method designed for the Chl-a prediction in Baiyangdian Lake could reduce the input variables and improve the accuracy of the model. As a feasible and useful approach for small dataset, it can be tested in other lakes in the future and also be tried in other aspects of the complex water system.

Acknowledgments This work is supported by Tianjin Municipal Education Commission research project (2017KJ125), National Natural Science Foundation of China (No. 41372373), and the innovation team training plan of the Tianjin Education Committee (TD12-5037). Comments and suggestions from two anonymous reviewers and the editor are greatly appreciated.

References

Acharjee A, Kloosterman B, de Vos RCH, Werij JS, Bachem CWB, Visser RGF, Maliepaard C (2011) Data integration and network reconstruction with~ omics data using random Forest regression in potato. Anal Chim Acta 705(1):56–63

Babovic V (2005) Data mining in hydrology. Hydrol Process 19(7): 1511–1515

Babovic V, Keijzer M (2000) Forecasting of river discharges in the presence of chaos and noise. Nato Science Series 2 Environmental Security 71:405–420

Babovic V, Cañizares R, Jensen HR, Klinting A (2001) Neural networks as routine for error updating of numerical models. J Hydraul Eng 127(3):181–193

Bao-Gang H, Yong W (2008) Evaluation criteria based on mutual information for classifications including rejected class. Acta Automat Sin 34(11):1396–1403

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. Geosci Model Dev 7(3):1247–1250

Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27

Chen Y, Fan C, Teubner K, Dokulil M (2003) Changes of nutrients and phytoplankton chlorophyll-a in a large shallow lake, Taihu, China: an 8-year investigation. Hydrobiologia 506(1):273–279

Conrad Lamon E, Qian SS (2008) Regional scale stressor-response models in aquatic ecosystems. JAWRA J Am Water Resour Assoc 44(3):771–781

Convertino M, Muñoz-Carpena R, Chu-Agor ML, Kiker GA, Linkov I (2014) Untangling drivers of species distributions: global sensitivity and uncertainty analyses of MaxEnt. Environ Model Softw 51:296–309

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3): 273–297

Downing JA, McCauley E (1992) The nitrogen: phosphorus relationship in lakes. Limnol Oceanogr 37(5):936–945

Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D, Marshall S (2014) A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol Meas 35(11):2191–2203

Freeman AM, Lamon EC, Stow CA (2009) Nutrient criteria for lakes, ponds, and reservoirs: a Bayesian TREED model approach. Ecol Model 220(5):630–639

Galelli S, Castelletti A (2013) Tree-based iterative input variable selection for hydrological modeling. Water Resour Res 49(7):4295–4310

Gao Y-F, Li BQ, Cai YD, Feng KY, Li ZD, Jiang Y (2013) Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection. Mol BioSyst 9(1):61–69

Genuer R, Poggi J-M, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recogn Lett 31(14):2225–2236

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. J Hydrol 509:379–386

Hejazi MI, Cai X (2009) Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm. Adv Water Resour 32(4):582–593

Huang N, Hu Z, Cai G, Yang D (2016) Short term electrical load forecasting using mutual information based feature selection with generalized minimum-redundancy and maximum-relevance criteria. Entropy 18(9):330

Jaiantilal A (2009) Classification and regression by randomforest-matlab. URL http://code.google.com/p/randomforest-matlab

Kohonen T (1998) The self-organizing map. Neurocomputing 21(1):1–6

Lee H, Lee JH (1995) Continuous monitoring of short term dissolved oxygen and algal dynamics. Water Res 29(12):2789–2796

Lee JH, Huang Y, Dickman M, Jayawardena AW (2003) Neural network modelling of coastal algal blooms. Ecol Model 159(2):179–201

Li F, Miao D, Pedrycz W (2017a) Granular multi-label feature selection based on mutual information. Pattern Recogn 67:410–423

Li X, Sha J, Li Y-M, Wang Z-L (2018) Comparison of hybrid models for daily streamflow prediction in a forested basin. J Hydroinf 20:191–205

Li X, Sha J, Wang Z-L (2017c) Chlorophyll-a prediction of lakes with different water quality patterns in China based on hybrid neural networks. Water 9(7):524

Li X, Sha J, Wang Z-l (2017d) A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. Hydrol Res 48(5):1214–1225

Li X, Zhang Y, Guo F, Gao X, Wang Y (2018) Predicting the effect of land use and climate change on stream macroinvertebrates based on the linkage between structural equation modeling and bayesian network. Ecol Indic 85:820–831

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18–22

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 Resour Manag 28(12):4095–4111

Morgan AM, Royer TV, David MB, Gentry LE (2006) Relationships among nutrients, chlorophyll-, and dissolved oxygen in agricultural streams in Illinois. J Environ Qual 35(4):1110–1117

Mutanga O, Adam E, Cho MA (2012) High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. Int J Appl Earth Obs Geoinf 18:399–406

Paerl HW, Paul VJ (2012) Climate change: links to global expansion of harmful cyanobacteria. Water Res 46(5):1349–1363

Park Y, Cho KH, Park J, Cha SM, Kim JH (2015) Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci Total Environ 502:31–41

Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

Rahmati O, Pourghasemi HR, Melesse AM (2016) Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran region, Iran. Catena 137:360–372

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

Sullivan AB, Snyder DM, Rounds SA (2010) Controls on biochemical oxygen demand in the upper Klamath River, Oregon. Chem Geol 269(1):12–21

Trafalis TB, Ince H (2000) Support vector machine for regression and applications to financial forecasting. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks 6:348–353

Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO (2012) Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng 59(5):1264–1271

Vojinovic Z, Kecman V, Babovic V (2003) Hybrid approach for modeling wet weather response in wastewater systems. J Water Resour Plan Manag 129(6):511–521

Wang X, Zhang J, Babovic V (2016) Improving real-time forecasting of water quality indicators with combination of process-based models and data assimilation technique. Ecol Indic 66:428–439

Were K, Bui DT, Dick ØB, Singh BR (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. Ecol Indic 52:394–403

Yang Y, Yin X, Yang Z (2016) Environmental flow management strategies based on the integration of water quantity and quality, a case study of the Baiyangdian wetland, China. Ecol Eng 96:150–161

Yang Y, Yin X, Yang Z, Sun T, Xu C (2017) Detection of regime shifts in a shallow lake ecosystem based on multi-proxy paleolimnological indicators. Ecological Indicators. https://doi.org/10.1016/j.ecolind.2017.05.059

Yu X, Liong S-Y, Babovic V (2004) EC-SVM approach for real-time hydrologic forecasting. J Hydroinf 6(3):209–223

Yuan R, Wang S, Wang P, Song X, Tang C (2017) Changes in flow and chemistry of groundwater heavily affected by human impacts in the Baiyangdian catchment of the North China Plain. Environ Earth Sci 76(16):571

Zhang Y, Tang L, Zou H, Yang Q, Yu X, Jiang J, Wu H, Yu R (2015) Identifying protein arginine methylation sites using global features of protein sequence coupled with support vector machine optimized by particle swarm optimization algorithm. Chemom Intell Lab Syst 146:102–107