Title

A support vector regression model to predict nitrate-nitrogen isotopic composition using hydro-chemical variables

Authors

Yang, Yue
Shang, Xu
Chen, Zheng
et al.

Publication Date

2021-07-01

DOI

10.1016/j.jenvman.2021.112674

Peer reviewed

Journal of Environmental Management 290 (2021) 112674

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: http://www.elsevier.com/locate/jenvman

Research article

A support vector regression model to predict nitrate-nitrogen isotopic composition using hydro-chemical variables

Yue Yang [a], Xu Shang [b,c], Zheng Chen [b,c], Kun Mei [b], Zhenfeng Wang [b], Randy A. Dahlgren [d], Minghua Zhang [b,c,d], Xiaoliang Ji [b,c,*]

[a] Zhejiang Provincial Key Laboratory for Water Environment and Marine Biological Resources Protection, College of Life and Environmental Science, Wenzhou University, Wenzhou, 325035, China

[b] Key Laboratory of Watershed Science and Health of Zhejiang Province, School of Public Health and Management, Wenzhou Medical University, Wenzhou, 325035, China

[c] Southern Zhejiang Water Research Institute (iWATER), Wenzhou, 325035, China

[d] Department of Land, Air and Water Resources, University of California, Davis, CA, 95616, USA

ARTICLE INFO

Keywords:
Nitrate pollution
Nitrate-nitrogen isotopic composition ($\delta^{15}N–NO_3^-$)
Prediction
Principal component analysis (PCA)
Support vector regression (SVR)
Machine learning model

ABSTRACT

Nitrate is a prominent pollutant in surface and groundwater bodies worldwide. Isotopes in nitrate provide a powerful approach for tracing nitrate sources and transformations in waters. Given that analytical techniques for determining isotopic compositions are generally time-consuming, laborious and expensive, alternative methods are warranted to supplement and enhance existing approaches. Hence, we developed a support vector regression (SVR) model and explored its feasibility to predict nitrogen isotopic composition of nitrate ($\delta^{15}N–NO_3^-$) in a rural-urban river system in Southeastern China. A total of 16 easily obtained hydro-chemical variables were measured in the wet season (September 2019) and dry season (January 2020) and used to develop the SVR prediction model. The grading method utilized ~75% (35) of the samples for model building while the remaining 11 samples assessed model performance. Principal component analysis (PCA) extracted 7 principal components for SVR model inputs as PCA reduces superfluous variables. We optimized tuning parameters in the SVR model using a grid search technique coupled with V-fold cross-validation. The optimized SVR model provided accurate $\delta^{15}N–NO_3^-$ predictions with a determination coefficient (R^2) of 0.88, Nash-Sutcliffe (*NS*) of 0.87, and mean square error (*MSE*) of 0.53‰ in the testing step, and performed much better than the corresponding multivariate linear regression model ($R^2 = 0.60$, $NS = 0.58$ and $MSE = 1.76$‰) and general regression neural network model ($R^2 = 0.66$, $NS = 0.65$ and $MSE = 1.45$‰). Overall, the SVR model provides a potential indirect method to predict environmental isotope values for water quality management that will complement and enhance the interpretation of direct measurements of $\delta^{15}N–NO_3^-$.

1. Introduction

Nitrate (NO_3^-) contamination in surface waters induces deterioration of aquatic ecosystem health (e.g., eutrophication, harmful algal blooms, loss of aquatic biodiversity, and hypoxia/anoxia) as well as human health risks (e.g., stomach cancer, diabetes, thyroid disorders, miscarriage, and "blue baby" syndrome) (Burow et al., 2010; Ji et al., 2017b; Nestler et al., 2011; World Health Organization, 2011). Nitrate levels in surface waters are regulated by the interplay between allochthonous pollution sources, such as municipal sewage, livestock excreta, nitrogen fertilizer, soil nitrogen and atmospheric deposition, and nitrogen cycling processes including ammonia volatilization, nitrification, denitrification and plant/microbial uptake (Husic et al., 2020; Shang et al., 2020; Wang et al., 2020). Identification of the major pollution sources and transformations of nitrate in surface waters is a primary objective for environmental and water quality agencies to develop remediation strategies to address nitrate contamination.

Measuring the isotopic composition in nitrate ($\delta^{15}N/\delta^{18}O–NO_3$) provides important information on nitrate sources and potential transformations (Fadhullah et al., 2020; Hu et al., 2019; Kendall and

* Corresponding author. Key Laboratory of Watershed Science and Health of Zhejiang Province (China), Wenzhou Medical University, Wenzhou, 325035, Zhejiang Province, China.
E-mail address: jixiao556677@wmu.edu.cn (X. Ji).

https://doi.org/10.1016/j.jenvman.2021.112674
Received 8 November 2020; Received in revised form 14 April 2021; Accepted 16 April 2021
Available online 23 April 2021

McDonnell, 1998; Xue et al., 2009). Merits of the isotope-based approach include less ancillary information requirements, high precision, and direct identification of pollution sources. Analyzing the isotopic composition of nitrate is an initial step in pollution source apportionment as it serves as an isotope-based source tracing method. An isotope ratio mass spectrometer (IRMS) is a necessary instrument for measuring stable isotopic compositions. However, IRMS is very expensive to purchase and maintain, limiting its availability for many institutions. Additionally, sample pre-treatment techniques, including ion-exchange, cadmium-azide reduction and bacteria denitrifier, are highly technical, laborious and time-consuming further increasing analysis cost. Given these issues, stable isotope values are often difficult to acquire in sufficient quantities for complex field studies, especially in developing countries. Therefore, an economical and efficient secondary (indirect) method is warranted to predict environmental isotope values and supplement direct measurements of the stable isotopes in nitrate.

Machine learning models are an emerging data-analysis/discovery method for addressing water resource applications, which employ artificial neural networks, adaptive neuro-fuzzy inference systems, extreme learning machines and support vector regression (SVR) models. Several studies document the potential of machine learning models for simulating and predicting streamflow, rainfall, groundwater levels, suspended sediment loads, evapotranspiration, dissolved oxygen content, biochemical oxygen demand, algal density and the isotopic composition of oxygen in water ($\delta^{18}O–H_2O$) (e.g., Adnan et al., 2020; Banadkooki et al., 2020; Cerar et al., 2018; Diez-Sierra and del Jesus, 2020; Ji et al., 2017a; Kim et al., 2020; Noori et al., 2015; Tikhamarine et al., 2020; Xiao et al., 2017; Yaseen et al., 2016; Yoon et al., 2011). For example, Cerar et al. (2018) investigated the performance of artificial neural networks for $\delta^{18}O–H_2O$ prediction in groundwater based on spatial characteristics of the watershed including elevation, distance from the sea and average annual precipitation. Their results demonstrated the efficacy of artificial neural network models to effectively predict $\delta^{18}O–H_2O$ values of groundwater. Despite the significant potential of machine learning models in the field of environmental isotope modeling, relevant studies assessing the application of machine learning models for $\delta^{15}N–NO_3^-$ value prediction are lacking. Hence, there is a compelling opportunity to test whether machine learning approaches can be effectively used to model/predict $\delta^{15}N–NO_3^-$ values using common and easily measured hydro-chemical variables as the input data.

Among different machine learning models, SVR model has gained popularity due to its superior generalization and accurate prediction abilities. The main advantage of SVR is its use of a kernel trick to minimize prediction errors and model complexity simultaneously when coping with complex nonlinear associations (Raghavendra and Deka, 2014). Furthermore, like other machine learning models, the SVR model is an input-output transformation, which can be structured without understanding the underlying mechanistic processes, leading to simple and practical model development. Additional merits of SVR models are (1) excellent performance using relatively small data sets, and (2) prevention of over-fitting the model, which is a critical shortcoming of artificial neural networks (Ji and Lu, 2018).

In view of the above considerations, the main objective of this study was to develop a machine learning model (i.e., SVR model) for predicting $\delta^{15}N–NO_3^-$ values based on common and easily measured hydro-chemical variables. We believe this study is the first attempt to examine the potential feasibility of machine learning models for prediction of $\delta^{15}N–NO_3^-$ values in surface waters. We expect that successful SVR models will serve as an efficient tool for accurate, rapid and inexpensive prediction of environmental isotopes that can supplement and enhance the interpretation of directly measured $\delta^{15}N–NO_3^-$ values in water quality studies.

2. Study area and data collection

2.1. Study area

The Wen-Rui Tang River watershed (27°51' – 28°02′ N, 120°27′–120°46′ E) is located in Wenzhou, Zhejiang province of Southeastern China and occupies a total drainage area of 740 km^2 (Fig. 1). The Wen-Rui Tang River is a major river system across a rural-suburban-urban interface in Wenzhou, where is a rapidly developing city with a population of ~9.2 million. The mainstream and total lengths of the Wen-Rui Tang River network are 33.8 km and ~1200 km, respectively. Climate is subtropical monsoon with mild, dry winters and hot, humid summers. Annual average temperature is ~18 °C and annual average rainfall is ~1800 mm, with 70% of precipitation occurring between April and September (Wang et al., 2019). The Wen-Rui Tang River is a typical coastal plain river network in Southeast China characterized by nearly stagnant water flows for long periods of the year (Wang et al., 2018). When floodgates to the adjacent Ou River are open during heavy rainfall events, the river ultimately flows into the East China Sea.

2.2. Data collection

Two synoptic water quality surveys collected water samples from 23 sites across the Wen-Rui Tang River watershed under contrasting hydrological conditions. The first sampling campaign was performed during the wet season (September 2019) and the second in the dry season (January 2020). In total, 46 samples were collected from a 30-cm depth in the center of a well-mixed channel segment.

Several hydro-chemical variables (i.e., water temperature (T), dissolved oxygen (DO), electrical conductivity (EC), turbidity (TUR), phycocyanin of cyanobacteria (PC), chlorophyll *a* (Chla)) were determined *in situ* using a portable multi-parameter water-quality sonde (YSI-EXO2, Xylem, USA). Nutrient (total nitrogen (TN), ammonium (NH_4^+), nitrate (NO_3^-), nitrite (NO_2^-), total phosphorus (TP), phosphate (PO_4^{3-})) concentrations were measured with a continuous-flow analyzer (Auto-analyser-3, Seal, Germany); chloride (Cl^-) was analyzed using ion chromatography (Compact IC plus 882, Metrohm, Switzerland); and total organic carbon (TOC), total carbon (TC) and total inorganic carbon (TIC) were determined with a TOC analyzer (TOC-L, Shimadzu, Japan). Detection limits for TN/TP, $NH_4^+/NO_3^-/NO_2^-/PO_4^{3-}$, Cl^-, and TOC/TC/IC were ~0.02 mg N or P/L, ~0.003 mg N or P/L, ~0.1 mg Cl/L and ~0.1 mg C/L, respectively.

The $\delta^{15}N–NO_3^-$ values were analyzed using the bacteria denitrifier method (Sigman et al., 2001; Casciotti et al., 2002) at the Chinese Academy of Agricultural Sciences (Beijing, China). Briefly, a denitrifying bacteria *Pseudomonas aureofaciens* (ATCC 13985, United States) transformed nitrate to gaseous nitrous oxide (N_2O) for detection of $^{15}N–N_2O$ using a continuous-flow isotope ratio mass spectrometer (Delta V, Thermo Fisher Scientific). The $\delta^{15}N–NO_3^-$ values are expressed in parts per thousand (‰) relative to atmospheric N_2 (Kendall et al., 2007) and have an analytical precision of ±0.2‰.

3. Model development

3.1. Data pre-processing

To calibrate and evaluate model efficacy, water samples were divided into training and testing data sets. The training data set was used for model construction, whereas the validation of model performance utilized the testing data set. The training samples should provide representative information about the sources of variability that are present in unknown samples across the watershed. Herein, we used a grading method to select representative training and testing samples as follows: (1) all samples were ranked based on $\delta^{15}N–NO_3^-$ values; (2) for every four samples, the first three samples were extracted into the training data set and the last sample was assigned to the testing data set;

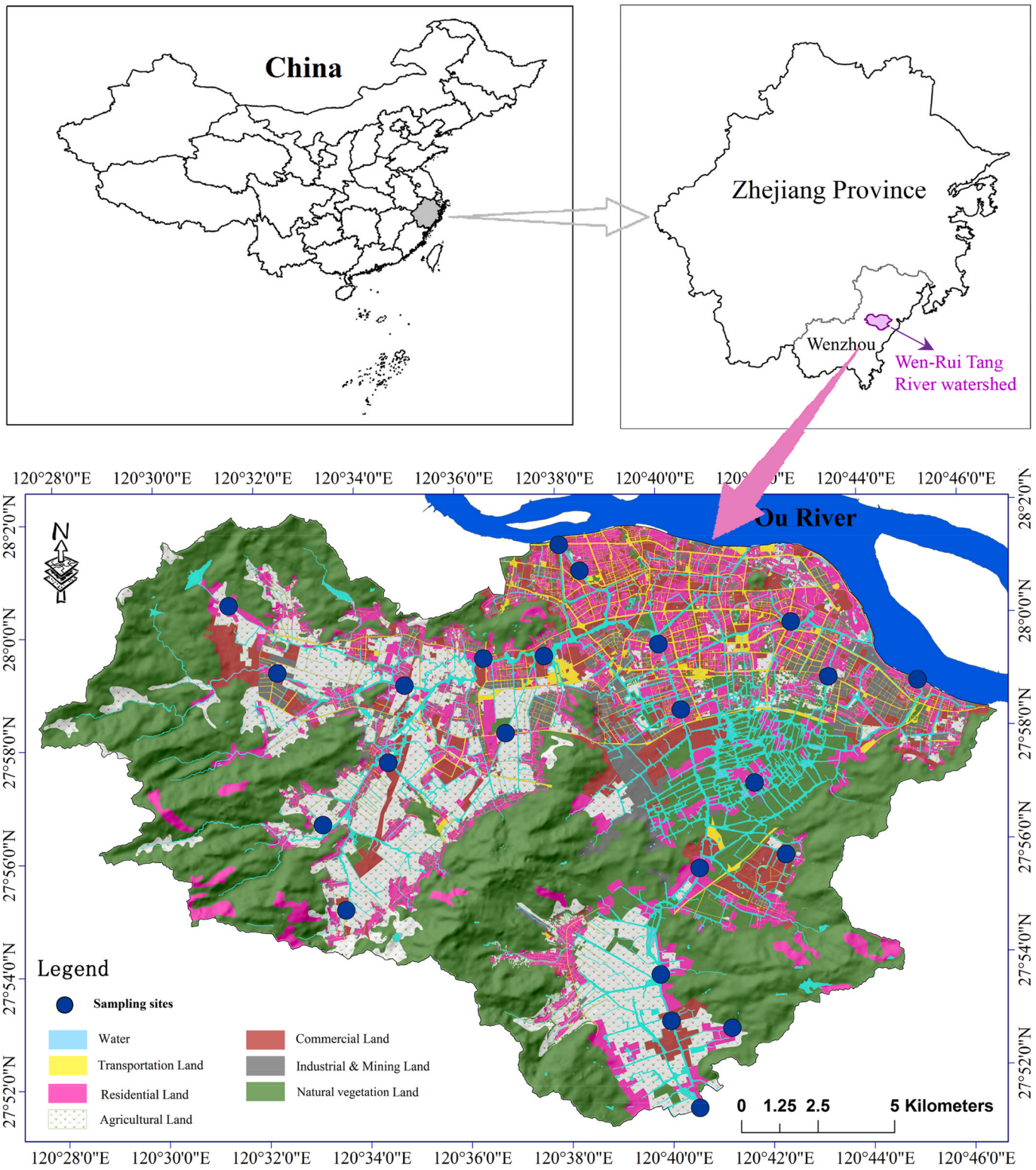

Fig. 1. Location of sampling sites in the Wen-Rui Tang River watershed.

and (3) the training and testing data sets used 75% and 25% of available samples, respectively.

Data normalization was performed to eliminate dimensional differences and transform raw data into a common scale, without distorting differences in the ranges of values or losing information. In this study, we employed z-score normalization to transform on all variables independently (Chen et al., 2020). The mathematical formula for z-score normalization is:

$$x_n = \frac{x - x_{mean}}{x_{SD}} \quad (1)$$

where, x_n is the normalized value; x is the original value; and x_{mean} and x_{SD} are the mean and standard deviation of the original variables.

3.2. Principal component analysis

The 16 measured hydro-chemical parameters were taken as potential input variables given their direct or indirect effect on ^{15}N values in aquatic ecosystems. When using a high number of input variables, the occurrence of irrelevant, redundant and noisy variables coupled with multicollinearity among variables may weaken model performance. Therefore, it is necessary to first extract key model input variables.

Principal component analysis (PCA) is a multivariate statistical technique extensively applied for data reduction in environmental and hydro-chemical studies. It gives information on the most meaningful variables, which describe the interpretation of the whole data set, summarize the statistical correlation among different variables with little loss of the original information (Helena et al., 2000). PCA is an objective method for calculating indices so that the variations in the data set are accounted by a new set of uncorrelated variables called principal components (Li et al., 2018; Sarbu and Pop, 2005). Principal components are originated from linear combinations of the original variables. Typically, the first principal component can explain most of the variations while the last principal component is responsible for the least of variations in all original variables.

3.3. Support vector regression model

Support vector machine (SVM) models, developed on the basis of statistical learning theory, were widely used for classification and regression problems (Mohammadpour et al., 2015). SVM models include support vector classification (SVC) models and SVR models. The basic theory of SVR model is to map the original data points from the input space into a higher or even infinite-dimensional feature of space where an optimal separating hyperplane is built (Lin et al., 2008). The distance to all data points is minimum from the constructed separating hyperplane. Numerous studies report on the full development and expression of SVR models (Raghavendra and Deka, 2014; Vapnik, 1998). Therefore, only a brief explanation of the SVR is provided below along with a schematic diagram in SVR (Fig. S1).

For a training data set $\{(x_i, y_i)|i = 1, 2, \ldots, n\}$, where $x_i \in R^D$ is a D-dimensional real input vector, $y_i \in R$ is the corresponding target value, and n is the total number of data patterns, the regression function of SVR model is expressed as follows:

$$f(x) = w^T \cdot \varphi(x) + b \tag{2}$$

where $w \in R^D$ is a weight vector, T stands for the transpose operator. b is a bias, φ is a nonlinear transfer function mapping the input vectors into a high dimensional feature space. The parameters w and b, which define the location of the separating hyperplane, can be derived by minimizing the regularized risk function as follows:

$$Minimize: \frac{1}{2} w^T \cdot w + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right) \tag{3}$$

Subject to

$$y_i - w^T \cdot \varphi(x_i) - b \le \varepsilon + \xi_i$$
$$w^T \cdot \varphi(x_i) + b - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i \ge 0, \xi_i^* \ge 0, i = 1, 2 \ldots, n$$

where C stands for the regularization parameter, ε (Epsilon) is error tolerance, and ξ_i and ξ_i^* are slack variables. Minimizing Eq. (3) is a constrained optimization problem. Introducing a dual set of Lagrange multiplier i.e., α_i and α_i^* allows the optimization problem to be solved by maximizing the quadratic programming algorithm as follows:

$$\sum_{i=1}^{n} y_i \left(\alpha_i - \alpha_i^*\right) - \varepsilon \sum_{i=1}^{n} \left(\alpha_i + \alpha_i^*\right) - \frac{1}{2} \sum_{i,j=1}^{n} \left(\alpha_i - \alpha_i^*\right) \left(\alpha_j - \alpha_j^*\right) \varphi(x_i)^T \cdot \varphi(x_j) \tag{4}$$

Subject to

$$\sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) = 0$$
$$0 \le \alpha_i \le C$$
$$0 \le \alpha_i^* \le C, i = 1, 2, \ldots, n$$

The solution to Eq. (4) is unique and optimal. Subsequent to the determination of Lagrange multipliers in Eq. (4), the parameters w and b in support vector machine regression function can be calculated under the Karush-Kuhn-Tucker optimality condition, where $w = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \varphi(x_i)$. Besides, the inner product $\varphi(x_i)^T \cdot \varphi(x)$ can be replaced by the so-called kernel function $K(x_i, x)$ under Mercer's condition. Therefore, the final form of SVR function can be expressed as follows:

$$f(x) = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) K(x_i, x) + b \tag{5}$$

It follows from this description that the kernel function plays a crucial role in the SVR algorithm. There are four options, namely polynomial, sigmoid, linear and radial basis function (RBF), that can be utilized as the SVR model kernel function. In this study, the RBF kernel function was chosen for the following reasons: (1) unlike the linear kernel, RBF is capable of modeling nonlinear relationships by mapping data points from the input space into a high dimensional feature space in a nonlinear fashion; (2) RBF requires fewer adjustable parameters compared to polynomial and sigmoid kernels, making it simple and practical (Keerthi and Lin, 2003); and (3) the superior performance of RBF has been demonstrated in several studies (Dibike et al., 2001; Keerthi and Lin, 2003). The kernel function RBF is defined as:

$$K(x_i, x) = \exp\left(-g \|x_i - x\|^2\right) \tag{6}$$

where g is the adjustable kernel parameter.

3.4. Support vector regression model parameter optimization

It is important to note that the performance of SVR models is strongly dependent on the RBF kernel parameter g in combination with the regularization parameter C. The parameter g defines the width of the kernel, which regulates the amplitude of the kernel function and, thereby, the generalization ability of the model. The parameter C regulates the trade-off between maximizing the margin and minimizing training errors. A minimal C value induces insufficient fitting of the training data, while a large C value gives rise to the algorithm to over-fit the training data (Wang et al., 2007).

This study utilized a grid search technique coupled with V-fold cross-validation to calculate the optimal C and g values (Hsu et al., 2007). For V-fold cross-validation, data were randomly partitioned into five non-overlapping subsets, each containing one-fifth of the data. Each subset served as the testing data for models trained on the other four-fifths of the data, resulting in five different pairs of training and testing data sets, with each observation appearing in one testing set and the four training sets not paired with that testing set. Therefore, the V-fold cross-validation procedure limits the over-training problem. The grid search algorithm is an unguided method and the mean square error (*MSE*) value is often used as a criterion to tune the parameters C and g. Grid search algorithm divides the search scope of the parameters to be optimized into grids and traverses all the grid points to search the optimal value.

3.5. Model performance assessment

Performance of the optimized $\delta^{15}N$–NO_3^- prediction model was evaluated using three commonly used performance metrics: determination coefficient (R^2), Nash-Sutcliffe model efficiency (*NS*) and mean square error (*MSE*). These indexes were computed as:

$$R^2 = \left(\frac{\sum_{i=1}^{n} \left(P_i - \overline{P}\right)\left(O_i - \overline{O}\right)}{\sqrt{\sum_{i=1}^{n} \left(P_i - \overline{P}\right)^2 \sum_{i=1}^{n} \left(O_i - \overline{O}\right)^2}} \right)^2 \quad (7)$$

$$NS = 1 - \frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} \left(O_i - \overline{O}\right)^2} \quad (8)$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (O_i - P_i)^2 \quad (9)$$

where n is the total number of data; O_i and P_i denote the measured and predicted $\delta^{15}N–NO_3^-$ values (‰), respectively; and $\overline{O}$ and $\overline{P}$ represent the mean of measured and predicted $\delta^{15}N–NO_3^-$ values (‰), respectively.

The R^2 represents the square of the correlation between the predicted and observed values (Kim et al., 2020). The *NS* is a measure of the ability of the model to predict observations (Shoaib et al., 2016). According to Shu and Ouarda (2008), model accuracy can be evaluated as very good when $NS > 0.8$. The *MSE* measures the average error associated with the model (Legates and McCabe, 1999). In sum, $R^2 = 1$, $NS = 1$ and $MSE = 0$‰ indicate perfect model performance.

3.6. Software

The correlation test figure was generated using the corrplot package in R (Ver. 3.0.2, R Core Team; https://CRAN.R-project.org/package=corrplot). Multivariate linear regression (MLR) model and z-score normalization were performed using SPSS (Ver. 17.0, SPSS Inc., Chicago, USA). PCA and general regression neural network (GRNN) modeling were performed using MATLAB (Ver. 2018b; MathWorks, Natick, USA). SVR modeling was conducted using the LIBSVM toolbox (Chang and Lin, 2011) working in MATLAB environment.

4. Application and results

4.1. Training and testing data set partitioning

The grading method selected 35 samples to establish the model while the remaining 11 samples were used as testing data to assess model performance. Basic statistics for the target variable ($\delta^{15}N–NO_3^-$) and hydro-chemical variables in training and testing data sets are provided in Table 1. The range (−2.38 – 13.37‰) and coefficient of variation (CV = 46.4%) for $\delta^{15}N–NO_3^-$ in the training data set was broad, and encompassed that of the testing data set (range = 2.22–9.97‰, CV = 32.6%).

4.2. Input variable reduction

This study used 16 hydro-chemical variables to predict nitrate-nitrogen isotopic composition in surface waters of the Wen-Rui Tang River network. Spearman rank correlation analysis among the different hydro-chemical variables examined the linear dependence between variables (Fig. 2). There were strong correlations between several water quality parameters. For instance, correlation coefficients (r values) between TP and PO_4^{3-}, Cl^- and EC, and PC and Chla were 0.90, 0.95 and 0.90, respectively; while correlation coefficients among carbonaceous components (e.g., TOC, TC, IC) were greater than 0.85. These results revealed that several hydro-chemical variables contained collinear and redundant information, which often leads to performance degradation of the calibrated models.

To eliminate superfluous inputs, we performed PCA on the original variables (16 hydro-chemical variables) to extract the useful and independent principal components. We performed a Kaiser-Meyer-Olkin analysis to determine the suitability of the data for PCA analysis. Kaiser-Meyer-Olkin analysis is a measure of sampling adequacy that indicates the proportion of common variance, i.e., variance caused by the underlying factors. High values (≥0.6) generally suggest that PCA may be useful, which was the case in this study: Kaiser-Meyer-Olkin = 0.60. Further, Bartlett's sphericity test confirmed that the data distribution was suitable for PCA owing to a *p*-value close to zero (≤0.01). Figure S2 depicts the changing trend for cumulative percent variance with an increasing number of principal components. The percent of cumulatively explained total variance values increased quickly as the number of principal components increased from 1 to 7, which was ascribed to the inclusion of useful variables, then held a relatively stable level as additional principal components were added. Herein, the scores of first seven principal components (termed PC_1, PC_2, …, PC_7) which explained > 95% of the total variance were selected as the input variables to construct the SVR model.

Table 1
Statistical summary of $\delta^{15}N–NO_3^-$ and hydro-chemical variables in the training and testing data sets.

	Mean	SD	Minimum	Maximum	CV (%)
Training data set (n = 35)					
$\delta^{15}N–NO_3^-$ (‰)	6.44	2.99	−2.38	13.37	46.4
T (°C)	19.3	5.9	12.5	27.7	30.7
DO (mg/L)	6.5	3.0	1.7	12.2	46.0
TN (mg/L)	3.11	1.60	0.34	7.37	51.6
NH_4^+ (mg/L)	1.76	1.67	0.07	6.50	95.1
NO_3^- (mg/L)	1.25	0.80	<0.01	2.75	64.0
NO_2^- (mg/L)	0.08	0.05	<0.01	0.19	69.2
TP (mg/L)	0.12	0.09	0.03	0.47	74.3
PO_4^{3-} (mg/L)	0.10	0.10	<0.01	0.44	98.9
TOC (mg/L)	5.2	2.5	2.2	12.4	47.1
TC (mg/L)	20.1	10.9	5.9	59.7	53.9
TIC (mg/L)	14.9	8.6	3.1	47.4	57.7
Cl^- (mg/L)	102.3	91.7	5.1	283.3	89.6
EC (ms/cm)	0.54	0.41	0.03	1.50	76.0
TUR (NTU)	25.3	24.1	1.6	85.6	95.0
PC (μg/L)	0.5	0.6	0.1	2.5	114.0
Chla (μg/L)	11.0	12.0	0.3	50.9	109.2
Testing data set (n = 11)					
$\delta^{15}N–NO_3^-$ (‰)	6.59	2.15	2.22	9.97	32.6
T (°C)	22.4	5.3	13.8	26.6	23.5
DO (mg/L)	6.9	1.9	3.6	9.8	28.1
TN (mg/L)	1.82	1.43	0.35	4.50	78.5
NH_4^+ (mg/L)	1.10	1.46	0.06	4.97	132.4
NO_3^- (mg/L)	0.89	0.67	<0.01	2.20	75.1
NO_2^- (mg/L)	0.07	0.05	0.01	0.15	73.5
TP (mg/L)	0.09	0.10	0.03	0.39	109.2
PO_4^{3-} (mg/L)	0.08	0.10	0.01	0.37	125.5
TOC (mg/L)	4.7	2.4	2.0	9.8	50.8
TC (mg/L)	17.7	9.4	5.9	34.8	53.0
TIC (mg/L)	13.0	7.2	3.4	25.1	55.0
Cl^- (mg/L)	72.8	82.9	3.4	252.1	113.8
EC (ms/cm)	0.35	0.31	0.04	0.95	88.3
TUR (NTU)	23.2	17.9	2.5	52.3	77.5
PC (μg/L)	0.6	1.1	0.2	4.0	174.4
Chla (μg/L)	13.3	24.2	0.4	85.1	182.2

SD standard deviation, *CV* coefficient of variation.

4.3. Optimization of SVR tuning parameters

Optimum tuning parameters are essential for proper SVR calibration. Thus, prior to developing the SVR model, we optimized the regularization parameter C and RBF kernel parameter g using a grid search technique coupled with V-fold cross-validation. The accuracy of grid search optimization depends on the parameter range in combination with the selected interval size (Singh et al., 2011). Commonly, higher efficiencies for obtaining an optimal solution are achieved by increasing the parameter range and decreasing the step size (Wang et al., 2007). Therefore, both the C and g parameters were evaluated widely within

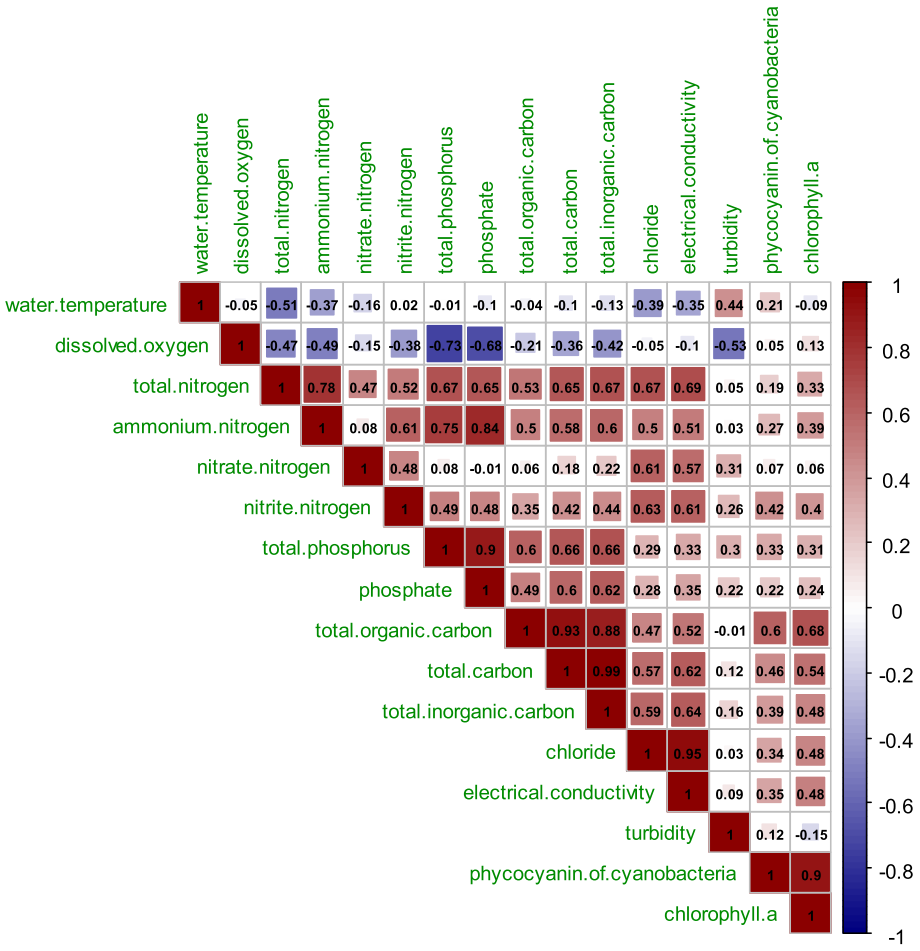

Fig. 2. Spearman rank correlation coefficients for hydro-chemical variables.

the region of 2^{-10} to 2^{10}, and the step size was set to $2^{0.5}$. The pairwise parameters with the lowest *MSE* for cross-validation values were considered as the optimum combination. Fig. 3 shows a three-dimensional view of optimization results for *C* and *g* by the grid search method with V-fold cross-validation. The optimal values were $C = 5.6569$ and $g = 0.0625$.

4.4. Performance of the SVR model

We used the optimum combination for the *C* and *g* parameters to train the SVR model for prediction of $\delta^{15}N\text{–}NO_3^-$ values (Fig. 4). The optimized SVR model had outstanding efficacy for predicting $\delta^{15}N\text{–}NO_3^-$ values with regard to the R^2 (0.88), *NS* (0.87) and *MSE* (0.53‰) metrics. Observed and predicted values were fully superposed and the differences between predicted and observed values were small, suggesting

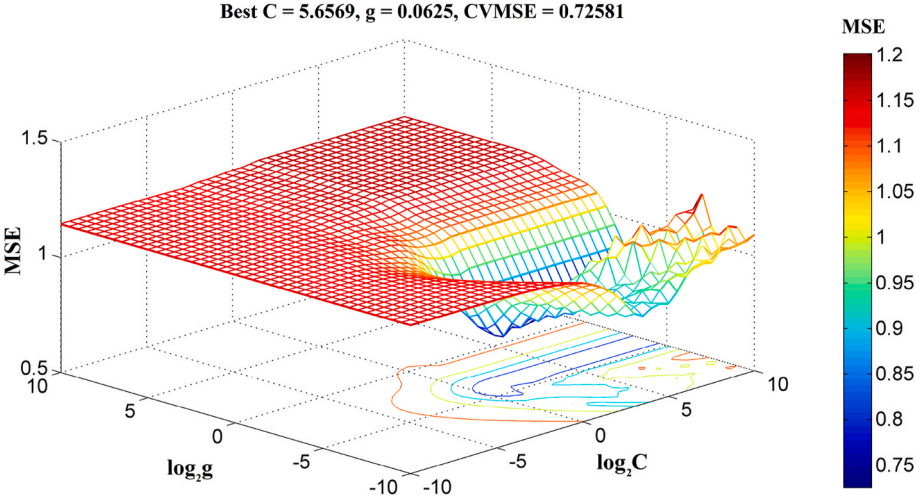

Fig. 3. Three-dimensional representation of support vector regression (SVR) model parameter optimization.

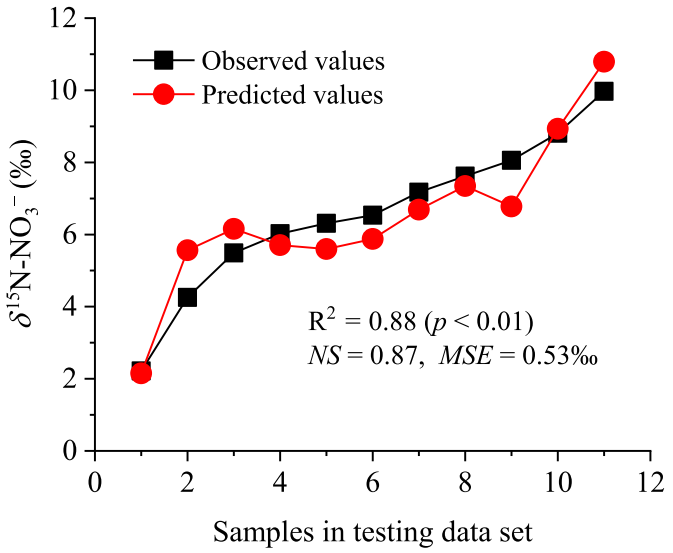

Fig. 4. Observed versus predicted $\delta^{15}N–NO_3^-$ values using SVR model in testing period.

strong efficacy for the SVR model prediction of $\delta^{15}N–NO_3^-$ values in surface waters of the Wen-Rui Tang River network.

4.5. Comparison of SVR, MLR and GRNN

The MLR model, which assumes the dependent variables are linearly dependent on two or more independent variables (Abrougui et al., 2019), often serves as a reference to assess other nonlinear models (like the SVR model used in this study). The MLR model (regression kind: Enter) was constructed as in Eq. (10) to predict $\delta^{15}N–NO_3^-$ values in the Wen-Rui Tang River network.

$$\delta^{15}N-NO_3^- = 0.009 - 0.136^{*}PC_1 + 0.324^{**}PC_2 - 0.001PC_3 - 0.074PC_4 + 0.226PC_5 + 0.204PC_6 + 0.249PC_7 \quad (10)$$

where * refers to $p < 0.05$, ** refers to $p < 0.01$.

Fig. 5a presents the observed and predicted $\delta^{15}N–NO_3^-$ values obtained from the MLR model during the testing stage. The performance of the established MLR model was unsatisfactory for $\delta^{15}N–NO_3^-$ prediction purposes in terms of R^2 (0.60), *NS* (0.58) and *MSE* (1.76‰). Observed and predicted values were not well superposed and the differences between predicted and observed values were often large. These results revealed the MLR model failed to effectively predict the isotope values for the riverine network, largely due to its linear structure. The SVR model was far superior to MLR model in predicting $\delta^{15}N–NO_3^-$ values in terms of R^2 (0.60 vs. 0.88), *NS* (0.58 vs. 0.87) and *MSE* (1.76‰ vs. 0.53‰). GRNN is a common artificial neural network which has wide applications in various linear and nonlinear regression problems. Here, the GRNN was trained ten times and the best network was retained for isotopic prediction. The smoothing factor was evaluated within the range of 0.1–2 at the step size of 0.1 and which was determined as 1.5. Finally, the GRNN model was obtained with R^2, *NS* and *MSE* in testing data set of 0.66, 0.65 and 1.45‰, respectively (Fig. 5b). Obviously, GRNN could provide better predictions than the MLR model. However, the performance of GRNN was much poorer than that of SVR, which was not available for prediction purpose. Overall the results demonstrated the superiority of SVR model compared to the MLR and GRNN models in the prediction of $\delta^{15}N–NO_3^-$ values.

5. Discussion

Machine learning models do not require any complex or explicit description of the underlying hydrologic/environmental processes in a mathematical form (Ahmed et al., 2019). Thus, they are useful for dealing with prediction problems in many scientific disciplines where the main concern is accurate predictions and not necessarily understanding the underlying mechanistic relationships. Although extensively employed to predict water quality parameters including DO (Antanasijević et al., 2013), chemical oxygen demand (Kisi and Parmar, 2016), biochemical oxygen demand (Kim et al., 2020), and chlorophyll

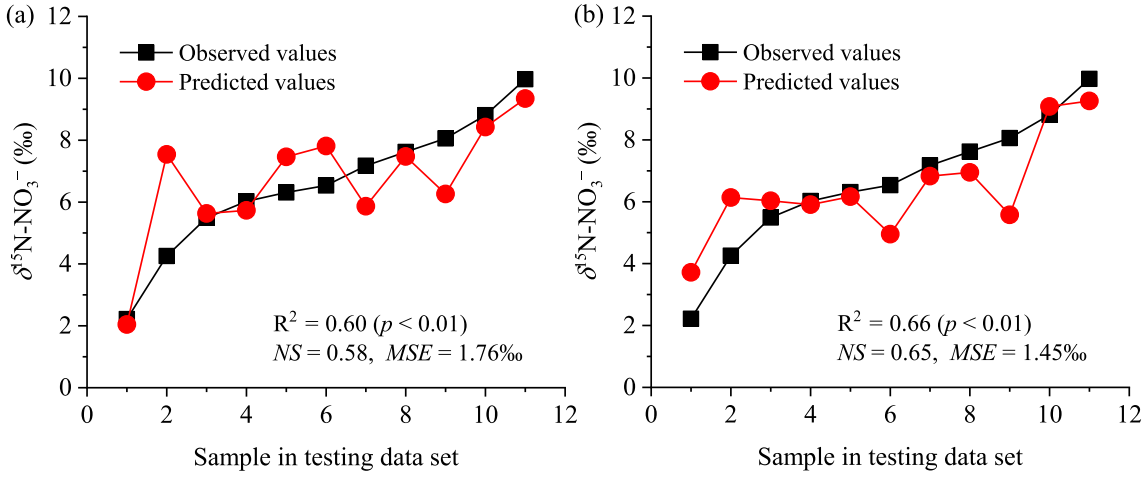

Fig. 5. Observed versus predicted $\delta^{15}N–NO_3^-$ values using (a) multivariate linear regression (MLR) model and (b) general regression neutral network (GRNN) model in testing period.

(Mamun et al., 2020), machine learning models are not yet used to predict environmental stable isotopes based on easily measured and low-cost hydro-chemical variables. Hence, this study fills an important gap in advancing machine learning techniques to studies of environmental stable isotopes.

Analyzing the composition of environmental isotopes using complex sample pre-treatment techniques (e.g., bacteria denitrifier) coupled with IRMS in the laboratory is not only expensive, but also time and labor consuming, which hinders its widespread application. Although the machine learning models do not replace experimental analysis, this study demonstrates that nitrate isotopes can be accurately predicted by more rapid, cost-effective, and easier to measure hydro-chemical variables *via* the alternative machine learning model approach. Thus, after appropriate data acquisition for calibration and validation, machine learning models can supplement and enhance the interpretation of measured isotope data.

The aquatic ecosystems are complex biogeochemical systems containing numerous chemical, physical and biological components experiencing a wide range of integrated transformation processes. Creating an association between water quality parameters and stable isotopes in water bodies is thereby a complex nonlinear problem, which is beyond the capability of linear models, such as MLR models. Thus, the SVR model outperformed the MLR model for $\delta^{15}N–NO_3^-$ prediction because of its ability to model complicated nonlinear relationships. Moreover, one of the most important characteristics of SVR is its ability to generalize well from a limited amount of training samples such as 35 samples used here. Compared to alternative methods such as artificial neural networks (like the GRNN model used here), SVR can yield comparable accuracy using a much smaller training sample size (Mountrakis et al., 2011). This is in line with the "support vector" concept that relies only on a few data points to define the position of the decision surface (Huang and Zhao, 2018; Kuter, 2021; Mountrakis et al., 2011). In light of this research, the SVR model had the ability to accurately modeling $\delta^{15}N–NO_3^-$ values in surface water relying on commonly measured hydro-chemical variables, and thereby serves as an indirect, rapid, and convenient tool for environmental stable isotope prediction.

We recommend researchers and water resource managers investigate the relationship between stable isotopic composition and water quality parameters using the SVR modeling approach established here, especially in regions where isotope compositions are difficult to obtain. As more and more waterbodies have high-frequency water quality monitoring programs in real time, many model inputs are easily obtained for investigation as input data. The SVR model based on readily available input variables enables us to predict $\delta^{15}N–NO_3^-$ values with high spatial and temporal resolution. This is of considerable significance and can aid full interpretation of nitrate source dynamics and transformations. Of course, sufficient $\delta^{15}N–NO_3^-$ data must be initially collected to develop and confirm model results as various systems are likely to have site-specific attributes. SVR models can be used to generate hypotheses concerning spatial/temporal patterns in isotopes that can be subsequently tested with direct isotope measurements. Models can further aid in optimizing the collection of water sample locations within watersheds for isotope analysis to maximize the effectiveness of the data acquisition strategy. Furthermore, this study demonstrates the potential efficacy for prediction of other environmental stable isotopes (e.g., $\delta^{18}O–H_2O$, $\delta^{13}C$, and $\delta^{11}B$) in waterbodies such as rivers, lakes, reservoirs, groundwater and oceans.

While excellent model performance was achieved in this study, some additional aspects need to be further investigated in future research. For instance, only 16 hydro-chemical variables were used as inputs in this study. Therefore, it is warranted to investigate other potentially important variables, such as hydrology (e.g., flow velocity and hydraulic retention time), microbial communities (e.g., species richness, total number of species, species evenness, and distribution of species), meteorology (e.g., precipitation, evaporation, relative humidity, water vapor pressure, total solar radiation, temperature and wind speed), and even social economic parameters (e.g., gross domestic product, municipal sewage generation, and population density). In subsequent studies, ancillary data should be collected to improve upon the conclusions drawn from this study. Additionally, due to the uncertainties associated with model parameters, structure and input data, methods to quantify the prediction uncertainty would be beneficial for data interpretation (Noori et al., 2015). Notably, the SVR modeling approach for $\delta^{15}N–NO_3^-$ requires extensive future testing across a wide range of environmental conditions to examine the robustness of the approach. Finally, it is advantageous to apply the model output to the interpretation of mechanistic associations with the various input variables to provide a process-level understanding of the model output. Hence, a major benefit of the model predictions will be to supplement and enhance the interpretation of limited stable isotope measurements within complex real-world settings.

6. Conclusions

This study is the first of its kind to investigate the efficacy of SVR model to predict $\delta^{15}N–NO_3^-$ values in surface waters using basic hydro-chemical variables. The SVR model exhibited very good performances metrics for R^2 (0.88), *NS* (0.87), and *MSE* (0.53‰) using a testing data set. The overall results demonstrated the efficacy of machine learning models, as an indirect method, for estimating nitrate isotopic compositions in surface waters. Machine learning models are a simple, low cost and time saving approach for estimating environment isotopes from commonly measured water quality parameters. This methodology supplements and enhances measured $\delta^{15}N–NO_3^-$ values to generate a better understanding of nitrate sources and transformation processes, especially in developing countries where the high-cost of stable isotope analysis precludes extensive isotope investigations. Future studies should target validation of this machine learning approach to other catchments across the globe, investigate additional water quality variables (as dependent and independent variable), incorporate model uncertainty metrics and investigate process-level interpretation of model results.

Credit author statement

Yue Yang: Methodology, Investigation, Writing – original draft, Writing – review & editing. Xu Shang: Conceptualization, Methodology. Zheng Chen: Investigation. Kun Mei: Investigation. Zhenfeng Wang: Investigation. Randy A. Dahlgren: Formal analysis, Writing – original draft, Writing – review & editing. Minghua Zhang: Supervision. Xiaoliang Ji: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (Grant No. 51979197), Second Tibetan Plateau Scientific Expedition and Research Program (STEP) (Grant No. 2019QZKK0903), Science and Technology Project of Wenzhou Municipal Science and Technology Bureau (Grant No. S20180005), and Science Research Funding of Wenzhou Medical University (Grant No. QTJ18032). We acknowledge the Environmental Stable Isotope Lab (Chinese Academy of Agricultural Sciences, Beijing, China) for their analytical support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2021.112674.

References

Abrougui, K., Gabsi, K., Mercatoris, B., Khemis, C., Amami, R., Chehaibi, S., 2019. Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). Soil Till. Res. 190, 202–208. https://doi.org/10.1016/j.still.2019.01.011.

Adnan, R.M., Liang, Z.M., Heddam, S., Zounemat-Kermani, M., Kisi, O., Li, B.Q., 2020. Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. J. Hydrol. 586, 124371. https://doi.org/10.1016/j.jhydrol.2019.124371.

Ahmed, A.N., Othman, F.B., Afan, H.A., Ibrahim, R.K., Fai, C.M., Hossain, M.S., Ehteram, M., Elshafie, A., 2019. Machine learning methods for better water quality prediction. J. Hydrol. 578, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084.

Antanasijević, D., Pocajt, V., Povrenović, D., Perić-Grujić, A., Ristić, M., 2013. Modelling of dissolved oxygen content using artificial neural networks: danube River, North Serbia, case study. Environ. Sci. Pollut. Res. 20, 9006–9013. https://doi.org/10.1007/s11356-013-1876-6.

Banadkooki, F.B., Ehteram, M., Ahmed, A.N., Teo, F.Y., Ebrahimi, M., Fai, C.M., Huang, Y.F., El-Shafie, A., 2020. Suspended sediment load prediction using artificial neural network and ant lion optimization algorithm. Environ. Sci. Pollut. Res. 27, 38094–38116. https://doi.org/10.1007/s11356-020-09876-w.

Burow, K.R., Nolan, B.T., Rupert, M.G., Dubrovsky, N.M., 2010. Nitrate in groundwater of the United States, 1991-2003. Environ. Sci. Technol. 44, 4988–4997. https://doi.org/10.1021/es100546y.

Casciotti, K.L., Sigman, D.M., Hastings, M.G., Bohlke, J.K., Hilkert, A., 2002. Measurement of the oxygen isotopic composition of nitrate in seawater and freshwater using the denitrifier method. Anal. Chem. 74, 4905–4912. https://doi.org/10.1021/ac020113w.

Cerar, S., Mezga, K., Zibret, G., Urbanc, J., Komac, M., 2018. Comparison of prediction methods for oxygen-18 isotope composition in shallow groundwater. Sci. Total Environ. 631–632, 358–368. https://doi.org/10.1016/j.scitotenv.2018.03.033.

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM T. Intel. Syst. Tec. 2 https://doi.org/10.1145/1961189.1961199, 27:1–27:27.

Chen, K.Y., Chen, H., Zhou, C.L., Huang, Y.C., Qi, X.Y., Shen, R.Q., Liu, F.R., Zuo, M., Zou, X.Y., Wang, J.F., Zhang, Y., Chen, D., Chen, X.G., Deng, Y.F., Ren, H.Q., 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Res. 171, 115454. https://doi.org/10.1016/j.watres.2019.115454.

Dibike, Y.B., Velickov, S., Solomatine, D., Abbott, M.B., 2001. Model induction with support vector machines: introduction and applications. J. Comput. Civ. Eng. 15, 208–216. https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208).

Diez-Sierra, J., del Jesus, M., 2020. Long-term rainfall prediction using atmospheric synoptic patterns in semiarid climates with statistical and machine learning methods. J. Hydrol. 586, 124789. https://doi.org/10.1016/j.jhydrol.2020.124789.

Fadhullah, W., Yaccob, N.S., Syakir, M.I., Muhammad, S.A., Yue, F.J., Li, S.L., 2020. Nitrate sources and processes in the surface water of a tropical reservoir by stable isotopes and mixing model. Sci. Total Environ. 700, 134517. https://doi.org/10.1016/j.scitotenv.2019.134517.

Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J.M., Fernandez, L., 2000. Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Res. 34, 807–816. https://doi.org/10.1016/S0043-1354(99)00225-0.

Hsu, C.W., Chang, C.C., Lin, C.J., 2007. A practical guide to support vector classification. https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Hu, M.M., Wang, Y.C., Du, P.C., Shui, Y., Cai, A.M., Lv, C., Bao, Y.F., Li, Y.H., Li, S.Z., Zhang, P.W., 2019. Tracing the sources of nitrate in the rivers and lakes of the southern areas of the Tibetan Plateau using dual nitrate isotopes. Sci. Total Environ. 658, 132–140. https://doi.org/10.1016/j.scitotenv.2018.12.149.

Huang, Y., Zhao, L., 2018. Review on landslide susceptibility mapping using support vector machines. Catena 165, 520–529. https://doi.org/10.1016/j.catena.2018.03.003.

Husic, A., Fox, J., Adams, E., Pollock, E., Ford, W., Agouridis, C., Backus, J., 2020. Quantification of nitrate fate in a karst conduit using stable isotopes and numerical modeling. Water Res. 170, 115348. https://doi.org/10.1016/j.watres.2019.115348.

Ji, X.L., Lu, J., 2018. Forecasting riverine total nitrogen loads using wavelet analysis and support vector regression combination model in an agricultural watershed. Environ. Sci. Pollut. Res. 25, 26405–26422. https://doi.org/10.1007/s11356-018-2698-3.

Ji, X.L., Shang, X., Dahlgren, R.A., Zhang, M.H., 2017a. Prediction of dissolved oxygen concentration in hypoxic river systems using support vector machine: a case study of Wen-Rui Tang River, China. Environ. Sci. Pollut. Res. 24, 16062–16076. https://doi.org/10.1007/s11356-017-9243-7.

Ji, X.L., Xie, R.T., Hao, Y., Lu, J., 2017b. Quantitative identification of nitrate pollution sources and uncertainty analysis based on dual isotope approach in an agricultural watershed. Environ. Pollut. 229, 586–594. https://doi.org/10.1016/j.envpol.2017.06.100.

Keerthi, S.S., Lin, C.J., 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput. 15, 1667–1689. https://doi.org/10.1162/089976603321891855.

Kendall, C., Elliott, E.M., Wankel, S.D., 2007. Tracing anthropogenic inputs of nitrogen to ecosystems. In: Stable Isotopes in Ecology and Environmental Science, second ed. Blackwell, New york, pp. 375–449.

Kendall, C., McDonnell, J.J., 1998. Isotope Tracers in Catchment Hydrology. Elsevier, Amsterdam.

Kim, S., Alizamir, M., Zounemat-Kermani, M., Kisi, O., Singh, V.P., 2020. Assessing the biochemical oxygen demand using neural networks and ensemble tree approaches in South Korea. J. Environ. Manag. 270, 110834. https://doi.org/10.1016/j.jenvman.2020.110834.

Kisi, O., Parmar, K.S., 2016. Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. J. Hydrol. 534, 104–112. https://doi.org/10.1016/j.jhydrol.2015.12.014.

Kuter, S., 2021. Completing the machine learning saga in fractional snow cover estimation from MODIS Terra reflectance data: random forests versus support vector regression. Remote Sens. Environ. 255, 112294. https://doi.org/10.1016/j.rse.2021.112294.

Legates, D.R., McCabe Jr., G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241. https://doi.org/10.1029/1998WR900018.

Li, Y.Q., Yan, C., Liu, W., Li, M.Z., 2018. A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. Appl. Soft Comput. 70, 1000–1009. https://doi.org/10.1016/j.asoc.2017.07.027.

Lin, S.W., Lee, Z.J., Chen, S.C., Tseng, T.Y., 2008. Parameter determination of support vector machine and feature selection using simulated annealing approach. Appl. Soft Comput. 8, 1505–1512. https://doi.org/10.1016/j.asoc.2007.10.012.

Mamun, M., Kim, J.J., Alam, M.A., An, K.G., 2020. Prediction of algal chlorophyll-a and water clarity in monsoon-region reservoir using machine learning approaches. Water 12, 30. https://doi.org/10.3390/w12010030.

Mohammadpour, R., Shaharuddin, S., Chang, C.K., Zakaria, N.A., Ab Ghani, A., Chan, N.W., 2015. Prediction of water quality index in constructed wetlands using support vector machine. Environ. Sci. Pollut. Res. 22, 6208–6219. https://doi.org/10.1007/s11356-014-3806-7.

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: a review. ISPRS J. Photogramm. 66, 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001.

Nestler, A., Berglund, M., Accoe, F., Duta, S., Xue, D., Boeckx, P., Taylor, P., 2011. Isotopes for improved management of nitrate pollution in aqueous resources: review of surface water field studies. Environ. Sci. Pollut. Res. 18, 519–533. https://doi.org/10.1007/s11356-010-0422-z.

Noori, R., Yeh, H.D., Abbasi, M., Kachoosangi, F.T., Moazami, S., 2015. Uncertainty analysis of support vector machine for online prediction of five-day biochemical oxygen demand. J. Hydrol. 527, 833–843. https://doi.org/10.1016/j.jhydrol.2015.05.046.

Raghavendra, N.S., Deka, P.C., 2014. Support vector machine applications in the field of hydrology: a review. Appl. Soft Comput. 19, 372–386. https://doi.org/10.1016/j.asoc.2014.02.002.

Sarbu, C., Pop, H.F., 2005. Principal component analysis versus fuzzy principal component analysis-A case study: the quality of Danube water (1985-1996). Talanta 65, 1215–1220. https://doi.org/10.1016/j.talanta.2004.08.047.

Shang, X., Huang, H., Mei, K., Xia, F., Chen, Z., Yang, Y., Dahlgren, R.A., Zhang, M.H., Ji, X.L., 2020. Riverine nitrate source apportionment using dual stable isotopes in a drinking water source watershed of southeast China. Sci. Total Environ. 724, 137975. https://doi.org/10.1016/j.scitotenv.2020.137975.

Shoaib, M., Shamseldin, A.Y., Melville, B.W., Khan, M.M., 2016. A comparison between wavelet based static and dynamic neural network approaches for runoff prediction. J. Hydrol. 535, 211–225. https://doi.org/10.1016/j.jhydrol.2016.01.076.

Shu, C., Ouarda, T.B.M.J., 2008. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. J. Hydrol. 396, 31–43. https://doi.org/10.1016/j.jhydrol.2007.10.050.

Sigman, D.M., Casciotti, K.L., Andreani, M., Barford, C., Galanter, M., Böhlke, J.K., 2001. A bacterial method for the nitrogen isotopic analysis of nitrate in seawater and freshwater. Anal. Chem. 73, 4145–4153. https://doi.org/10.1021/ac010088e.

Singh, K.P., Basant, N., Gupta, S., 2011. Support vector machines in water quality management. Anal. Chim. Acta 703, 152–162. https://doi.org/10.1016/j.aca.2011.07.027.

Tikhamarine, Y., Malik, A., Souag-Gamane, D., Kisi, O., 2020. Artificial intelligence models versus empirical equations for modeling monthly reference evapotranspiration. Environ. Sci. Pollut. Res. 27, 30001–30019. https://doi.org/10.1007/s11356-020-08792-3.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York.

Wang, J., Du, H.Y., Liu, H.X., Yao, X.J., Hu, Z.D., Fan, B.T., 2007. Prediction of surface tension for common compounds based on novel methods using heuristic method and support vector machine. Talanta 73, 147–156. https://doi.org/10.1016/j.talanta.2007.03.037.

Wang, Y.J., Peng, J.F., Cao, X.F., Xu, Y., Yu, H.W., Duan, G.Q., Qu, J.H., 2020. Isotopic and chemical evidence for nitrate sources and transformation processes in a plateau lake basin in Southwest China. Sci. Total Environ. 711, 134856. https://doi.org/10.1016/j.scitotenv.2019.134856.

Wang, Z.F., Su, B.B., Xu, X.Q., Di, D., Huang, H., Mei, K., Dahlgren, R.A., Zhang, M.H., Shang, X., 2018. Preferential accumulation of small (<300 μm) microplastics in the sediments of a coastal plain river network in eastern China. Water Res. 144, 393–401. https://doi.org/10.1016/j.watres.2018.07.050.

Wang, Z.F., Zhou, J.Y., Zhang, C., Qu, L.Y., Mei, K., Dahlgren, R.A., Zhang, M.H., Xia, F., 2019. A comprehensive risk assessment of metals in riverine surface sediments across the rural-urban interface of a rapidly developing watershed. Environ. Pollut. 245, 1022–1030. https://doi.org/10.1016/j.envpol.2018.11.078.

World Health Organization, 2011. Guidelines for Drinking-Water Quality, fourth ed. (Geneva).

Xiao, X., He, J.Y., Huang, H.M., Miller, T.R., Christakos, G., Reichwaldt, E.S., Ghadouani, A., Lin, S.P., Xu, X.H., Shi, J.Y., 2017. A novel single-parameter approach for forecasting algal blooms. Water Res. 108, 222–231. https://doi.org/10.1016/j.watres.2016.10.076.

Xue, D.M., Botte, J., De Baets, B., Accoe, F., Nestler, A., Taylor, P., Van Cleemput, O., Berglund, M., Boeckx, P., 2009. Present limitations and future prospects of stable isotope methods for nitrate source identification in surface-and groundwater. Water Res. 43, 1159–1170. https://doi.org/10.1016/j.watres.2008.12.048.

Yaseen, Z.M., Jaafar, O., Deo, R.C., Kisi, O., Adamowski, J., Quilty, J., El-Shafie, A., 2016. Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. J. Hydrol. 542, 603–614. https://doi.org/10.1016/j.jhydrol.2016.09.035.

Yoon, H., Jun, S.C., Hyun, Y., Bae, G.O., Lee, K.K., 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J. Hydrol. 396, 128–138. https://doi.org/10.1016/j.jhydrol.2010.11.002.