Science of the Total Environment 753 (2021) 141821

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Multi-factor analysis of algal blooms in gate-controlled urban water bodies by data mining

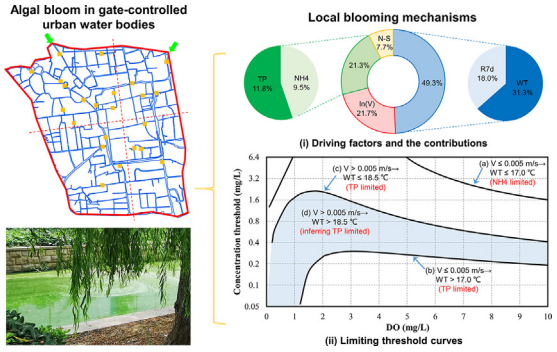
Ke Li [a,1], Te Xu [a,1], Jinying Xi [a,c], Haifeng Jia [a,b,c,*], Zhengjuan Gao [a,b,c], Zhaoxia Sun [a,b,c], Dingkun Yin [a], Linyuan Leng [a]

[a] School of Environment, Tsinghua University, Beijing, 100084, China
[b] Jiangsu Collaborative Innovation Center of Technology and Material of Water Treatment, Suzhou, China
[c] Research Institute for Environmental Innovation (Suzhou), Tsinghua, Suzhou 215163, China

HIGHLIGHTS

- Multi-factor effect in gate-controlled urban water bloom is intricate.
- Designed holistic framework uncovered the blooming mechanisms step by step.
- Collinearity and mutual causality in linear regression were treated prudently.
- Limiting threshold curves were depicted for joint regulation of bloom.
- Revealed compensation law appeals for stricter regulation for global warming.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 3 June 2020
Received in revised form 3 August 2020
Accepted 18 August 2020
Available online 21 August 2020

Editor: Ouyang Wei

Keywords:
Algal bloom
Gate-controlled
Multi-factor analysis
Limiting threshold
Data mining

ABSTRACT

Intense human disturbance has made algal bloom a prominent environmental problem in gate-controlled urban water bodies. Urban water bodies present the characteristics of natural rivers and lakes simultaneously, whose algal blooms may manifest multi-factor interactions. Hence, effective regulation strategies require a multi-factor analysis to understand local blooming mechanisms. This study designed a holistic multi-factor analysis framework by integrating five data mining techniques. First, the Kolmogorov–Smirnov test was conducted to screen out the possible explanatory variables. Then, correlation analyses and principal component analyses were performed to identify variable collinearity and mutual causality, respectively. After collinearity and mutual causality were treated prudently by using orthogonalization and instrumental variables, multilinear regression can be properly conducted to quantify factor contributions to algae growth. Lastly, a decision tree was used innovatively to depict the limiting threshold curves of each driving factor that restricts algae growth under different circumstances. The driving factors, their contributions, and the limiting threshold curves compose the complete blooming mechanisms, thus providing a clear direction for the targeted regulation task. A typical case study was performed in Suzhou, a Chinese city with an intricate gate-controlled river network. Results confirmed that climatic factors (i.e., water temperature and solar radiation), hydrodynamic factors (i.e., flow velocity), nutrients (i.e., phosphorus and nitrogen), and external loadings contributed 49.3%, 21.7%, 21.3%, and 7.7%, respectively, to algae growth. These results indicate that a joint regulation strategy is urgently required. Future studies can focus on coupling the revealed mechanisms with an ecological model to provide a comprehensive toolkit for the optimization of an adaptive joint regulation plan under the background of global warming.

* Corresponding author at: School of Environment, Tsinghua University, Beijing, China.
E-mail address: jhf@tsinghua.edu.cn (H. Jia).
[1]These two authors contributed equally to the research.

https://doi.org/10.1016/j.scitotenv.2020.141821
0048-9697/

1. Introduction

With human disturbance, such as pollutant discharge and gate control, many rivers and lakes in the world (e.g., the Rhine River in Europe and the Dianchi Lake in China) have experienced persistent algal blooms, causing serious negative impacts on aquatic ecosystems (Zheng et al., 2006; Friedrich and Pohlmann, 2009; Xia et al., 2019). For decades, multiple actions such as source point pollution control, water transfer, and ecological remediation have been taken by governments to control algal bloom (Hu et al., 2008; Xie et al., 2009; Xu et al., 2011; Jiang et al., 2020; Zhu et al., 2021). However, substantial work still has to be accomplished (Wang et al., 2019). One of the difficulties of algal bloom regulation is the mismatch of regulation measures and blooming mechanisms. Although many studies have confirmed that hydrodynamic factors, nutrients, and climatic factors can affect the growth of algae (Hilton et al., 2006; Zhou et al., 2017; Cheng et al., 2019; Huo et al., 2019), different water bodies could vary in terms of driving factors and individual contributions due to different local conditions. Therefore, a multi-factor analysis must be prioritized to identify the mechanisms for a new case before conducting any aimless attempt.

River–lake classification typifies two representative kinds of algal blooms in natural water bodies; however, algal blooms in urban water bodies show the characteristics of rivers and lakes simultaneously (Hilton et al., 2006). Similar to rivers, the perimeter–area ratio of urban water bodies is much larger than that of lakes and reservoirs, bringing intense substance exchange, including point or nonpoint source nutrients, in and out of the water body. Similar to lakes and reservoirs, urban water bodies often have poor hydrodynamic conditions, i.e., low velocity and small fluctuations, as a result of the regulation of gates, weirs, and pumps for flood prevention and landscaping needs (Bae and Seo, 2018). Therefore, blooming mechanisms in gate-controlled urban water bodies could be complex and may manifest multi-factor characteristics, which are totally different from those in natural water bodies.

In China, local governments are determined to improve the sensory quality of urban water environments. Staged achievements have been made in the urban black-odor water body remediation work (MOHURD, 2015). However, this work could not solve the eutrophication problem, which coupled to poor hydrodynamic conditions and improved transparency, would turn urban water bodies even more suitable for algae growth (Hu et al., 2019). Even worse, algal blooms can possibly ruin the achievement of urban black-odor water body remediation because of the decay of large quantities of dead algae. On this basis, algal bloom regulation will become one of the highlights of future urban water environment management policies in China (Hu et al., 2019). Among all the Chinese cities, those in the plain river network area require algal bloom regulation most urgently. The plain river network area covers the most developed and densely populated regions, such as the Yangtze River Delta (YRD) and the Pearl River Delta. Water as a cultural element of these cities has penetrated thousands of households through the river network. Large-scale algal blooms could have a strong negative impact on the residents and the economy.

Existing studies are mostly concerned about the mechanisms of algal blooms in natural water bodies, such as large rivers, lakes, and estuaries (Xiao et al., 2017; Shen et al., 2019; Sun et al., 2019), whereas only a few has investigated algal blooms in gate-controlled urban water bodies. Hence, this study proposes a holistic multi-factor analysis framework by integrating multiple data mining techniques for identifying the local mechanisms of algal blooms in gate-controlled urban water bodies. Suzhou, a YRD city, is selected for a typical case study. Many studies have already attempted to use one or a few unsupervised data mining techniques which require no pre-existing labels to identify the driving factors of algal blooms in natural water bodies and their contributions (Zhou et al., 2017; Cheng et al., 2019; Wang et al., 2019). However, to our furthest knowledge, no research has provided a limiting threshold curve that is practical for algal bloom regulation. Though unsupervised techniques can help dig out the patterns within the dataset with minimum prior knowledge, supervised techniques are more reliable in a classification problem (i.e. identifying the blooming samples) with labelled data. Thus, Supervised techniques are included in this study to distinguish the threshold of each driving factor below or over which algal blooms are more likely to outbreak. Limiting threshold curves have been maturely applied in some environmental problems, such as Empirical Kinetic Modeling Approach for ozone control (Gipson et al., 1981) and Noise Criteria curves for noise control (Beranek, 1957). Our limiting threshold curve makes an important supplement to the factor contributions to compose the complete blooming mechanisms, thus providing a clear direction for the regulation work.

2. Methodology

Data mining techniques based on statistics and machine learning allows the model to learn the relationships among variables from numerous observations (Anderson et al., 2010; Wang et al., 2019). These models are not only computationally efficient but also flexible in terms of structure, making it adaptable to local blooming characteristics (Shen et al., 2019). Among data mining techniques, the unsupervised ones can dig out the linear or nonlinear relationships among variables and identify the key influential factors of blooms, whereas the supervised ones can distinguish blooming (i.e., Chl-a over 30 μg/L defined in this paper) and non-blooming samples to help predict whether the regulation strategy can meet the environmental goals.

This study establishes a data mining framework of five techniques. The framework is intended to be performed on a dataset consisting of a response variable representing algal growth and possible influential factors including climate factors, hydrodynamic factors, nutrients external loadings, and other variables. The response variable can be chlorophyll *a* (Chl-a) or algae density, and Chl-a is selected in this study. The variables representing influential factors are listed in Section 3.2 and Table 1.

The roadmap is displayed in Fig. 1. First, the Kolmogorov–Smirnov (K–S) test is conducted for the preliminary screening of possible explanatory variables affecting algae growth for multilinear regression (MLR). Then, correlation analysis and principal component analysis (PCA) are performed to identify collinearity and mutual causality among variables, respectively. Collinearity and mutual causality can weaken the performance of an MLR model. Orthogonalization is performed to address collinearity, and instrumental variables are constructed to address mutual causality. Afterward, MLR is performed to estimate standardized regression coefficients (SRCs). On the basis of SRCs, the individual contribution of each driving factor to algae growth are properly quantified. Meanwhile, reasonable forms of explanatory variables are also determined for the decision tree. Lastly, the decision tree uncovers driving factors' limiting threshold curves that restrict algae growth.

Although the K–S test, correlation analysis, PCA, MLR, and decision tree are all well-developed techniques that can be found in many textbooks, this study effectively organized them to complement each other and uncover well-knit blooming mechanisms, including the driving factors, their individual contributions and the limiting threshold curves. The framework can be widely applied to different cases to investigate local blooming mechanisms that benefit the algal bloom regulation.

2.1. K–S test: preliminary screening of possible explanatory variables

The K–S test discriminates if a variable in two sets are differently distributed without any prior information about variable distributions (Conover, 1999). Existing studies have applied the K–S test to compare the behavior of aquatic organisms in different conditions (Harvey and Menden-Deuer, 2011), evaluate a water environment model (Duda et al., 2012), and check data distributions (Parinet et al., 2010). In this study, K–S test is adopted to examine if a variable is differently distributed in blooming and non-blooming sets to qualify a possible explanatory variable.

Table 1
A summary of the monitored dataset.

Variables		Abbreviation	Periods	Frequency	Valid samples	Source
Response variable	Chlorophyll a ($\mu g \cdot L^{-1}$)	Chl-a	From 2017.11.04 to 2019.01.16	Biweekly	536	Manual sampling, Spectrophotometric method
Climatic factors	Water temperature (°C)	WT	From 2017.01.01 to 2018.12.31	Daily	536	Suzhou weather station
	7-day average solar radiation ($W \cdot m^{-2}$)	R7d	From 2017.01.01 to 2018.12.31	Daily	513	Suzhou weather station
	7-day precipitation (mm)	P7d	From 2017.01.01 to 2018.12.31	Daily	513	Suzhou weather station
Hydrodynamic factors	Flow velocity ($m \cdot s^{-1}$)	V	From 2017.10.01 to 2018.12.31	Daily	452	Automatic monitoring station
	Water level fluctuation (%)	ΔH	From 2017.10.01 to 2018.12.31	Daily	452	Automatic monitoring station
Nutrients	Total nitrogen ($mg \cdot L^{-1}$)	TN	From 2017.11.04 to 2019.01.16	Biweekly	276	Manual sampling, filtrated Alkaline potassium persulfate digestion - UV spectrophotometric method
	Ammonia nitrogen ($mg \cdot L^{-1}$)	NH4	From 2017.11.04 to 2019.01.16	Biweekly	515	Manual sampling, filtrated Nessler's reagent colorimetric method
	Total phosphorus ($mg \cdot L^{-1}$)	TP	From 2017.11.04 to 2019.01.16	Biweekly	507	Manual sampling, filtrated Ammonium molybdate spectrophotometric method
	Nitrogen-to-phosphorus ratio	N/P	From 2017.11.04 to 2019.01.16	Biweekly	507	Calculated with NH4 and TP
External loadings	South-or-north dummy variable	N-S	/	/	536	/
	East-or-west dummy variable	W-E	/	/	536	/
Others	Dissolved oxygen ($mg \cdot L^{-1}$)	DO	From 2017.11.04 to 2019.01.16	Biweekly	536	Manual sampling, Winkler's method

First, the cumulative frequency distribution function of each influential factor x in the two sets are calculated and denoted as F_1 and F_2. Then, the K–S distance is defined by the maximum difference between F_1 and F_2 and expressed as Eq. (1). A multiplier constructed by the number of valid samples in the two sets is added to give a greater K–S distance to the factor with more valid samples. Although following analysis may take different forms of variables, Although the following analysis steps may take different forms of variables, a rank-preserved transformation like logarithm does not affect the calculation of K–S distance. From this perspective, the K–S test is an appropriate method for variable prescreening.

$$d = \max_x \{F_1(x) - F_2(x)\} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad (1)$$

2.2. Correlation analysis: identification of variable collinearity

The variables selected by the K–S test may be collinear and thus must be properly handled to ensure a valid MLR. Variable collinearity means high intercorrelations within a set of explanatory variables. An MLR model with severe variable collinearity can yield an unreliable estimation of coefficients (Farrar and Glauber, 1967). Correlation analysis can help detect variable collinearity. Two variables with a correlation coefficient larger than 0.7 are often considered collinear. Collinearity can be examined by the variance inflation factor (VIF) as well (Kennedy, 1992).

Correlation coefficients can be calculated using the Pearson or Spearman methods. With variables usually in original form, the Pearson method measures the linear correlation between variables x and y, which is expressed as Eq. (2) (Pearson, 1895). The Spearman method

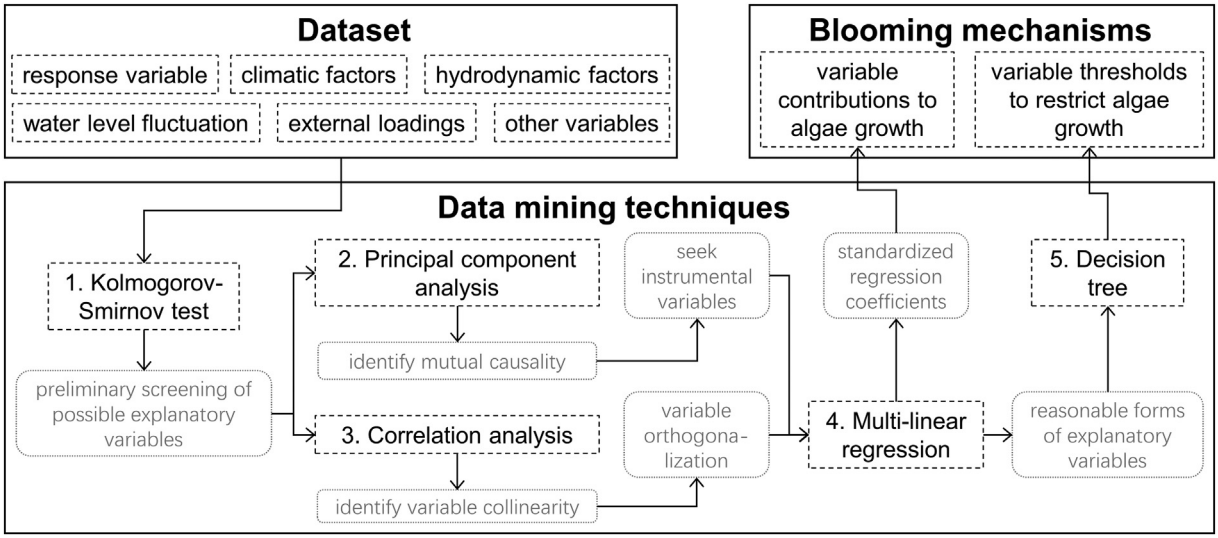


Fig. 1. Multi-factor analysis roadmap.

uses ranks (i.e., orders) of x and y, R and S to calculate the correlation coefficient, which is expressed as Eq. (3) (Spearman, 1904).

$$\rho(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{2}$$

$$\rho_s(x,y) = \frac{\sum_i (R_i - \bar{R})\left(S_i - \bar{S}\right)}{\sqrt{\sum_i (R_i - \bar{R})^2 \sum_i \left(S_i - \bar{S}\right)^2}} \tag{3}$$

The Pearson coefficient is sensitive to extreme values as extreme values sharply deviate from a linear pattern and decrease correlation coefficient excessively. During the blooming phase, many extreme values are monitored, thus attenuating the applicability of the Pearson method. On the contrary, extreme values do not change the variable ranks and thus make no difference to the Spearman coefficient. However, monitoring data tend to fall within narrow intervals. Potential monitoring errors could lead to discrepancies between the monitored ranks and the actual situation, resulting in the malfunction of the Spearman coefficient. Therefore, a compromise in which the two coefficients can be used together for correlation analysis is encouraged in this study.

2.3. PCA: identification of mutual causality between two variables

Mutual causality between a response variable and an explanatory variable is likely to undermine the exogeneity prerequisite for MLR. PCA is introduced to identify mutual causality between two variables. PCA transforms the data into another orthogonal space where each dimension represents a principal component. Principal components are linear combinations of original variables that can reflect the relationships among variables, including mutual causality (Trevor et al., 2001).

PCA assumes that the information contained in the data is reflected in the variance and attempts to preserve the maximum information during transformation. PCA identifies linear combinations of the variables that display the variance in a descending order, and in this process, the latter ones should be orthogonal to the former ones, as shown in Fig. 2(a). These linear combinations are noted as principal components, of the same number as the variables at most. Mathematically, each principal component corresponds to an eigenvector ($\boldsymbol{v}$) of the covariance matrix of the variables, expressed as Eq. (4). Given that the eigenvalues (λ) are proportional to the variance of the principal components, one's ratio to the sum of all represents the variance contribution of the corresponding principal component.

$$\begin{bmatrix} \mathrm{cov}(y,y) & \mathrm{cov}(y,x_1) & \cdots & \mathrm{cov}(y,x_n) \\ \mathrm{cov}(x_1,y) & \mathrm{cov}(x_1,x_1) & \cdots & \mathrm{cov}(x_1,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(x_n,y) & \mathrm{cov}(x_n,x_1) & \cdots & \mathrm{cov}(x_n,x_n) \end{bmatrix} \boldsymbol{v} = \lambda \boldsymbol{v} \tag{4}$$

PCA preserves most of the information in the first few components with the largest eigenvalues, and each of them reflects various relationships among variables (Reid and Spencer, 2009). This study intends to use the first two components to investigate mutual causality. That is, opposite relationships between Chl-a and an explanatory variable often imply an existing mutual causality.

2.4. MLR: determining the reasonable form and the contribution of each driving factor to algae growth

2.4.1. Theory

MLR uses a linear equation to describe the relationships between potential explanatory variables ($x_1,\ldots,x_n$) and response variable y (i.e., Chl-a), expressed as Eq. (5), where ε is the residual item, and the unknown parameters ($b_0, b_1,\ldots,b_n$) are estimated via the least square method (Wooldridge, 1960). With the help of t-test on the estimated parameters and model fitness index (i.e., the adjusted R^2), significant explanatory variables (driving factors) of algae growth and their reasonable forms can be eventually determined. Lastly, the individual contribution of each driving factor can be calculated based on SRCs, as expressed in Eqs. (6) and (7) (Saltelli et al., 2004).

$$y = b_0 + b_1 x_1 + \ldots + b_n x_n + \varepsilon \tag{5}$$

$$b_i' = \left| \hat{b}_i \times \frac{\sigma_{x_i}}{\sigma_y} \right| \tag{6}$$

$$\varphi_i^{SRC} = \frac{\hat{b}_i'}{\sum_i \hat{b}_i'} \tag{7}$$

An appropriate application of MLR must satisfy three prerequisites (Wooldridge, 1960). First, the predicted residuals must be individually identically distributed as $N(0,\sigma)$. Second, severe collinearity seldom exists among explanatory variables, which would otherwise yield unreliable results. Third, all the included explanatory variables must be exogenous relative to the response variable, i.e., $\rho(x_i,\varepsilon)$, which would otherwise lead to biased estimators. Mutual causality may undermine the exogeneity prerequisite. Although numerous studies preferred pure MLR to explain and predict the variation trend of Chl-a (Cho et al., 2009; Franklin et al., 2020; Liu et al., 2014), the above prerequisites were often neglected.

2.4.2. Jarque–Bera test

The Jarque-Bera (J-B) test on the predicted residual helps check whether the first prerequisite is fulfilled (Jarque and Bera, 1980). The J-B test statistic is defined by Eq. (8), where n is the number of observations, and S and K are the sample skewness and sample kurtosis, which are calculated by Eqs. (9) and (10), respectively. If the test statistic is

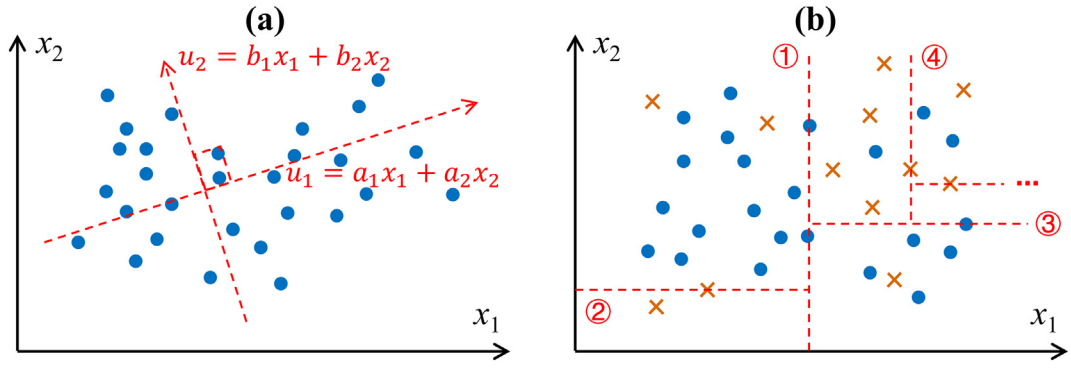


Fig. 2. Simple examples of PCA (a) and decision tree (b).

significantly greater than zero, then the predicted residual does not obey a normal distribution.

$$JB = \frac{n}{6}\left[S^2 + \frac{1}{4}(K-3)^2\right] \tag{8}$$

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \overline{\varepsilon})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \overline{\varepsilon})^2\right]^{3/2}} \tag{9}$$

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \overline{\varepsilon})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \overline{\varepsilon})^2\right]^{2}} \tag{10}$$

2.4.3. Mutual causality and instrumental variables

A valid approach to address mutual causality is to introduce an instrumental variable w that meets $\rho(x_i, w) \neq 0$ and $\rho(w, \varepsilon) = 0$. By performing linear regression of x_i on w, and then substituting $\hat{x}_i$ for x_i to perform MLR on y, expressed as Eqs. (11) to (13), the mutual causality can be removed as $\rho(\hat{x}_i, \varepsilon) = \rho(w, \varepsilon) = 0$. Correlation analysis on w and the predicted residual can help to check if the exogeneity is fulfilled.

$$x_i = a_0 + a_1 w + \delta \tag{11}$$

$$\hat{x}_i = \hat{a}_0 + \hat{a}_1 w \tag{12}$$

$$y = b_0 + \ldots + b_{i-1}x_{i-1} + b_i\hat{x}_i + b_{i+1}x_{i+1} + \ldots + b_n x_n + \varepsilon \tag{13}$$

2.4.4. Collinearity and orthogonalization

Given that the collinear variables contain the same information about the response variable, their independent impacts on the response variable are less reliably estimated in comparison with those in the non-collinear case. To eliminate collinearity, the orthogonalization procedure can create transformed variables that are uncorrelated with each other (Klein and Chow, 2013). For any two correlated variables x_i and x_j in the Euclidean space, orthogonalization is conducted as Eqs. (14) and (15). After orthogonalization, the information contained in the correlated part is entirely removed from x_j; thus, $\rho(\tilde{x}_i, \tilde{x}_j) = 0$.

$$\tilde{x}_i = x_i \tag{14}$$

$$\tilde{x}_j = x_j - \frac{\langle x_j, \tilde{x}_i\rangle}{\langle \tilde{x}_i, \tilde{x}_i\rangle}\tilde{x}_i \tag{15}$$

2.5. Decision tree: determining the limiting threshold of each driving factor

Decision trees are a simple but effective technique for solving classification problems (Breiman et al., 1984), and have been used to classify algal bloom species from remote sensing images (Ghatkar et al., 2019). However, no research has ever used decision tree to investigate blooming mechanisms.

Decision trees make a tree-like classification routine by successively looking for the optimal division of the optimal factor to achieve the best purity of one sample subset. In the end, all samples are subdivided in the form of a binary tree (Fig. 2(b)). The purity of the sample subset is measured by information entropy, which is expressed as Eq. (16), where p and $1 - p$ are the proportions of the two categories in the sample subset, respectively, and w provides one category with a proper weight for an imbalanced classification problem. If the sample subset has only one category, then $S = 0$; If the two categories each counts a half, then $S = 1$. For a decision tree that is well trained, i.e., can effectively distinguish which category a sample belongs to, the corresponding division criteria are the limiting thresholds of the driving factors. Following the division criteria, an unknown sample ends in a leaf of the tree, whose majority is the predicted category of the sample.

$$S = -\left[\frac{wp}{(w-1)p+1}\log_2\frac{wp}{(w-1)p+1} + \frac{1-p}{(w-1)p+1}\log_2\frac{1-p}{(w-1)p+1}\right] \tag{16}$$

Decision trees are a supervised learning method for which overfitting should be avoided during training. Continuously partitioning the training set until each leaf has only one sample results in a perfectly correct classification although at the cost of a terribly poor prediction ability for any unknown sample. Such result occurs because the monitoring errors of individual samples would eventually cover the overall features of the sample set. Therefore, the maximum depth of the tree or the minimum size of the leaf need to be assigned to stop tree-building before overfitting.

A five-fold cross validation method is applied to decide the optimal values of hyperparameters, including the maximum depth of the tree, the minimum size of the leaf, and the class weight w, with the help of grid search. The dataset is split into five sets. Each of them is kept for model validation with the other four sets for model training. The average performance on five validation sets discriminates the best combination of hyperparameters. Five-fold cross validation can make the most of data, is tolerant to noise, and can ease the overfitting problem (Kohavi, 1995).

3. Case study overview

3.1. Study site

Suzhou is a famous YRD city (Fig. 3(a)). The city is typical of the plain river network area, whose terrain slopes gently with over 20,000 rivers and a water surface rate of 15.4%. The red polylines in Fig. 3(b) outlines the central urban district with a total area of 78 km^2. The subtropical monsoon climate brings about an annual average temperature of approximately 16.6 °C and an average precipitation of approximately 1163 mm. In addition, the recent four decades witnessed a temperature rise of nearly 0.1 °C/year. For the sake of flood prevention and landscape needs, the government has built eight hydro-junctions, two large weirs, two diversion channels, and numerous pumps and gates to control the water in and out of the central urban district, as shown in Fig. 3(c). However, such intense interventions on the hydrodynamic factors along with the increasing pollution and climate warming have made algal bloom a severe threat for the ecosystem in the urban river network as well as the Lake Taihu. Although a few regulation strategies like water diversion and off-site processing have been implemented recently, the monitored Chl-a still exceeds the regulation goal, i.e., over 30 μg/L on average. This situation requires an urgent understanding of the local blooming mechanisms to guide further regulation work.

3.2. Dataset

This study monitored 23 typical river sections in the central urban area, as labelled in Fig. 3(c). After the first trial on Nov. 4, 2017 at several sections, the monitoring work persisted from January 16, 2018 to January 16, 2019, biweekly. During the first three months, monitoring work were disrupted at times at some sections. Finally, a total of 536 samples were collected. Considering the monitoring cost and complexity, only about 25% of the samples were selected for microscopic counting of algae species. The results showed that the Cyanobacteria accounted for about 60%–97% of the total number of algae cells. Chl-a was chosen to represent algae growth. Beside Chl-a, other monitored variables included hydrodynamic factors (i.e., flow velocity (V) and water level fluctuation (ΔH)), climatic factors (i.e., water temperature (WT), seven-day precipitation (P7d), and seven-day average solar radiation (R7d)), nutrients (i.e., total nitrogen (TN), ammonia nitrogen (NH4),

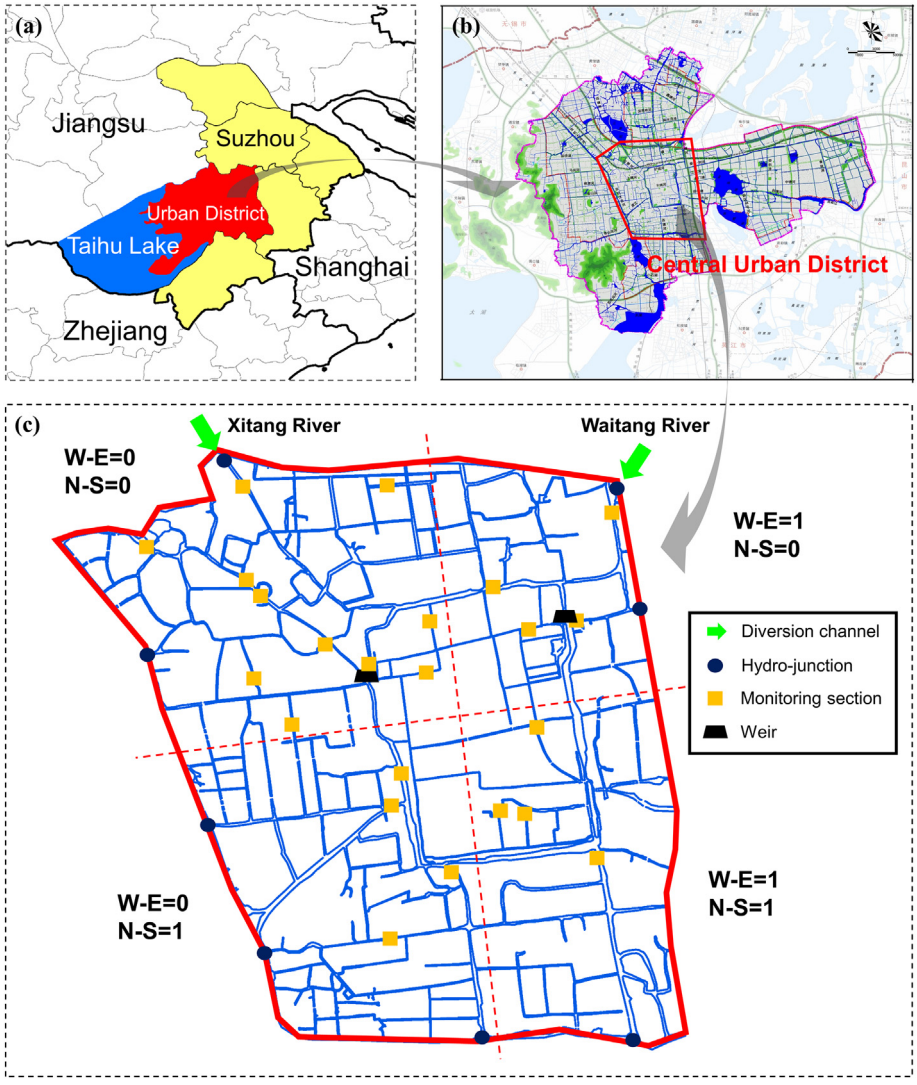


Fig. 3. Descriptions of the study sites.

and total phosphorus (TP)), and dissolved oxygen (DO). TN, TP, and NH4 were tested after having the water sample filtered to exclude the nitrogen and phosphorus contained in the algal cell. The atomic ratios of nitrogen to phosphorous (N/P ratio) were also calculated for analysis. As the number of missing values of TN counts almost a half, the N/P ratio was calculated with NH4 after examining the linear correlation of TN and NH4.

Dummy variables indicating categories are often introduced to combine quantitative and qualitative information in a regression model (Wooldridge, 1960). They can help to better describe response variable and omitting them can cause biased estimates of other coefficients in linear regression. Two dummy variables, namely N–S and W–E, were introduced by dividing the central urban area into four regions, i.e., northwest (N–S = 0 and W–E = 0), northeast (N–S = 0 and W–E = 1), southwest (N–S = 1 and W–E = 0), and southeast (N–S = 1 and W–E = 1) (Fig. 3(c)). In consideration of the concentric urban expansion history of the city, differences in other information, such as land use, water surface rate, drainage system, and human activities among the four regions, could be ignored. Thus, N–S and W–E mainly reflect whether and which diversion plays a role, respectively. The contrasts between the background and external loadings in the two water diversions are displayed in Fig. 4.

Detailed information of the dataset is summarized in Table 1.

4. Results and discussion

4.1. What are the possible explanatory variables to algae growth?

The blooming and non-blooming samples were taken as the two sets for the K–S test. The K–S distance and its significance are shown in Fig. 5. Variables significant at 10% level or above were selected (with red check mark) for the following analysis: For climatic factors, WT and R7d were kept. For hydrodynamic factors, only V was kept. For nutrients, TN, NH4, and TP were all kept. For external loadings, N–S was kept for greater importance to Chl-a compared with W–E. DO is also included in the subsequent analysis. The reason of higher significance of N–S is shown in Fig. 4: Chl-a concentrations in water diversions were higher than that in the background, thus significantly increasing the Chl-a concentration in the northern area, except in May 2018. In this month, no significant difference between the Chl-a concentrations were observed in the two diversions. Thus, whether implementing diversion or not played a more important role than the difference between the two implemented diversions.

4.2. Which variables have serious collinearity that may disturb MLR?

Prior to correlation analysis, a logarithm was taken for Chl-a and V to narrow their orders of data magnitude. The results are displayed in Fig. 6,

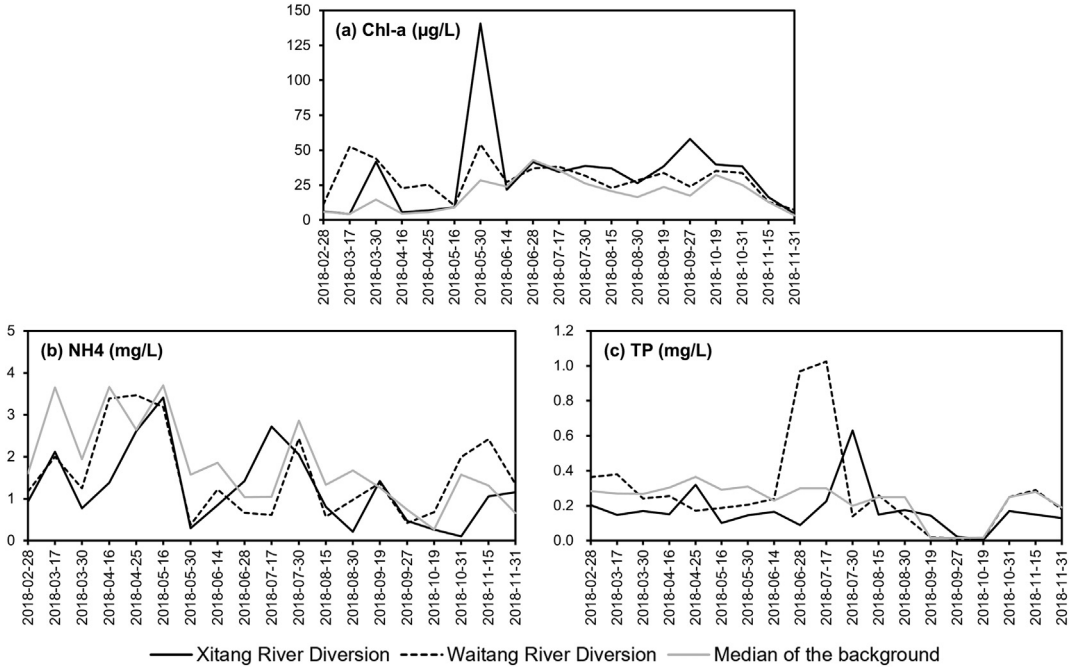


Fig. 4. Contrasts between the background and water diversion loadings.

where the upper right triangular part presents the Spearman coefficients, and the left lower triangular part shows the Pearson coefficients.

The variables significantly correlated to ln(Chl-a) for both methods were ln(V), R7d, WT, and N–S. Two pairs of variables, namely, TN-NH4 and R7d-WT, had a correlation coefficient of over 0.7. For the first pair of collinear variables, given that TN contained nearly half of the missing values (Table 1), the subsequent analysis steps directly omitted it. For the second pair of collinear variables, one can choose either omitting R7d or eliminating the information that R7d affects WT by orthogonalization, considering that solar radiation would affect water temperature. Orthogonalization result is expressed as Eq. (17), where WT′ denotes water temperature that is free from the effect of solar radiation. In addition, NH4 and TP also brings a glimmer of concern about collinearity. Nevertheless, after orthogonalization and omitting TN, the VIFs of NH4 and TP were only 2.12 and 1.95, respectively (Fig. 6); these values are both less than 10, thus indicating an ignorable collinearity (Kennedy, 1992).

$$WT' = WT - 0.0823 \times R7d \quad (17)$$

Although DO is a product of algae photosynthesis rather than an influential factor of algae growth, this section kept the results of DO. Its correlations with nutrients provide a valid approach for reducing possible mutual causality between Chl-a and nutrients, which is discussed in Section 4.3.

4.3. Which variables have a serious mutual causal link with Chl-a?

All variables excluding DO (WT was replaced by WT′) were adopted for PCA. The first two principal components, whose total variance contributions were 51.56%, are depicted in Fig. 7. The first principal component accounted for 32.72% contribution, in which the coefficients of WT, R7d, NH4, and TP had the same signs with ln(Chl-a), whereas the coefficients of ln(V) and N–S had the opposite signs with ln(Chl-a). This component well generalized the local blooming mechanisms that high temperature, intense solar radiation, and affluent nutrients all promoted algae growth, whereas high velocity restricted algae growth. The negative correlation between ln(Chl-a) and N–S is discussed in Section 4.4. The second principal component accounted for 18.84% contribution, in which the coefficients of ln(Chl-a) and nutrients had opposite signs.

Different signs between ln(Chl-a) and nutrients in the first and the second components implied the mutual causality between the two, especially during algal bloom or decay phases. Without question,

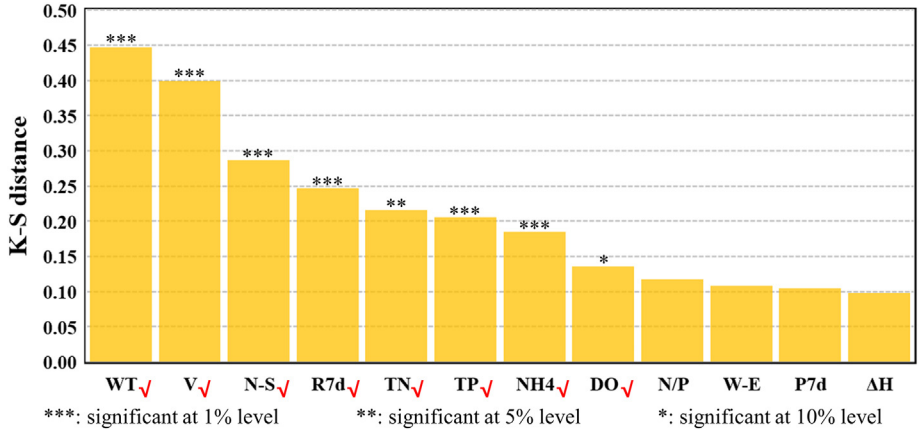


Fig. 5. Comparison of K–S distances.

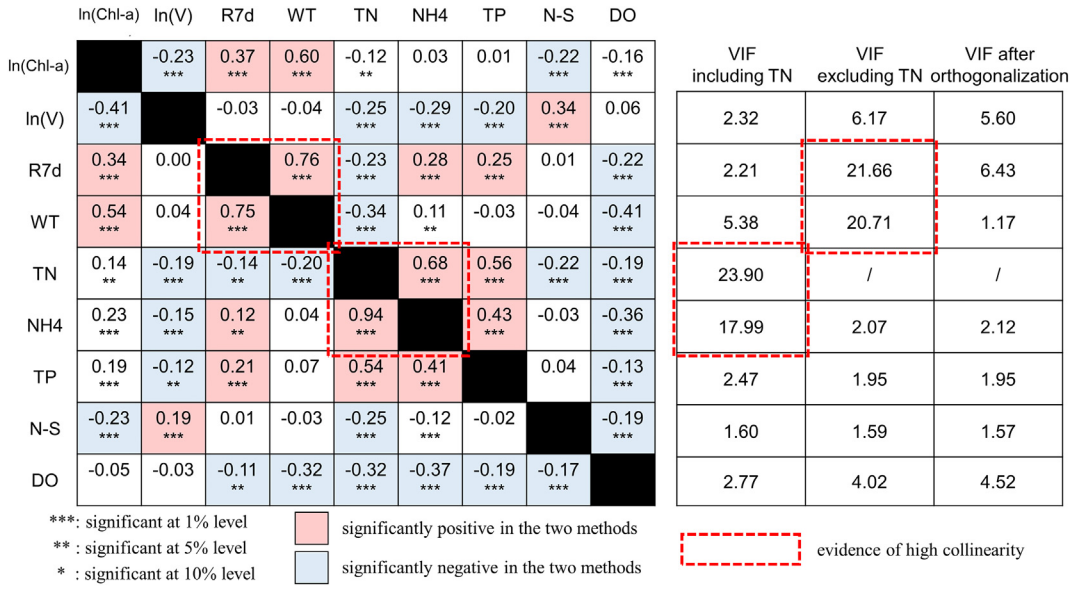


Fig. 6. Results of correlation analysis. The upper triangular parts are the results of the Spearman method, and the lower triangular parts are the results of the Pearson method. VIFs are attached on the right of the correlation coefficients in a tabular format.

nutrients promote algae growth. However, during the bloom or decay phase, algae overconsume or release nutrients, thus overturning the causality between Chl-a and nutrients. Therefore, instrumental variables should be introduced in place of the original nutrient terms, i.e., NH4 and TP, to reduce the mutual causality in the bloom and decay phases of algae. According to literature (Wu et al., 2012; Crossman et al., 2019), DO is a good index for distinguishing the bloom and decay phases. In fact, during the bloom phase, algae photosynthesis produces oxygen and oversaturates DO; whereas during the decay phase, the decomposition process of algal cells depletes DO. Hence, algae growth has opposite effects on the variations of DO and nutrients; this difference in effects could be proven by the significantly negative correlations between DO and nutrients in this study (Fig. 6) and indicates that the cross terms of DO and nutrients can be possible instrumental variables.

Based on the above discussion, stepwise regression on four cross terms of nutrients and DO (with or without logarithmic transformation) was performed. Two instrumental variables NH4′ and TP′ as expressed in Eqs. (18) and (19) were determined. Whether NH4′ and TP′ have a stronger explanatory power to algae growth compared with NH4 and TP would be verified in the next section, i.e., MLR.

$$\begin{aligned} NH4' &= 1.4737 + 0.0998 \times DO \times \ln(NH4) \\ &\quad + (0.1810 \times DO - 0.5030 \times \ln(DO)) \times NH4, R^2 \\ &= 0.67 \end{aligned} \tag{18}$$

$$\begin{aligned} TP' &= 0.3149 + 0.0619 \times \ln(DO) \times \ln(TP) \\ &\quad + (0.1892 \times DO - 0.3398 \times \ln(DO)) \times TP, R^2 \\ &= 0.71 \end{aligned} \tag{19}$$

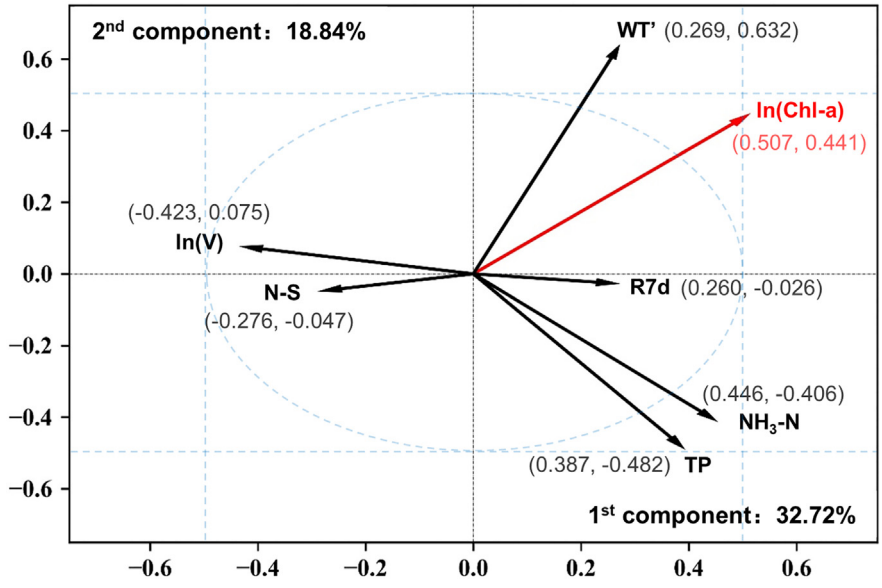


Fig. 7. Results of the first two components in PCA.

Table 2
Comparison of different MLR models.

Variables		(1)	(2)	(3)	(4)
Climatic factors	WT	0.0951***	0.0785***		
	WT′			0.0951***	0.1000***
	R7d	−0.0027**		0.0052***	0.0046***
Nutrients	NH4	0.0311**	0.0284**	0.0311**	
	NH4′				0.0496***
	TP	0.2372**	0.1854*	0.2372**	
	TP′				0.4576***
Hydrodynamic factor	ln(V)	−0.2184***	−0.2226***	−0.2184***	−0.1971***
External loadings	N-S	−0.2446***	−0.2559***	−0.2446***	−0.2397***
Constant		0.2257	0.2131	1.0494***	1.1137***
Adjusted R^2		0.432	0.425	0.432	0.457
AIC		1091	1096	1091	1072
BIC		1119	1120	1119	1110

Note: *, **, and *** mean the estimated coefficient is significant at 10%, 5%, and 1% levels, respectively.

4.4. What are the driving factors and their contributions to algae growth?

Based on the previous analysis steps, MLR considered WT, R7d, NH4, TP, ln(V), and N–S, including the adjusted substitutes (i.e., WT′, NH4′, and TP′ as the explanatory variables), with ln(Chl-a) as the response variable. The best model was determined by model comparison, as shown in Table 2.

Model (1) was a baseline model with all variables in original forms. The sign of R7d was negative in Model (1), opposite to our priori knowledge. The possible reason is the existence of collinearity between WT and R7d. Therefore, Model (2) directly omitted R7d, whereas Model (3) replaced WT with WT′. The three fitness indices, i.e., adjusted R^2, Akaike information criterion (AIC), and Bayesian information criterion (BIC), all showed preference for Model (3) over Model (2), as Model (2) might suffer from omitted relevant variable bias. Compared with Model (1), the adjusted R^2, AIC, and BIC in the two models were identical because orthogonalization brings no extra information to Model (3). However, on account of elimination of collinearity, coefficient of R7d in Model (3) showed greater significance and the coefficient became positive, which is in conformity with common sense.

Based on Model (3), Model (4) further adopted instrumental variables to replace the original nutrient terms. Compared with Model (3), Model (4) increased adjusted R^2 by 2.5% and reduced AIC and BIC by 19 and 9, respectively. So far, the explanatory power of Model (4) has been comparable with complex machine learning and mechanism models (Wu and Xu, 2011; Park et al., 2015). Moreover, in comparison with that in Model (3), the regression coefficients of nitrogen and phosphorus in Model (4) increased by 59.5% and 92.9%, respectively. These results showed stronger positive effects of nutrients on algae growth that were weakened by mutual causality, which was diminished effectively by the instrumental variables in Model (4). In addition, no significant correlation was detected between the instrumental variables and the predicted residuals of Model (4), indicating that NH4′ and TP′ satisfied the prerequisites of a qualified instrumental variable. The *p*-value of the J–B test was almost zero, indicating that the predicted residuals followed a normal distribution. Therefore, Model (4) satisfied all the three prerequisites of a reasonable MLR model. Hence, this study preferred Model (4) as the final model. Meanwhile, the reasonable forms of the explanatory variables, i.e., ln(V), WT′, R7d, NH4′, TP′, and N–S, were also determined. N–S had a negative estimated coefficient, indicating that an exogenous algae input would result in a worse aquatic ecosystem in the northern area (i.e., N–S = 0). In fact, compared with the background of the central urban district, water in the diversions had higher Chl-a concentrations (Fig. 4(a)). Therefore, off-site processing is urgently required to facilitate the function of water diversion.

The SRC contribution ratios of the six driving factors are shown in Fig. 8. Sorted from largest to smallest, the order of the SRC contribution ratios is as follows: climatic factors, hydrodynamic factor, nutrients, and external loadings. As MLR considered all explanatory variables simultaneously, SRC contribution ratios could exclude the variable interactions and therefore reflect the individual contribution of each driving factor (Saltelli et al., 2004).

Nearly half of the contributions came from climatic factors, thus emphasizing the importance of temperature and solar radiation to algae growth. Although the regulation work can hardly control the climatic factors directly, the potential negative impact of the global warming trend on the current aquatic ecosystem in Suzhou cannot be ignored. In addition, a few indirect regulation approaches are still accessible; an example is green infrastructure (GI) placement, which is greatly beneficial to cooling the local climate by mitigating the urban heat island effect (Norton et al., 2015). The effects of hydrodynamic factors and the nutrients were nearly equal, and the external loadings contributed 7.7%. These results raised the need of multifactor joint regulation.

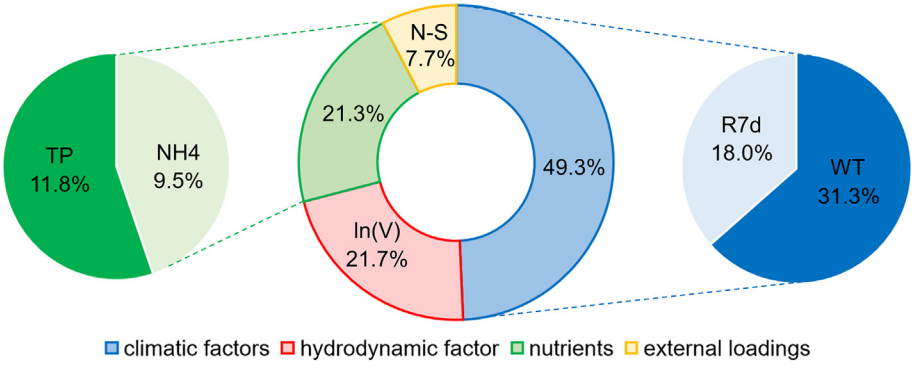


Fig. 8. Contributions of the influential variables to algae growth.

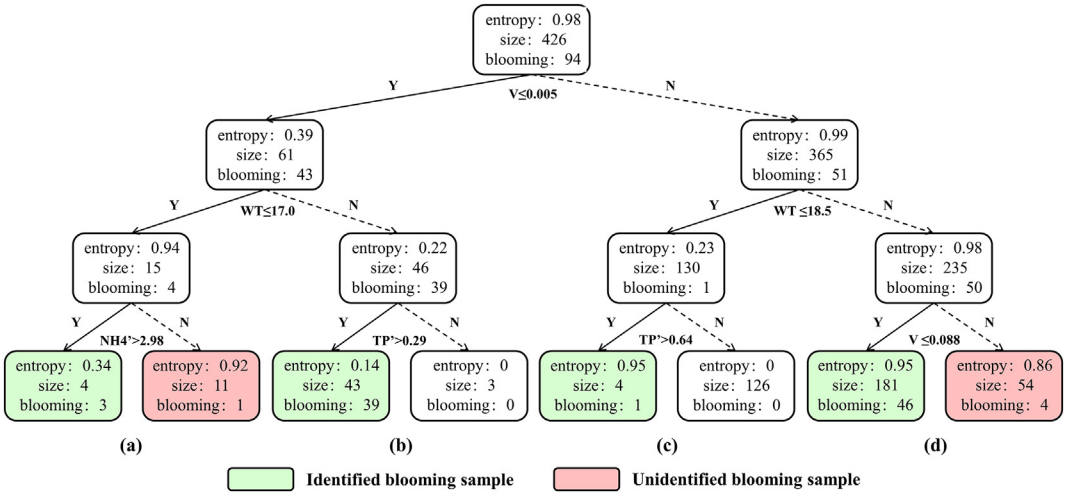


Fig. 9. Results of decision tree.

4.5. What is the limiting threshold of each driving factor?

Chl-a and the six explanatory variables were adopted for the construction of decision tree. Considering that the logarithmic form does not affect the classification result, this study included WT, V, and N–S in original form for convenience of the regulation work. MLR proved that the instrumental variables of nutrients had a stronger explanatory power for algal bloom compared with the original forms; thus, the decision tree involved the instrumental variables of NH4 and TP for classification. A total of 110 samples, including 21 blooming ones, were excluded because each of them contained at least one missing feature. The five-fold cross validation, together with the grid search methods, optimized the three hyperparameters, namely, maximum depth of the tree, minimum size of the leaf, and weight of the blooming samples, as 3, 3, and 5, respectively. The results are shown in Fig. 9. N–S and R7d were not selected as division criteria by the decision tree. Of the 94 blooming samples, only five samples, i.e., 5.3%, were misclassified.

The decision tree first chose V to divide the samples into nearly laminar or turbulent flow conditions. The proportion of blooming samples in the left subtree, i.e., nearly laminar flow condition, was much higher than that in the right subtree. WT continued to divide each subtree into low or high temperature conditions. Among them, subtrees (b) and (d) in the high temperature condition covered 94.7% of the blooming samples. Lastly, subtrees (a)–(d) were further divided by NH4′, TP′, TP′, and V, respectively. Interestingly, the TP′ threshold of subtree (b) was 55% less than the TP′ threshold of subtree (c). Note that subtree (b) was of nearly laminar flow and high temperature, whereas subtree (c) was of turbulent flow and low temperature. Therefore, in subtree (b) whose hydrodynamic condition and climatic factor were conducive to algae growth, a low concentration of phosphorus would suffice to induce blooming. This result qualitatively revealed the local compensation law of different influential factors. Huo et al. (2019) made similar conclusions based on mathematical models and scenario analysis.

By numerically solving Eqs. (18) and (19), the thresholds of NH4′ and TP′ turned into the threshold curves of NH4 and TP as a concave function of DO (Fig. 10), thus restricting algae growth under different V and WT conditions. The two ends of the curves corresponded to the blooming and decay phases, and in conformity, presented stricter nutrient criteria to restrict algae growth. Although subtree (d) did not have a nutrient threshold, it was under the turbulent flow–high temperature condition, just between the conditions in subtrees (b) and (c). Hence, subtree (d) could also be in the TP limited state, with its curve between curves (b) and (c).

4.6. Prospect of multi-factor joint regulation on algal bloom

The above five sequential steps have revealed the complicated mechanisms of algal bloom by quantifying the factor contributions

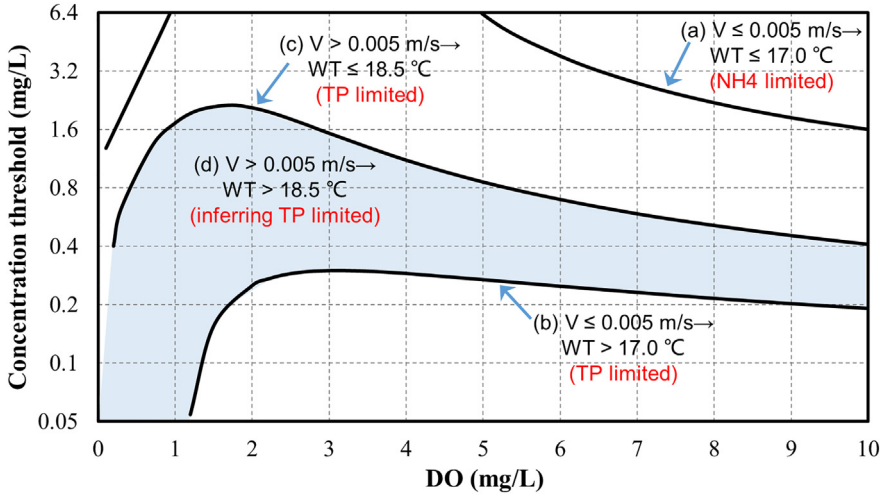


Fig. 10. Threshold curves of nutrients under different circumstances.

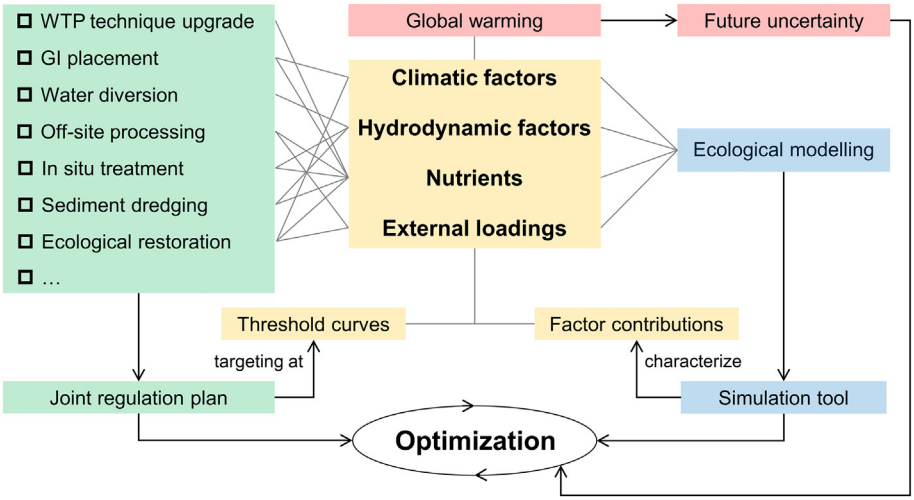


Fig. 11. A guiding roadmap of application of blooming mechanisms in multi-factor joint algal regulation.

and the limiting threshold curves. On this basis, this study called for a multifactor joint algal regulation and drew a guiding roadmap, as shown in Fig. 11. Corporative measures from different sections should be considered to address multiple factors that affect algae growth simultaneously. For example, wastewater treatment plant (WTP) technique upgrading reduces the source-point pollutant loadings, whereas sediment dredging lessens endogenous pollution, as well as facilitates hydrodynamic conditions by decreasing the channel roughness. The limiting threshold curves of the driving factors provide the joint regulation plan with a clear target. An ecological model is an important simulation tool for the optimization of a joint regulation plan. Regardless of whether it is based on machine learning or mechanisms, the model should characterize the local blooming mechanisms well. Moreover, the prominent effect of climatic factors pinpoints the challenge from global warming, suggesting future uncertainty as an important part in the optimization of an adaptive regulation plan.

5. Conclusions

With the help of multi-factor analysis by integrating multiple data mining techniques, this study concludes that the algal blooms in the river network of the Suzhou central urban districts were susceptible to climatic factors (i.e., water temperature and solar radiation), hydrodynamic factor (i.e., flow velocity), nutrients (i.e., phosphorus and nitrogen), and external loadings; this order is sorted from largest to smallest on the basis of contribution ratios. In addition, the limiting threshold curves of each driving factor that might restrict algae growth under different circumstances were identified for the targeted algae regulation. Multifactor features make the blooming mechanisms in gate-controlled urban water bodies different from those in natural water bodies. For many other gate-controlled urban water bodies, the local blooming mechanisms can also be explored by sequentially applying the five techniques. Then, a joint regulation plan could be designed and optimized by associating an ecological model.

In addition to the multi-factor mechanisms, the results also identified global warming as one of the greatest threats to aquatic ecosystems and sketched a compensation law of different factors. Given that direct regulation on water temperature is substantially difficult, stricter regulation strategies on the other factors are necessary under the background of global warming. Limited by the availability of project data, this study only used one-year biweekly data. Nevertheless, abundant monitoring sites compensate for the insufficient time span. Future studies based on long-term monitoring are favorable to further quantify a local compensation law. Adaptive joint regulation on algal bloom under the guidance of local blooming mechanisms is also a promising direction for the improvement of the sensory quality of urban water environments.

CRediT authorship contribution statement

All persons who have made substantial contributions to the work reported in the manuscript have been listed as co-authors and their individual contributions are:

Haifeng Jia: Corresponding author, Resources, Supervision, Project administration, Funding acquisition.

Ke Li: Software, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing.

Te Xu: Conceptualization, Software Formal analysis, Data Curation, Writing - Review & Editing, Visualization.

Jinying Xi: Resources, Monitoring, Investigation, validation.

Zhengjuan Gao: Investigation.

Zhaoxia Sun: Investigation.

Dingkun Yin: Investigation.

Linyuan Leng: Investigation.

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.

Acknowledgement

K. Li and T. Xu contributed equally to the research. The authors thank other members in the group of Prof. J.Y. Xi from Tsinghua University very much for their help in field monitoring. This research was supported by the National Natural Science Foundation of China (No. 41890823, No. 71961137007), and the Major Science and Technology Program for Water Pollution Control and Treatment (No. 2017ZX07205003).

References

Anderson, C.R., Sapiano, M.R.P., Prasad, M.B.K., et al., 2010. Predicting potentially toxigenic Pseudo-nitzschia blooms in the Chesapeake Bay. J. Marine Syst. 83, 127–140.

Bae, S., Seo, D., 2018. Analysis and modeling of algal blooms in the Nakdong River, Korea. Ecol. Model. 372, 53–63.

Beranek, L.L., 1957. Revised criteria for noise in buildings. Noise Control 3, 19.

Breiman, L., Friedman, J., Olshen, R., et al., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Cheng, B., Xia, R., Zhang, Y., et al., 2019. Characterization and causes analysis for algal blooms in large river system. Sustain. Cities Soc. 51, 101707.

Cho, K.H., Kang, J., Ki, S.J., et al., 2009. Determination of the optimal parameters in regression models for the prediction of chlorophyll-a: a case study of the Yeongsan Reservoir, Korea. Sci. Total Environ. 407, 2536–2545.
Conover, W.J., 1999. Practical Nonparametric Statistics. John Wiley & Sons, New York.
Crossman, J., Futter, M., Elliott, J., et al., 2019. Optimizing land management strategies for maximum improvements in lake dissolved oxygen concentrations. Sci. Total Environ. 652, 382–397.
Duda, P.B., Hummel, P.R., Donigian Jr., A.S., et al., 2012. BASINS/HSPF: model use, calibration, and validation. T. ASABE 55, 1523–1547.
Farrar, D.E., Glauber, R.R., 1967. Multicollinearity in regression analysis: the problem revisited. Rev. Econ. Stat. 49, 92–107.
Franklin, J.B., Sathish, T., Vinithkumar, N.V., et al., 2020. A novel approach to predict chlorophyll-a in coastal-marine ecosystems using multiple linear regression and principal component scores. Mar. Pollut. Bull. 152, 110902.
Friedrich, G., Pohlmann, M., 2009. Long-term plankton studies at the lower Rhine Germany. Limnologica 39 (1), 14–39.
Ghatkar, J.G., Singh, R.K., Shanmugam, P., 2019. Classification of algal bloom species from remote sensing data using an extreme gradient boosted decision tree model. Int. J. Remote Sens. 40, 9412–9438.
Gipson, G.L., Freas, W., Kelly, R., et al., 1981. Guideline for Use of City-Specific EKMA in Preparing Ozone SIPs. US EPA, Research Triangle Park, North Carolina.
Harvey, E.L., Menden-Deuer, S., 2011. Avoidance, movement, and mortality: the interactions between a protistan grazer and Heterosigma akashiwo, a harmful algal bloom species. Limnol. Oceanogr. 56, 371–378.
Hilton, J., O'Hare, M., Bowes, M.J., et al., 2006. How green is my river? A new paradigm of eutrophication in rivers. Sci. Total Environ. 365 (1), 66–83.
Hu, W., Zhai, S., Zhu, Z., et al., 2008. Impacts of the Yangtze River water transfer on the restoration of Lake Taihu. Ecol. Eng. 34, 30–49.
Hu, H., Sun, Y., Chen, Z., et al., 2019. Topics and long-term governance model of urban water environment governance. Environ. Eng. 37 (10), 6–15 (in Chinese).
Huo, S., He, Z., Ma, C., et al., 2019. Stricter nutrient criteria are required to mitigate the impact of climate change on harmful cyanobacterial blooms. J. Hydrol. 569, 698–704.
Jarque, C.M., Bera, A.K., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ. Lett. 6 (3), 255–259.
Jiang, X., Gao, G., Zhang, L., et al., 2020. Denitrification and dissimilatory nitrate reduction to ammonium in freshwater lakes of the Eastern Plain, China: influences of organic carbon and algal bloom. Sci. Total Environ. 710, 136303.
Kennedy, P., 1992. A Guide to Econometrics. Blackwell, Oxford.
Klein, R.F., Chow, V.K., 2013. Orthogonalized factors and systematic risk decomposition. The Quarterly Review of Economics and Finance 53, 175–187.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence. vol. 2. Morgan Kaufmann Publishers Inc, Montreal, Quebec, Canada, pp. 1137–1143.
Liu, Y., Wang, Y., Sheng, H., et al., 2014. Quantitative evaluation of lake eutrophication responses under alternative water diversion scenarios: a water quality modeling based statistical analysis approach. Sci. Total Environ. 468-469, 219–227.
Ministry of Housing and Urban-Rural Development of PRC, 2015. Urban Black Odor Water Body Remediation Work Guide.
Norton, B.A., Coutts, A.M., Livesley, S.J., et al., 2015. Planning for cooler cities: a framework to prioritise green infrastructure to mitigate high temperatures in urban landscapes. Landscape Urban Plan 134, 127–138.
Parinet, J., Rodriguez, M.J., Sérodes, J., 2010. Influence of water quality on the presence of off-flavour compounds (geosmin and 2-methylisoborneol). Water Res. 44, 5847–5856.
Park, Y., Cho, K.H., Park, J., et al., 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.
Pearson, K., 1895. Notes on regression and inheritance in the case of two parents. Proc. R. Soc. Lond. 58, 240–242.
Reid, M.K., Spencer, K.L., 2009. Use of principal components analysis (PCA) on estuarine sediment datasets: the effect of data pre-treatment. Environ. Pollut. 157, 2275–2281.
Saltelli, A., Tarantola, S., Campolongo, F., et al., 2004. Sensitivity analysis in practice. J. Am. Stat. Assoc. 101 (473), 398–399.
Shen, J., Qin, Q., Wang, Y., et al., 2019. A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading. Ecol. Model. 398, 44–54.
Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol. 15 (1), 72–101.
Sun, X., Liu, T., Wang, J., 2019. A Bayesian structural model for predicting algal blooms. Int. J. Forecast. 38, 788–802.
Trevor, H., Robert, T., Jerome, F., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.
Wang, J., Yang, C., He, L., et al., 2019. Meteorological factors and water quality changes of Plateau Lake Dianchi in China (1990–2015) and their joint influences on cyanobacterial blooms. Sci. Total Environ. 665, 406–418.
Wooldridge, J.M., 1960. Introductory Econometrics: A Modern Approach. South-Western Cengage Learning, Mason, OH.
Wu, G., Xu, Z., 2011. Prediction of algal blooming using EFDC model: case study in the Daoxiang Lake. Ecol. Model. 222, 1245–1252.
Wu, Y., Yu, Y., Li, X., et al., 2012. Biomass production of a Scenedesmus sp. under phosphorous-starvation cultivation condition. Bioresour. Technol. 112, 193–198.
Xia, R., Zhang, Y., Wang, G., et al., 2019. Multi-factor identification and modelling analyses for managing large river algal blooms. Environ. Pollut. 254 (B), 13056.
Xiao, X., He, J., Huang, H., et al., 2017. A novel single-parameter approach for forecasting algal blooms. Water Res. 108, 222–231.
Xie, X., Qian, X., Hang, Y., et al., 2009. Effect on Chaohu Lake water environment of water transfer from Yangtze River to Chaohu Lake. Res. Environ. Sci. 22, 897–901.
Xu, F., Wang, J., Chen, B., et al., 2011. The variations of exergies and structural exergies along eutrophication gradients in Chinese and Italian lakes. Ecol. Model. 222, 337–350.
Zheng, J., Zhong, C., Deng, C., 2006. Discussion on definition of algal bloom. Water Resources Protection 22 (5), 45–48 (in Chinese).
Zhou, Z., Yu, R., Zhou, M., 2017. Resolving the complex relationship between harmful algal blooms and environmental factors in the coastal waters adjacent to the Changjiang River estuary. Harmful Algae 62, 60–72.
Zhu, X., Dao, G., Tao, Y., et al., 2021. A review on control of harmful algal blooms by plant-derived allelochemicals. J. Hazard. Mater. 401 (accepted).