Sensitivity analysis of external conditions based on the MARS-Sobol method: case study of Tai Lake, China

Ruichen Xu, Yong Pang and Zhibing Hu

ABSTRACT

This study utilized the ECO Lab model calculation samples of Tai Lake, in combination with robust analysis and the GCV test, to promote a faster intelligent application of machine learning and evaluate the MARS machine learning method. The results revealed that this technique can be better trained with small-scale samples, as indicated by the R^2 values of the water quality test results, which were all >0.995. In combination with the Sobol sensitivity analysis method, the contribution degree of the parameterized external conditions as well as the relationship with the water quality were examined, which indicated that TP and TN are primarily related to the external input water quality and flow, while Chl-a is related to inflow (36.42%), TP (26.65%), wind speed (25.89%), temperature (8.38%), thus demonstrating that the governance of Chl-a is more difficult. In general, the accuracy and interpretability of MARS machine learning are more in line with the actual situation, and the use of the Sobol method can save computer calculation time. The results of this research can provide a certain scientific basis for future intelligent management of lake environments.

Key words | cluster analysis, MARS, sensitivity analysis, Sobol, Tai Lake

Ruichen Xu
Yong Pang (corresponding author)
Zhibing Hu
Key Laboratory of Integrated Regulation and Resource Development on Shallow Lakes, Ministry of Education,
Hohai University,
Nanjing 210098,
China
and
College of Environment,
Hohai University,
Nanjing 210098,
China
E-mail: *ypang@hhu.edu.cn*

HIGHLIGHTS

- Introduce a MARS – machine learning method coupled with a Sobol sensitive analysis approach.
- Coupled methods can solve the same problems with less time.
- The declared goal of this research is to provide a certain scientific basis for future intelligent management of lake environments.

INTRODUCTION

After the Tai Lake cyanobacteria crisis of 2007, a more serious cyanobacteria crisis broke out 10 years later, in 2017, indicating that there are still some shortcomings in the original management of the watershed (Zhang *et al.* 2020). However, since management of the watershed is very complex, involving many administrative conflicts and natural areas, determining how to rely on either measured data or model results in order to achieve intelligent management is a difficult problem that needs to be solved (Daneshfaraz *et al.* 2020; Xu *et al.* 2020). The basis of intelligent management is the need to analyze and explain the connection between measured data or model calculations and actual phenomena. Therefore, research on the sensitivity of external conditions and water quality results is very important.

As computer performance has improved, the accuracy of the original water quality model has been greatly enhanced, and the current research on parameter sensitivity has also achieved great breakthroughs (Liang *et al.* 2020; Liu & Ding 2020). Koo *et al.* (2020), for example, used the Sobol method to analyze the SWAT model parameters and

doi: 10.2166/ws.2020.359

nitrate flux sensitivity, eliminating some unimportant parameters and thereby saving time for later calculations (Koo *et al.* 2020). Through the combined use of the Morris and Sobol methods, Garcia *et al.* (2016) developed a new system for selecting important parameters while reducing the amount of calculation (García-Nieto *et al.* 2016). Jiang (Jiang *et al.* 2018) used the global sensitivity analysis (GSA) method to study and analyze the water quality parameters of different areas in Tai Lake, finding that the sensitivity of the parameters affected by the water quality in different regions exhibits certain differences, and the relationship between the growth of algae with temperature and water quality is mutually transformable. When the water quality concentration is sufficient, algae growth has a closer relationship with temperature; otherwise, it is still mainly affected by nutrient salt concentration.

The Sobol method is a global sensitivity analysis approach that can quantify parameter sensitivity and correlation, although it is not widely used in the early stages of analysis due to the huge amount of calculation (Wang 2018; Jordan *et al.* 2020). At present, given the development and application of machine learning methods, the goal of this study was to employ the multivariate adaptive regression splines (Demneh 2019) (MARS)-Sobol method to reduce the workload and improve future management efficiency. The MARS method, which was formally proposed by the American statistician Jerome Friedman in 1991 (Kuter *et al.* 2018), can handle large amounts of data and high dimensionality, and also features the advantages of fast calculation and accurate modeling. Deo (Deo *et al.* 2017) conducted an in-depth study of the relationship between rainfall runoff and regional drought with the help of the MARS method and established a refined model with the dual factors of geography and seasonality via the measurement data of long-term drought and the determination of related parameters, thus providing a successful intelligent application case for later regional drought research. Garcia *et al.* (García-Nieto *et al.* 2016) successfully constructed the MARS-ABS prediction model by combining the MARS method and an artificial bee colony algorithm. Since the estimated coefficients of total phosphorus (TP) and chlorophyll-a (Chl-a) in the lake body were both >0.8 and had good physical, chemical, and biological interpretability, this was a relatively successful and innovative research project.

Based on the measured data of Tai Lake for the past 10 years, this study parameterized the external conditions of the lake such as temperature, wind speed, water entering and exiting, and water quality. With the help of the Latin hypercube sampling (LHS) method, 250 random combinations were selected within the range of values, and the ECO Lab water quality model was then utilized. The TP, TN, and Chl-a were simulated, and the results obtained were trained and tested using the MARS method. A total of 6,000 groups were then sampled according to the Sobol sequence, combined with the Sobol method, to study the sensitivity between external conditions and water quality, and finally passed to the mini-batch k-means clustering algorithm for analysis of the overall conclusion when the

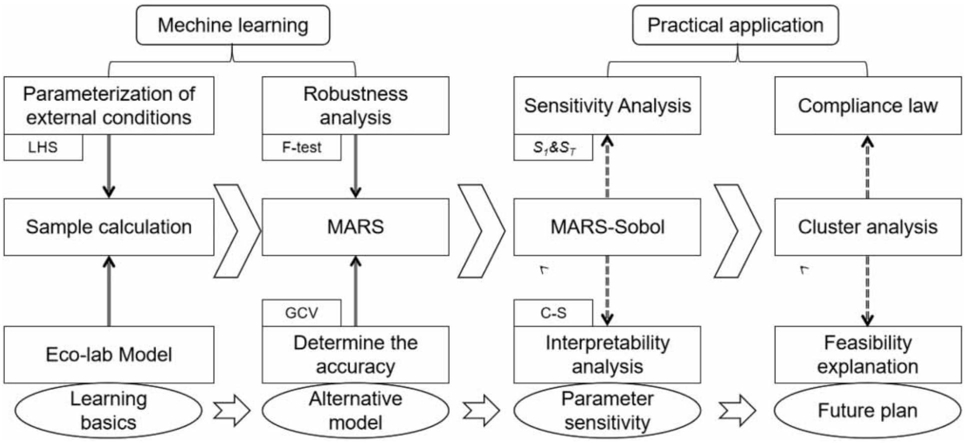


Figure 1 | Research method of this study (mainly including machine learning and practical applications). Definitions: LHS: Latin hypercube sampling; F-test: Joint hypothesis test; GCV: Generalized cross-validation; S_1 and S_T: First-order sensitivity index and total order sensitivity index, respectively; C-S: comparative study.

lake algae reached the standard level. This manuscript is mainly composed of four sections (Figure 1): (1) Research area; (2) Research methods; (3) Results and discussion; and (4) Conclusions and prospects.

STUDY AREA AND METHODS

Study area

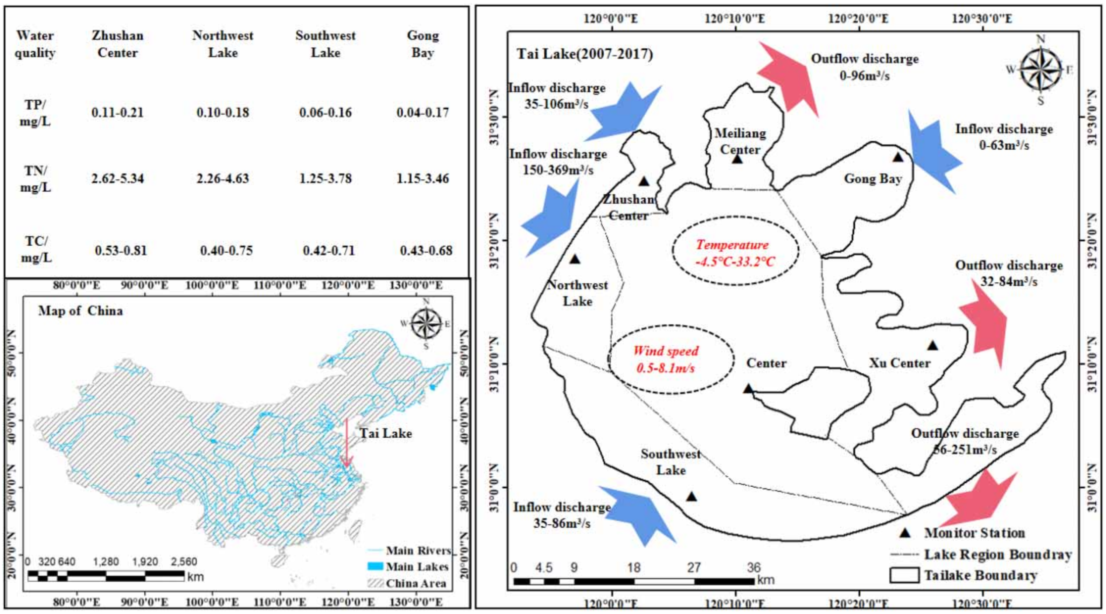
Tai Lake (30°05′–32°08′N, 119°08–122°55′E) is located in the lower reaches of the Yangtze River in China (Yao *et al.* 2020). It has a total area of approximately 2,338 km^2 and is a typical large-scale shallow-water freshwater lake, with an average depth of about 1.9 m. In the past 10 years, the air temperature has ranged from −4.5 °C to 33 °C, the wind speed at a height of 10 m has varied from 0.5 to 8.1 m/s, the average annual rainfall was ~1,222 mm, and the annual average evaporation was 1,051 mm. There are approximately 219 rivers that flow into and out of the surrounding area. The main lake areas can be divided into seven regions: the lake center, Zhushan Bay, Meiliang Bay, Gong Bay, Northwest Lake, Southwest Lake, and East Lake (Tang *et al.* 2020), and there is a corresponding hydrological and water quality synchronization monitoring site (Figure 2).

Based on the relevant measured data for the past 10 years from the Jiangsu Monitoring Center, Taihu Basin Administration of Ministry of Water Resources (http://www.tba.gov.cn/) and the China Meteorological Data Network (http://data.cma.cn/), this study parameterized the amount of water entering and leaving the lake, the water quality, temperature, and wind speed, and linked the LHS method to a total of 250 sampling combinations for external input conditions within the range of 50–150% of the average value. The ECO Lab water quality model simulated and calculated the TP, TN, and Chl-a of the seven main monitoring stations. The actual measured external conditions are illustrated in Figure 2. The obtained simulation results were used as the basic data for later research on alternative models.

Methods

ECO Lab model established

The ECO Lab model is based on a three-dimensional unsteady hydrodynamic model (Waldman *et al.* 2017). It is formed from a hydrodynamic-ecological model coupled with ecological modules, including algae, oxygen cycle, nitrogen cycle, phosphorus cycle, and carbon cycle. The



Water quality	Zhushan Center	Northwest Lake	Southwest Lake	Gong Bay
TP/ mg/L	0.11-0.21	0.10-0.18	0.06-0.16	0.04-0.17
TN/ mg/L	2.62-5.34	2.26-4.63	1.25-3.78	1.15-3.46
TC/ mg/L	0.53-0.81	0.40-0.75	0.42-0.71	0.43-0.68

Figure 2 | Study area and range of the external conditions from 2007 to 2017.

model employed in this study featured a Cartesian coordinate grid of 5,881 rectangular cells, each of which had a length of 300–500 m. To better simulate the lake bottom terrain, σ coordinates were used in the vertical direction, which was divided into three layers on average. Based on hydrostatic continuity and to avoid the pressure gradient error caused by the σ coordinates, the slope of the lake bottom should be <0.33. The model calculation time step was 3,600 s, and the simulation time was 365 d.

Establishing the alternative model based on the MARS method

Multivariate adaptive regression splines (MARS) is a prediction method proposed by the American statistician Jerome Friedman in 1991. This method uses the tensor product of the spline function as the basis function and is divided into three steps: forward process, backward pruning process, and model selection (Metya *et al.* 2017). Furthermore, the generalized cross-validation (GCV) criterion is adopted, and the fitting path is adjusted according to the dynamic characteristics of the fitting object and the interaction between variables, which can fully fit functions of different dimensions.

The MARS model is based on an input variable X with a dimension of n × d and a dependent continuous variable Y with a dimension of n × 1; there is no need to establish preliminary assumptions about X and Y in advance. The model uses piecewise polynomials to construct a smoothly connected basis function (BF) and divides X into different intervals. The piecewise point p is called a node, and its spline curve is a bilateral truncated power function, as expressed by Equation (1):

$$\begin{cases} [+(x-p)]_+^q = \begin{cases} (p-x)^q, & x \geq p \\ 0, & \text{otherwise} \end{cases} \\ [-(x-p)]_+^q = \begin{cases} (p-x)^q, & x \leq p \\ 0, & \text{otherwise} \end{cases} \end{cases} . \tag{1}$$

If N basis functions are considered, the MARS model can be expressed as

$$\hat{y} = \hat{f}_N(x) = c_0 + \sum_{i=1}^{N} c_i BF_i(x), \tag{2}$$

where c_0 is a constant; c_i is the coefficient of the i^{th} basis function BF_i; q ($q > 0$) is the power of the spline function, which determines the smoothness of the spline curve; and x is the data in the input variable $\vec{X}$:

$$GCV = \frac{1}{n} \frac{\sum_{k=1}^{n} (y_k - \hat{y}_k)^2}{\left(1 - \frac{C(M)}{n}\right)^2} \tag{3}$$

In Equation (3), n is the number of predictors; the k^{th}-dependent variable y_k is excluded in order to construct the model; $\hat{y}_k$ is the predicted value of y_k; and $C(M)$ is the amount of function complexity correction.

This study used MATLAB programming tools, and based on the dimension n × d of the input variable $\vec{X}$, as well as the selection of independent variables and nodes, the basis function consisted of several pairs of piecewise polynomials given by Equations (1) and (2). Hence, screening out a suitable basis function was the key to the establishment of the MARS model. The modeling process was divided into three steps.

Step 1: Start with a basic model with only one constant term, and then use the direct truncation process to split the sample function. Taking into consideration the interaction of variables, continue to increase the number of BFs, and improve the accuracy of the model until the residual sum of squares reaches the minimum value or the number of BFs reaches the maximum value, resulting in an overfitting model.

Step 2: Delete the basis functions with small contributions using the backward pruning process, and continuously modify the coefficients of the remaining items. If the accuracy of the model can be guaranteed, delete the redundant BFs; otherwise, keep them.

Step 3: Finally, compare the series of models obtained from the backward pruning process and select the optimal model via the GCV criterion in Equation (3). When the accuracy of the model increases, the GCV value decreases.

Sensitivity analysis of external conditions based on the Sobol method

The Sobol method (Jordan *et al.* 2020) was first proposed by Ilya M. Sobol in 1993. It is a widely used quantitative global sensitivity analysis method. As opposed to the qualitative sensitivity analysis method, the Sobol method can directly provide the sensitivity of the model parameters through the calculation of the sensitivity quantitative index. The specific calculation process is as follows (Figure 3).

Cluster analysis based on the mini-batch k-means method

The mini-batch k-means algorithm (Jia *et al.* 2020) uses a method called mini-batch (batch processing) to calculate the distance between data points. The advantage of the mini-batch technique is that it is not necessary to use all the data samples in the calculation process. Instead, a portion of each sample from different types of samples is extracted to represent their respective types for calculation. Since the calculation sample size is small, the running time will be reduced accordingly, although this type of sampling will inevitably bring about a decrease in accuracy. This technique is employed when the dataset is huge.

In fact, this approach is not only applied to k-means clustering but also widely used in machine learning and deep learning algorithms such as gradient descent and deep networks.

In this study, in order to more clearly determine the characteristics of the influence of different external conditions on algae growth, the clustering method was used to analyze and count the 200 groups of data in the later period. The flow of this algorithm is similar to k-means and consists of the following steps.

Step 1: First, extract part of the dataset, and use the k-means algorithm to construct a model with k clustering points.

Step 2: Continue to extract part of the sample data of the dataset in the training dataset, add it to the model, and assign it to the nearest cluster center point.

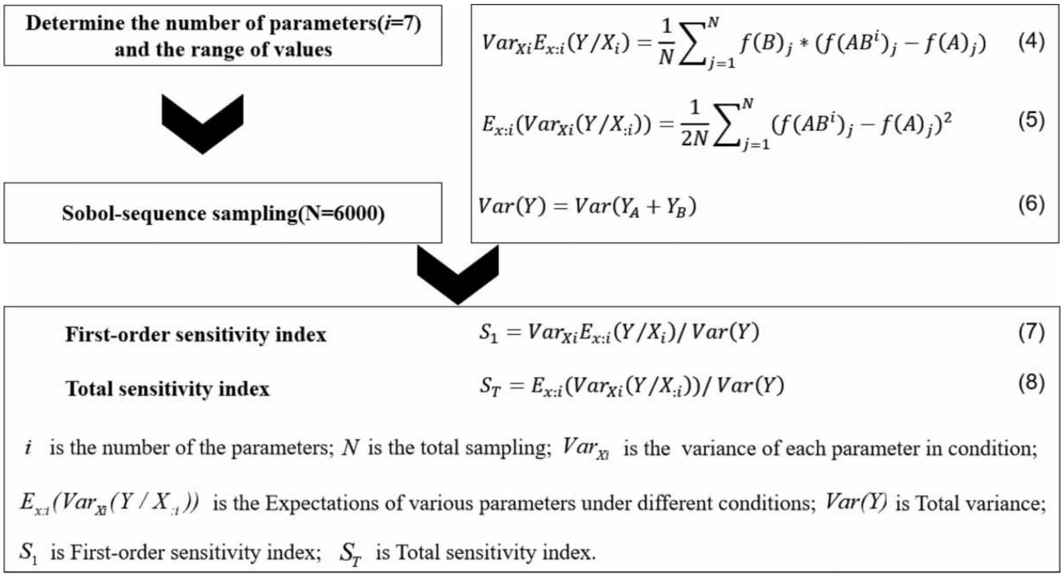
Step 3: Update the center point value of the cluster.

Step 4: Iterate the second and third steps of the loop until the center point is stable or the number of iterations is reached, then stop the calculation operation.

RESULTS AND DISCUSSION

ECO Lab model evaluation

In order to comprehensively consider the credibility of the water quality model, this study compared the water quality simulation results with the actual values, primarily by utilizing the mean relative error (MRE), root mean square error (RMSE), correlation coefficient (R^2), Nash-Sutcliffe efficiency (NSE), and comprehensive prognostic index (CPI) evaluation system to analyze the error and correlation between the measured value M and the simulated value S


Determine the number of parameters(i=7) and the range of values
Sobol-sequence sampling(N=6000)
First-order sensitivity index
Total sensitivity index

$$Var_{Xi}E_{x:i}(Y/X_i) = \frac{1}{N}\sum_{j=1}^{N} f(B)_j * (f(AB^i)_j - f(A)_j) \quad (4)$$

$$E_{x:i}(Var_{Xi}(Y/X_{:i})) = \frac{1}{2N}\sum_{j=1}^{N} (f(AB^i)_j - f(A)_j)^2 \quad (5)$$

$$Var(Y) = Var(Y_A + Y_B) \quad (6)$$

$$S_1 = Var_{Xi}E_{x:i}(Y/X_i)/\,Var(Y) \quad (7)$$

$$S_T = E_{x:i}(Var_{Xi}(Y/X_{:i}))/\,Var(Y) \quad (8)$$

i is the number of the parameters; N is the total sampling; Var_{Xi} is the variance of each parameter in condition; $E_{x:i}(Var_{Xi}(Y/X_{:i}))$ is the Expectations of various parameters under different conditions; $Var(Y)$ is Total variance; S_1 is First-order sensitivity index; S_T is Total sensitivity index.

Figure 3 | Sobol sensitivity analysis method.

(Qiu *et al.* 2020). The specific equations are as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - M_i)^2} \tag{9}$$

$$\mathrm{MRE} = \frac{1}{N}\sum_{i=1}^{N}|S_i - M_i| \tag{10}$$

$$R^2 = \frac{\sum_{i=1}^{N}(S_i - \bar{S})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{N}(S_i - \bar{S})^2 \sum_{i=1}^{N}(M_i - \bar{M})^2}} \tag{11}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(S_i - M_i)^2}{\sum_{i=1}^{N}(S_i - \bar{M})^2} \tag{12}$$

$$\mathrm{CPI}_i = \sum_{i=1}^{4}\alpha_i S_i^k, \tag{13}$$

where N is the total number of simulations; i is the simulation number; S_i is the value of the i^{th} simulation; M_i is the value of the i^{th} measured value; $\bar{M}$ is the simulated average value; and $\bar{S}$ is the measured average value. i takes 1–4; that is, traverses R^2, RMSE, MRE, and NSE; according to the relationship between the four parameter size trends and model accuracy, for RMSE and MRE, the coefficient α_i is -1; for NSE and R^2, the coefficient α_i is 1; and CPI_i is the comprehensive prognostic index of model i. The larger the CPI, the higher the model prediction accuracy.

The results revealed that the water quality simulation of Tai Lake by the ECO Lab model was highly credible, with a comprehensive error <20% (Figure 4) and a $\mathrm{CPI} > 1.5$ (Table 1), indicating that it could better invert the actual water quality in 2017. At the same time, it can be seen from the simulation results that the change trends of Chl-a and TP were relatively close, and the relationship with TN was very weak. DO was almost inversely proportional to temperature. These results are not only consistent with the trends of the actual measurement results but also consistent with the research conclusions of Wang *et al.* (2017). This proves once again that the ECO Lab water quality model

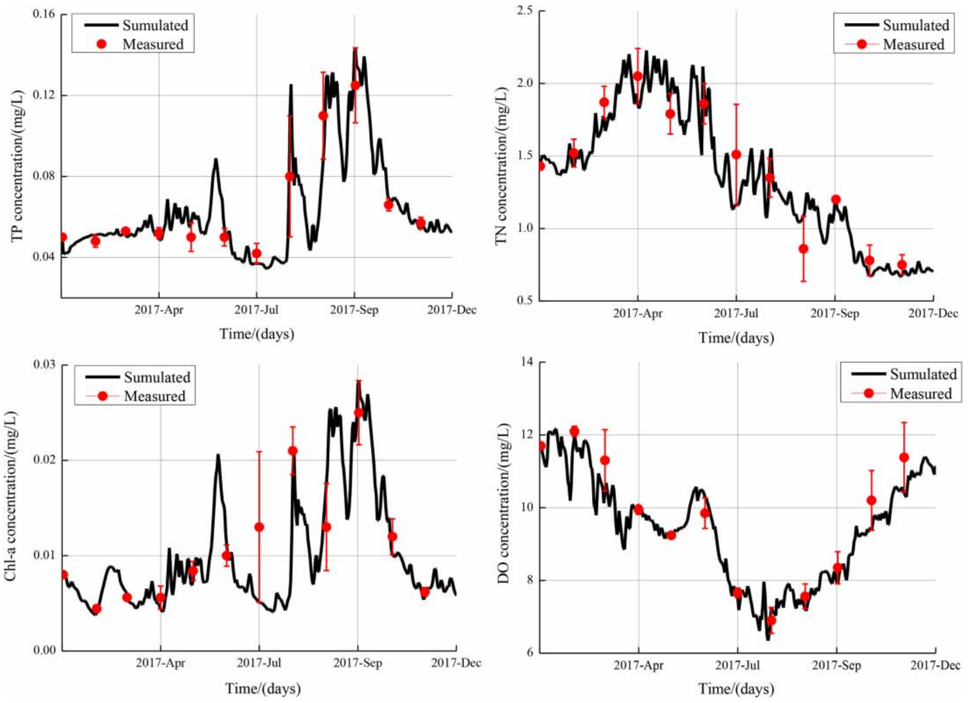


Figure 4 | Monthly error evaluation chart of the four major indicators (TP, TN, Chl-a, and DO) in the seven regions of Tai Lake in 2017.

Table 1 | Evaluation of model calculation results of the four major indicators in the seven regions of Tai Lake

	RMSE	MRE	R^2	NSE	CPI
TP	0.008	10.71%	83.99%	0.93	1.65
TN	0.13	9.98%	87.08%	0.95	1.59
Chl-a	0.002	16.98%	82.64%	0.91	1.56
DO	0.35	4.01%	93.01%	0.97	1.52

can provide basic support for the subsequent mechanism research.

Robustness test and MARS substitution model establishment

For a more scientifically sound analysis of the significant differences in the simulation results, this study used an *F*-test (Sanderson & Windmeijer 2016) to measure the homogeneity of variance. The assumptions were H_0: No significant difference exists in the simulated data; and H_1: A significant difference exists in the simulated data, $\alpha = 0.05$. The specific equations are as follows:

$$S_T^2 = \sum_{i=1}^{s}\sum_{j=1}^{n_j}(X_{ij} - \bar{X})^2 / (n-1) \tag{14}$$

$$F = \frac{S_T^2}{S_T^{2'}} \tag{15}$$

where S_T is the sum of the squared total deviations, s is the number of evaluations, n_j is the number of levels, X_{ij} is the i^{th} evaluation value at the j^{th} level, $\bar{X}$ is the average of all evaluation values, and n is the degrees of freedom; the rejection field is $F \geq F_\alpha\ (n-1,\ n-1)$.

The $N = 25$, 50, 100, 150, and 200 groups of parameter samples were simulated and calculated, after which the Morris index corresponding to each parameter and the *F*-test value between independent samples were obtained (Figure 5). The convergence of the Morris index of the parameter group was closely related to the number of parameter samples; however, when the number of samples was >150 groups, the basic conditions of R^2, p value >0.90, and the GCV criterion could all be satisfied at the same time, indicating that calculating the Morris index required a higher computational workload. A sensitivity analysis can be performed on multi-dimensional parameters for a small sample size, to ensure its high accuracy (Wang *et al.* 2020). It shows that this method can obtain high accuracy with a small number of samples, can save a lot of calculation time in the intermediate process on the basis of the later stage, and also can provide scientific support for the final research goal (Sima *et al.* 2018). In order to ensure the reliability of the study, 200 groups of parameter samples will be utilized for the next simulation study.

Further optimization of the BF and number of nodes revealed that when the TN and TP basis functions were set to 20 and the number of nodes was set to three, the accuracy reached a steady state. In addition, when the basis function of Chl-a was set to 50 and the number of nodes was 3, the accuracy reached a steady state. The calculation accuracies of TP, TN, and Chl-a were all >0.995 (Figure 6). While ensuring the accuracy of the replacement model, the impact on the computer load was minimal, and the calculation time for 1,000 operations was only five minutes. Overall, the computing performance was greatly improved. At the same time, in terms of the basis function, we initially found that the BFs of TP and TN were relatively simple, and the correlation between external conditions was weak, while the BF of Chl-a was more complex, and the correlation between external conditions was also strong, indicating that the MARS model was in the learning stage. Different judgments were made based on the number of parameters and the magnitude of the interaction, which could simultaneously reflect the linear and nonlinear relationships and increase the credibility of the interpretability of the actual situation (Conoscenti *et al.* 2015). At the same time, the study of Huang (Huang *et al.* 2019) also shows that MARS method also has certain advantages in nonlinear prediction, which can avoid common problems such as overfitting, and the large savings in time also provides more possibilities for the later intelligent management of big data. Subsequently, this research will combine the analysis with the Sobol method in order to further analyze the sensitivity of the external conditions, thereby determining whether the interpretability of the alternative model can meet the actual research needs.

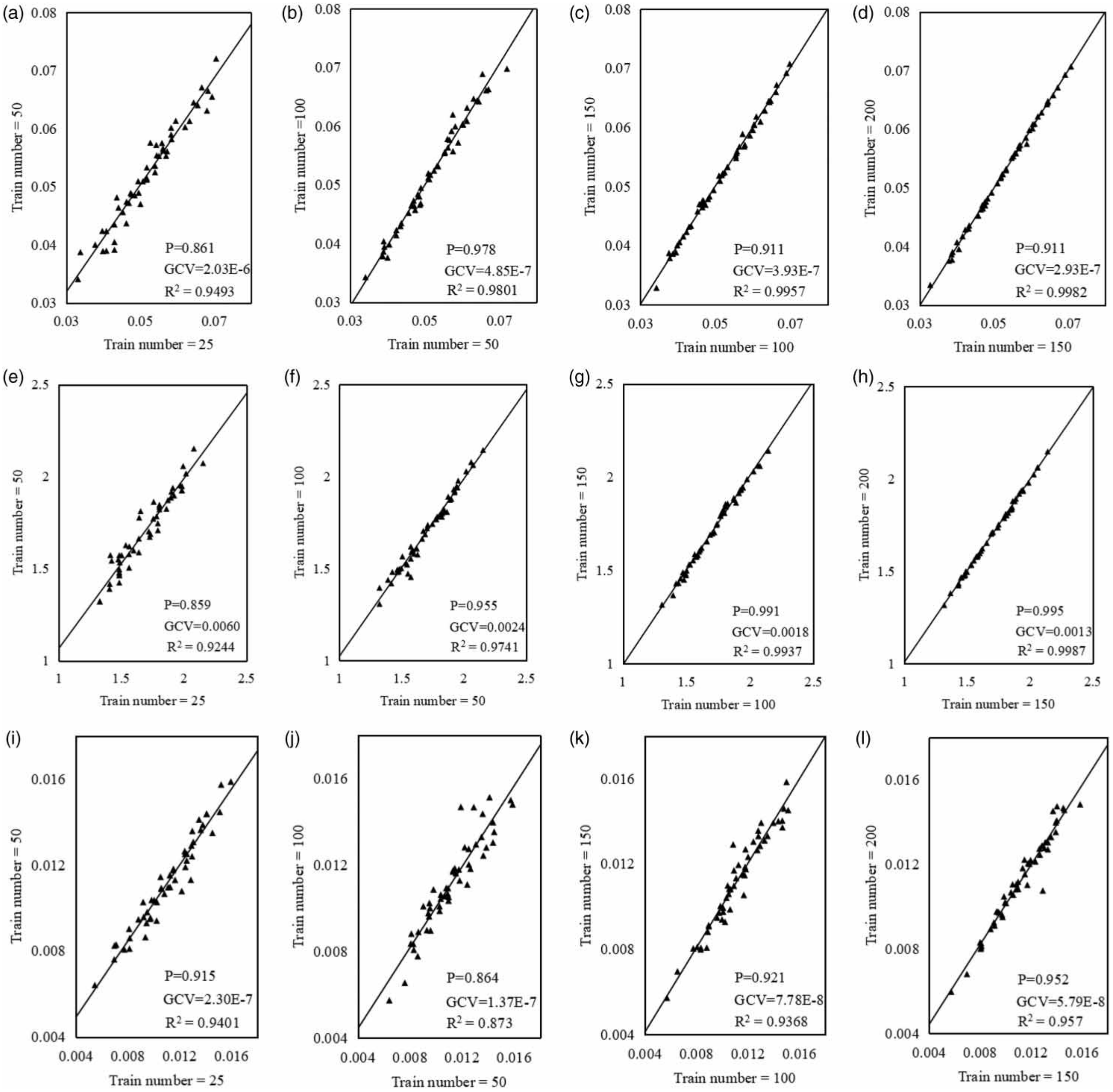


Figure 5 | Robust analysis results of the Morris index with different sample sizes: sample sizes of a–d are $N = 5$ on the X-axis and $N = 10$ on the Y-axis, while sample sizes of e–h are $N = 10$ on the X-axis and $N = 20$ on the Y-axis; (a–d) TP, total phosphorus concentration; (e–h) TN, total nitrogen concentration; (i–l) Chl-a concentration.

Sensitivity analysis of external conditions and thoughts on future governance

Tai Lake is currently in a stage of severe eutrophication, and the outbreak of algae blooms has seriously affected the safety of the drinking water, according to the 'Technical Specification for Water Bloom Remote Sensing and Ground Monitoring and Evaluation' issued by the Ministry of Ecology and Environment of China (http://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/other/qt/202002/t20200213_762889.shtml). For specific requirements, this study set the Chl-a concentration of

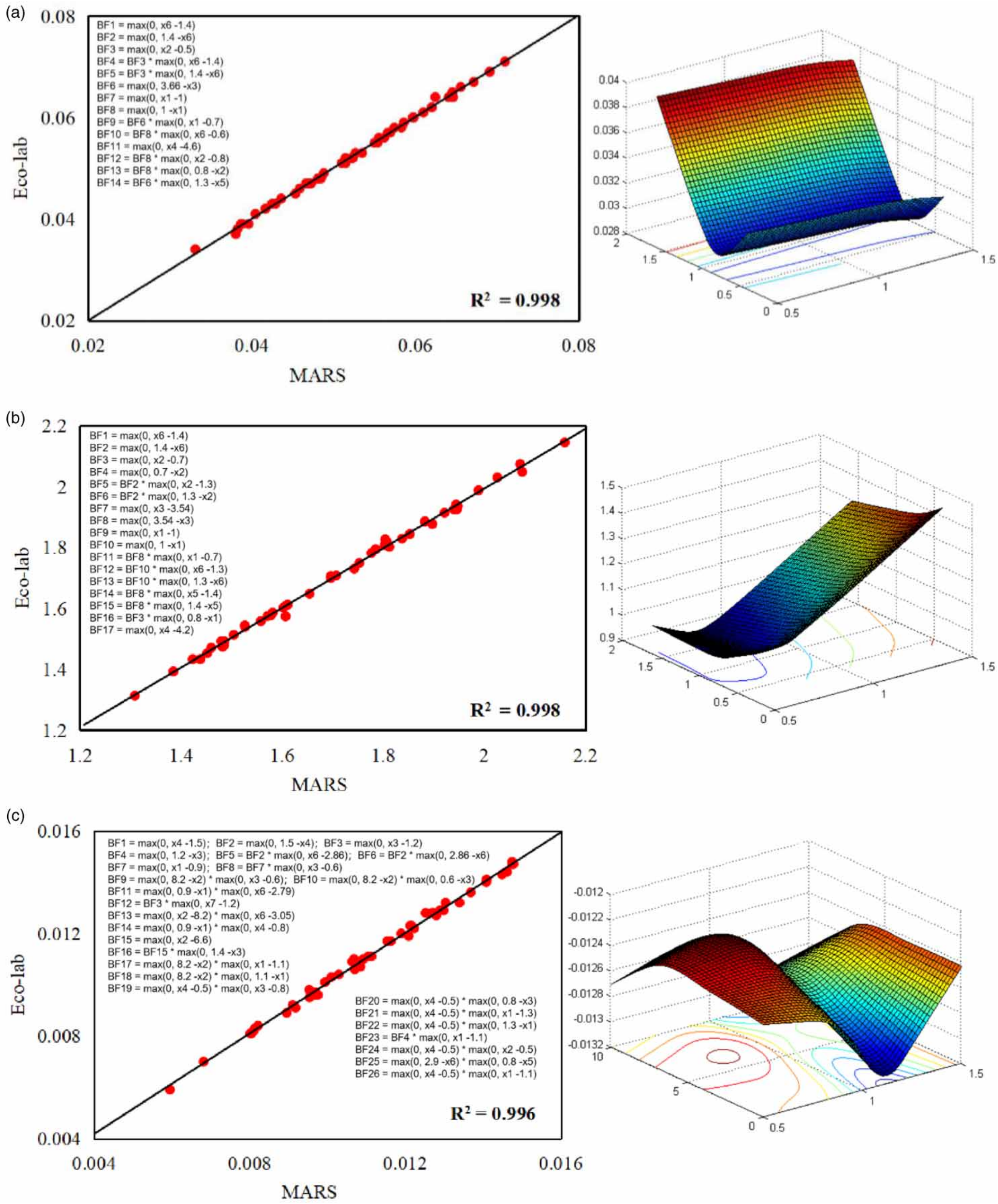


Figure 6 | Replacement model accuracy and MARS three-dimensional calculation results under the best basis function and number of nodes, in which (a) is the calculation result of the TP test samples, (b) is the calculation result of the TN test samples and (c) is the calculation result of the Chl-a test samples.

<10 mg/L as the assessment standard for algae water quality. Based on the cluster analysis of 200 groups of results calculated by the ECO Lab model, it was found that in 66 groups of compliance scenarios, algae growth was mainly affected by three external conditions – wind speed, flow rate, and TP concentration, as shown in Figure 7. This conclusion is similar to that of Jalil *et al.* (2018), indicating that the algae in Tai Lake is mainly affected by hydrodynamics and phosphorus flux. However, since wind speed and temperature are natural factors that are not currently controlled by humans, the future compliance of algae requires dual control of the hydrodynamics and nutrients of Tai Lake by closely combining flow and phosphorus flux. The overall stable compliance scenario indicates that the total phosphorus entering the lake has already exceeded the acceptable threshold. The amount of TP needs to be reduced by 30–50% in order to prevent large-scale outbreaks of algae, which is basically consistent with the research conclusions (Xu *et al.* 2016; Liu *et al.* 2020).

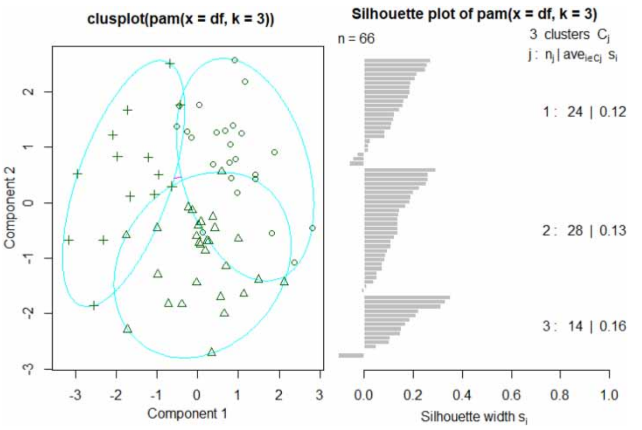
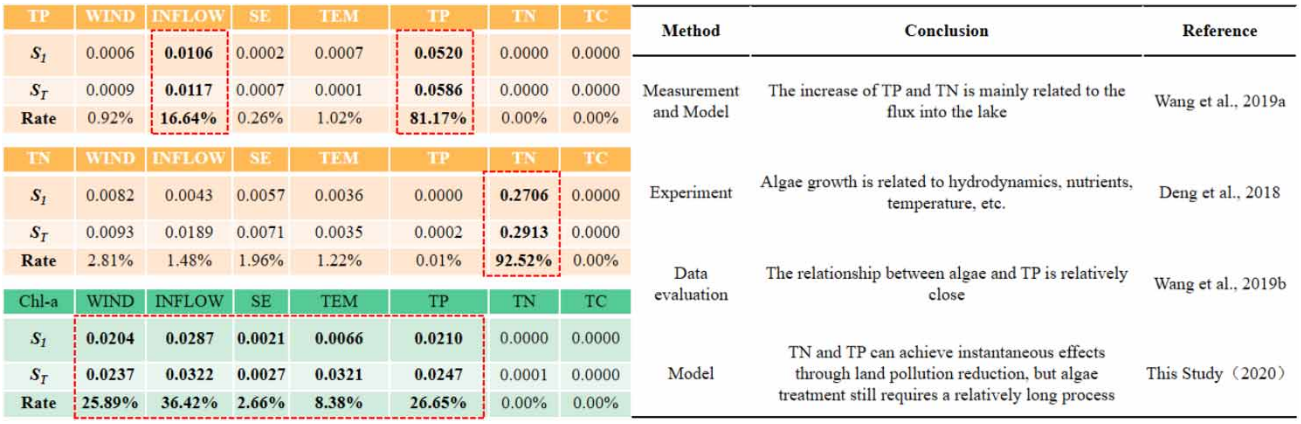


Figure 7 | Study on the comparison and laws of the schemes for reaching or not meeting the standards based on the clustering method.

Studies have shown that TP and TN are affected in far simpler ways than algae. TP and TN are directly affected by external input pollution flux, the influences of both reaching more than 90% (Figure 8). This is similar to conclusions of the experimental research on Tai Lake performed by many scientists (Deng *et al.* 2018; Wang *et al.* 2019a, 2019b), indicating that under the premise of maintaining a dynamic balance of the internal sources of nutrients, TP and TN are still mainly affected by external flux input. This influence is not closely related to meteorological conditions such as temperature and wind, although the Chl-a concentration, which reflects the growth of algae, is greatly affected by external comprehensive factors. This is because algae are a type of living organism and are therefore more significantly affected by the water environment of eutrophic lakes. In particular, under conditions of suitable temperature and nutrients, they are primarily affected by wind speed and flow velocity, the influence of which can reach more than 60%. In addition, Chl-a is closely related to TP (Wu *et al.* 2019) but has almost no relationship with TN, indicating that TN in Taihu Lake is not a factor that limits the growth of Chl-a. Therefore, more attention should be paid to the dynamic changes of TP in the future. Moreover, in the Sobol calculation results, we found parameters with large first-order sensitivity, and the total-order sensitivity



TP	WIND	INFLOW	SE	TEM	TP	TN	TC
S_1	0.0006	**0.0106**	0.0002	0.0007	**0.0520**	0.0000	0.0000
S_T	0.0009	**0.0117**	0.0007	0.0001	**0.0586**	0.0000	0.0000
Rate	0.92%	**16.64%**	0.26%	1.02%	**81.17%**	0.00%	0.00%

TN	WIND	INFLOW	SE	TEM	TP	TN	TC
S_1	0.0082	0.0043	0.0057	0.0036	0.0000	**0.2706**	0.0000
S_T	0.0093	0.0189	0.0071	0.0035	0.0002	**0.2913**	0.0000
Rate	2.81%	1.48%	1.96%	1.22%	0.01%	**92.52%**	0.00%

Chl-a	WIND	INFLOW	SE	TEM	TP	TN	TC
S_1	**0.0204**	**0.0287**	**0.0021**	**0.0066**	**0.0210**	0.0000	0.0000
S_T	**0.0237**	**0.0322**	**0.0027**	**0.0321**	**0.0247**	0.0001	0.0000
Rate	**25.89%**	**36.42%**	**2.66%**	**8.38%**	**26.65%**	0.00%	0.00%

Method	Conclusion	Reference
Measurement and Model	The increase of TP and TN is mainly related to the flux into the lake	Wang et al., 2019a
Experiment	Algae growth is related to hydrodynamics, nutrients, temperature, etc.	Deng et al., 2018
Data evaluation	The relationship between algae and TP is relatively close	Wang et al., 2019b
Model	TN and TP can achieve instantaneous effects through land pollution reduction, but algae treatment still requires a relatively long process	This Study（2020）

Figure 8 | Sensitivity analysis of external conditions based on the MARS-Sobol method and comparison with relevant research carried out on Tai Lake in recent years.

was also large, indicating that this parameter has a strong correlation with other parameters (Jaxa-Rozen & Kwakkel 2018).

According to the above research results, the MARS-Sobol method used in this study has good interpretability when calculating multiple parameters and multiple dimensions (Zhang *et al.* 2015). In addition, although it takes 45 min to calculate a Sobol sample alone, when combined with the MARS method, it only takes about five minutes to calculate 1,000 groups of samples, thus greatly reducing the calculation load of the workstation without changing the actual situation. Furthermore, the MARS method requires fewer samples than the back propagation artificial neural network (BP-ANN) method (Sun *et al.* 2019). In the future, it is strongly recommended that more in-depth research be conducted on more complex models or measured data.

CONCLUSIONS AND PROSPECTS

(1) The ECO Lab model was found to fully reflect the actual situation in the water quality simulation of Tai Lake, and the CPI was >1.5. The learning accuracy of the MARS method was related to the number of training samples, basis functions, number of nodes, and other factors. In the later stage, appropriate parameter sizes should be selected according to the specific situation in order to adjust the learning accuracy of the MARS method. In general, the MARS method requires a small number of training samples, and the learning accuracy of 150 groups can reach more than 0.990. At the same time, this research proves that it is suitable for the learning of high-latitude parameter groups.

(2) The MARS-Sobol method has the dual analysis functions of sensitivity and correlation. The results revealed that the factors with strong sensitivity are strongly correlated; different external conditions will have a significant impact on the water quality of Tai Lake. When the Chl-a is the output, the sensitivity ranking is inflow (36.42%) >TP (26.65%) > wind speed (25.89%) > temperature (8.38%). When TP is the output indicator, the sensitivity ranking is TP (81.17%) > inflow (16.64%); when TN is used as the output index, the sensitivity is closely related to TN concentration (92.52%). In general, the impact of a single external condition on Chl-a is less than that on TP and TN, and the growth of algae depends largely on the magnitude of the hydrodynamic force. Therefore, the government should devote much more time to treating the algae problem.

(3) The clustering results of 200 groups of data revealed that the growth of algae is mainly affected by three external conditions: wind speed, flow rate, and TP concentration. In other words, algae are more significantly affected by hydrodynamic forces and phosphorus flux in Tai Lake. Since the temperature and water level cannot be controlled properly by humans, this study suggests that the water flow and phosphorus flux should be the dual controls of the hydrodynamic and nutrient levels in the future.

ACKNOWLEDGEMENTS

The authors thank the Chinese National Science Foundation (Grant No. 51879070). This work was supported by the Fundamental Research Funds for the Central Universities, the World-class Universities (Disciplines), and the Characteristic Development Guidance Funds for the Central Universities. This research was also funded by the Major Science and Technology Program for Water Pollution Control and Treatment of China (Grant No. 2018ZX07208003). We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Conoscenti, C., Ciaccio, M., Caraballo-Arias, N. A., Gómez-Gutiérrez, Á., Rotigliano, E. & Agnesi, V. 2015 Assessment of susceptibility to earth-flow landslide using logistic regression and multivariate adaptive regression splines: a case of the Belice River basin (western Sicily, Italy). *Geomorphology* **242**, 49–64.

Daneshfaraz, R., Bagherzadeh, M., Esmaeeli, R., Norouzi, R. & Abraham, J. 2020 Study of the performance of support vector machine for predicting vertical drop hydraulic parameters in the presence of dual horizontal screens. *Water Supply* (in press), doi: 10.2166/ws.2020.279.

Demneh, S. E. R. K. 2019 A comparison of artificial intelligence models for the estimation of daily suspended sediment load: a case study on the Telar and Kasilian rivers in Iran. *Water Supply* **19** (1), 165–178.

Deng, J., Zhang, W., Qin, B., Zhang, Y., Paerl, H. W. & Salmaso, N. 2018 Effects of climatically-modulated changes in solar radiation and wind speed on spring phytoplankton community dynamics in Lake Taihu, China. *PloS One* **13** (10), 1–16.

Deo, R. C., Kisi, O. & Singh, V. P. 2017 Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. *Atmospheric Research* **184**, 149–175.

García-Nieto, P. J., García-Gonzalo, E., Alonso Fernández, J. R. & Díaz Muñiz, C. 2016 Using evolutionary multivariate adaptive regression splines approach to evaluate the eutrophication in the Pozón de la Dolores lake (Northern Spain). *Ecological Engineering* **94**, 136–151.

Huang, H., Ji, X., Xia, F., Huang, S., Shang, X., Chen, H., Zhang, M., Dahlgren, R. A. & Mei, K. 2019 Multivariate adaptive regression splines for estimating riverine constituent concentrations. *Hydrological Processes* **34** (5), 1213–1227.

Jalil, A., Li, Y., Du, W., Wang, W., Wang, J., Gao, X., Khan, H. O. S., Pan, B. & Acharya, K. 2018 The role of wind field induced flow velocities in destratification and hypoxia reduction at Meiling Bay of large shallow Lake Taihu, China. *Environmental Pollution* **232**, 591–602.

Jaxa-Rozen, M. & Kwakkel, J. 2018 Tree-based ensemble methods for sensitivity analysis of environmental models: a performance comparison with Sobol and Morris techniques. *Environmental Modelling & Software* **107**, 245–266.

Jia, D., Liu, H., Zhang, J., Gong, B., Pei, X., Wang, Q. & Yang, Q. 2020 Data-driven optimization for fine water injection in a mature oil field. *Petroleum Exploration and Development* **47** (3), 674–682.

Jiang, L., Li, Y., Zhao, X., Tillotson, M. R., Wang, W., Zhang, S., Sarpong, L., Asmaa, Q. & Pan, B. 2018 Parameter uncertainty and sensitivity analysis of water quality model in Lake Taihu, China. *Ecological Modelling* **375**, 1–12.

Jordan, M., Millinger, M. & Thrän, D. 2020 Robust bioenergy technologies for the German heat transition: a novel approach combining optimization modeling with Sobol' sensitivity analysis. *Applied Energy* **262**, 114534.

Koo, H., Chen, M., Jakeman, A. J. & Zhang, F. 2020 A global sensitivity analysis approach for identifying critical sources of uncertainty in non-identifiable, spatially distributed environmental models: a holistic analysis applied to SWAT for input datasets and model parameters. *Environmental Modelling & Software* **127**, 104676.

Kuter, S., Akyurek, Z. & Weber, G.-W. 2018 Retrieval of fractional snow covered area from MODIS data by multivariate adaptive regression splines. *Remote Sensing of Environment* **205**, 236–252.

Liang, H., Gao, S. & Hu, K. 2020 Global sensitivity and uncertainty analysis of the dynamic simulation of crop N uptake by using various N dilution curve approaches. *European Journal of Agronomy* **116**, 126044.

Liu, W. & Ding, L. 2020 Global sensitivity analysis of influential parameters for excavation stability of metro tunnel. *Automation in Construction* **113**, 103080.

Liu, L., Dong, Y., Kong, M., Zhou, J., Zhao, H., Wang, Y., Zhang, M. & Wang, Z. 2020 Towards the comprehensive water quality control in Lake Taihu: correlating chlorphyll a and water quality parameters with generalized additive model. *Science of the Total Environment* **705**, 135993.

Metya, S., Mukhopadhyay, T., Adhikari, S. & Bhattacharya, G. 2017 System reliability analysis of soil slopes with general slip surfaces using multivariate adaptive regression splines. *Computers and Geotechnics* **87**, 212–228.

Qiu, R., Wang, Y., Wang, D., Qiu, W., Wu, J. & Tao, Y. 2020 Water temperature forecasting based on modified artificial neural network methods: two cases of the Yangtze River. *Science of the Total Environment* **737**, 139729.

Sanderson, E. & Windmeijer, F. 2016 A weak instrument [Formula: see text]-test in linear IV models with multiple endogenous variables. *Journal of Econometrics* **190** (2), 212–221.

Sima, N. Q., Harmel, R. D., Fang, Q. X., Ma, L. & Andales, A. A. 2018 A modified F-test for evaluating model performance by including both experimental and simulation uncertainties. *Environmental Modelling & Software* **104**, 236–248.

Sun, Q., Zhang, M. & Yang, P. 2019 Combination of LF-NMR and BP-ANN to monitor water states of typical fruits and vegetables during microwave vacuum drying. *LWT* **116**, 108548.

Tang, C., Li, Y., He, C. & Acharya, K. 2020 Dynamic behavior of sediment resuspension and nutrients release in the shallow and wind-exposed Meiliang Bay of Lake Taihu. *Science of the Total Environment* **708**, 135131.

Waldman, S., Bastón, S., Nemalidinne, R., Chatzirodou, A., Venugopal, V. & Side, J. 2017 Implementation of tidal turbines in MIKE 3 and Delft3D models of Pentland Firth & Orkney Waters. *Ocean & Coastal Management* **147**, 21–36.

Wang, Y. W. Z. H. L. 2018 Sensitivity analysis of the Chaohu Lake eutrophication model with a new index based on the Morris method. *Water Supply* **18** (4), 1375–1387.

Wang, J., Zhao, Q., Pang, Y., Li, Y., Yu, Z. & Wang, Y. 2017 Dynamic simulation of sediment resuspension and its effect on water quality in Lake Taihu, China. *Water Science and Technology: Water Supply* **17** (5), 1335–1346.

Wang, M., Strokal, M., Burek, P., Kroeze, C., Ma, L. & Janssen, A. B. G. 2019a Excess nutrient loads to Lake Taihu: opportunities for nutrient reduction. *Science of the Total Environment* **664**, 865–873.

Wang, J., Fu, Z., Qiao, H. & Liu, F. 2019b Assessment of eutrophication and water quality in the estuarine area of Lake Wuli, Lake Taihu, China. *Science of the Total Environment* **650** (Pt 1), 1392–1402.

Wang, C., Peng, M. & Xia, G. 2020 Sensitivity analysis based on Morris method of passive system performance under ocean conditions. *Annals of Nuclear Energy* **137**, 107067.

Wu, T., Qin, B., Brookes, J. D., Yan, W., Ji, X. & Feng, J. 2019 Spatial distribution of sediment nitrogen and phosphorus in Lake Taihu from a hydrodynamics-induced transport perspective. *Science of the Total Environment* **650** (Pt 1), 1554–1565.

Xu, H., Paerl, H. W., Zhu, G., Qin, B., Hall, N. S. & Zhu, M. 2016 Long-term nutrient trends and harmful cyanobacterial bloom potential in hypertrophic Lake Taihu, China. *Hydrobiologia* **787** (1), 229–242.

Xu, R., Pang, Y., Hu, Z., Zhu, T. & Kaisam, J. P. 2020 Influence of water diversion on spatial and temporal distribution of flow field and total phosphorus (TP) concentration field in Taihu Lake. *Water Supply* **20** (3), 1059–1071.

Yao, X., Zhang, Y., Zhang, L., Zhu, G., Qin, B., Zhou, Y. & Xue, J. 2020 Emerging role of dissolved organic nitrogen in supporting algal bloom persistence in Lake Taihu, China: emphasis on internal transformations. *Science of the Total Environment* **736**, 139497.

Zhang, W., Goh, A. T. C., Zhang, Y., Chen, Y. & Xiao, Y. 2015 Assessment of soil liquefaction based on capacity energy concept and multivariate adaptive regression splines. *Engineering Geology* **188**, 29–37.

Zhang, T., Qin, M., Wei, C., Li, D., Lu, X. & Zhang, L. 2020 Suspended particles phoD alkaline phosphatase gene diversity in large shallow eutrophic Lake Taihu. *Science of the Total Environment* **728**, 138615.

First received 20 October 2020; accepted in revised form 27 November 2020. Available online 11 December 2020