Simulation of Chlorophyll a Concentration in Donghu Lake Based on ABC-SVM and Water Quality Indexes

Xiaodong Tang
Huazhong University of Science and Technology

Mutao Huang (✉ huangmutao@hust.edu.cn)
Huazhong University of Science and Technology

Article

Keywords: Chlorophyll a concentration, Artificial Bee Colony algorithm, Support Vector Machine, Donghu Lake

Posted Date: May 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1654109/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Simulation of Chlorophyll a Concentration in Donghu Lake Based on ABC-SVM and Water Quality Indexes

Xiaodong Tang 1, Mutao Huang 2, *

1 School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

2 School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

* Corresponding Author: huangmutao@hust.edu.cn
Abstract
Chlorophyll a concentration is an important index of eutrophication, and simulation of chlorophyll a concentration is of great significance to the monitoring and control of lake eutrophication. The aim of the study is to explore a new method to build an effective model to simulate chlorophyll a concentration in Donghu Lake. Based on measured data of water quality indexes, ABC-SVM model for simulating chlorophyll a concentration is built by using Artificial Bee Colony (ABC) algorithm to improve Support Vector Machine (SVM). Moreover, accuracy, sensitivity and universality of the model are analysed. Modeling results show that the ABC-SVM model has high accuracy and good simulation effect, $R^2$ and RMSE in testing process is 0.96 and 10.40 μg/L, ABC optimization increases $R^2$ by 0.04 and reduces RMSE by 4.12 μg/L compared with SVM model. Sensitivity analysis results demonstrate chlorophyll a concentration is more sensitive to TP than other water quality indexes. In addition, universality analysis results reveal that ABC-SVM model has good universality and can be used to simulate the chlorophyll a concentration of Donghu Lake at different times. Overall, we have built an efficient simulation model, which provides a new idea and method for chlorophyll a concentration simulation.

Keywords: Chlorophyll a concentration; Artificial Bee Colony algorithm; Support Vector Machine; Donghu Lake

Introduction
Lake eutrophication is a global environmental problem, 30% to 40% of lakes and reservoirs worldwide are affected to varying degrees of eutrophication at the end of 20th century. Chlorophyll a concentration, as an important index of phytoplankton biomass as well as eutrophication, is of great significance for the study of primary productivity, eutrophication and algal bloom. Therefore, the simulation of chlorophyll a concentration in lake is very important for environmental protection and social development.

The simulation of chlorophyll a concentration mainly includes remote sensing inversion and water quality modeling. Remote sensing inversion has gradually developed into an indispensable method for chlorophyll a simulation because it is timely, widely covered and costs relatively lower. It is a method to simulate chlorophyll a concentration by building an inversion model based on chlorophyll a concentration measured data and remote sensing images. At the same time, the introduction of machine learning algorithm makes remote sensing inversion more diverse and effective. Philipp et al. applied a one-dimensional convolutional neural network for Estimating Chlorophyll a Concentration of Inland Water Bodies and the real-world SpecWa dataset to estimate the chlorophyll a values of unknown inland water bodies successfully. Tang et al. adopted BPNN, ELM, SVM and multiple datasets to build chlorophyll a concentration inversion models of Donghu Lake and
compared the accuracy of each model\textsuperscript{15}. Yu et al. combined Wavelet Domain Threshold Denoising (WDTD), Wavelet Mean Fusion (WMF) and Long-Short Term Memory (LSTM) to a WDTD-LSTM-WMF long-term prediction model and used it to simulate the historical change process of chlorophyll a concentration in Dianchi Lake and predicted the future trend of chlorophyll a concentration\textsuperscript{16}. However, remote sensing inversion has some limitations in the simulation of chlorophyll a concentration in lake. Cloud and a longer revisit cycle limit the images acquisition and application in time series analysis\textsuperscript{17,18,19}. And, remote sensing of medium to small lakes requires moderate (~ 300 m) to high (10 ~ 30 m) spatial resolution sensors, the resolution of commonly used and easily available satellite images usually cannot meet the requirements for water quality inversion of small lake\textsuperscript{21,22}.

In addition, inversion model has poor universality and it is only applicable to a specific area as inland water body has the different characteristics of regional and seasonal\textsuperscript{23,24,25}. Compared with remote sensing inversion, water quality modeling is more comprehensive, and the modeling and simulation process is more complex. It considers the physical, chemical and biological processes of substances in lake, and builds a comprehensive dynamic simulation system to simulate the migration and transformation process of water quality indexes\textsuperscript{26}. Water quality numerical models such as EFDC\textsuperscript{27}, WASP\textsuperscript{28,29}, MIKE\textsuperscript{30,31} and Delft-3D\textsuperscript{32} are widely used in the study of lake eutrophication, and large number of research achievements have been made in simulation of chlorophyll a concentration in lake\textsuperscript{33}. Nevertheless, there are some difficulties that need to be solved when using numerical model to simulate chlorophyll a concentration. In the modeling process, sufficient data and suitable model are the basis of model construction, and it is difficult and cumbersome to obtain appropriate parameters as the complicated hydrodynamic conditions and growth/extinction mechanism of chlorophyll a\textsuperscript{34}. In the simulation process, the error of observation data will directly affect the accuracy of simulation results, and systematic errors cannot be avoided. Therefore, it is necessary to explore new methods to simulate chlorophyll a concentration in lakes.

Although the factors leading to lake eutrophication differ from one lake to another as the variable environment\textsuperscript{35}, the same thing is that chlorophyll a concentration is the important index of lake eutrophication. Studies have shown that lake eutrophication is mainly caused by excessive nutrients runoff to the lake and serious eutrophication is easy to lead to algal bloom\textsuperscript{36}. In addition, the growth and reproduction of algae are affected by light, water temperature (T) and pH\textsuperscript{37,38}. At the same time, the rapid growth of algae will lead to some changes in lake water quality, such as the decrease of transparency and dissolved oxygen (DO)\textsuperscript{39}. Furthermore, there is a positive correlation between living algae concentration and chlorophyll a concentration in lake. Therefore, chlorophyll a concentration in lake has a specific biochemical relationship with light, T, total nitrogen (TN), total phosphorus (TP) and other physical and chemical indexes\textsuperscript{40}.
In this paper, we attempt to use artificial bee colony algorithm to optimize support vector machine, explore the relationship between chlorophyll a concentration and water quality indexes (DO, T, TN, TP), and build ABC-SVM model to simulate the chlorophyll a concentration in Donghu Lake. The contributions of this work are presented as follows: (a) use ABC to optimize parameters of SVM and build ABC-SVM model for simulation of chlorophyll a concentration in Donghu Lake based on the data of DO, T, TN and TP; (b) use ABC-SVM model to simulate chlorophyll a concentration and verify the accuracy of the model; (c) analyse the sensitivity of chlorophyll a concentration to DO, T, TN and TP in ABC-SVM model simulation; (d) analyse the universality of the model at different times.

**Materials and methods**

**Study area and data acquisition**

Donghu Lake (30° 22′ -30° 40′ N, 114° 09′ -114° 39′ E) is located in the northeast of Wuhan, Hubei Province. It is a typical shallow lake in the middle reaches of the Yangtze River and is one of the largest urban lakes in China. It is mainly composed of Guozheng Lake, Tangling Lake, Miaohu Lake, Tuanhu Lake, Houhu Lake, Lingjiao Lake and Shuiguo Lake. Donghu Lake plays an important role in human society and has many functions, such as domestic water, industrial water, irrigation, flood prevention and storage, aquaculture, natural landscape, nevertheless, it is polluted by all kinds of production wastewater and domestic sewage discharged by human activities.

Figure 1. Distribution of sampling points in Donghu Lake.

Sampling points are set up in Donghu Lake on November 15, 2017 (20171115), December 17, 2017 (20171217), March 26, 2018 (20180326) and October 26, 2018 (20181026). The sampling points are presented in figure 1, the
water samples are taken from 0.5 meters underwater at every sampling point and put into labelled water sample bottles. The coordinates of each sampling point are recorded by portable GPS locator. DO, T of each water sample are measured in site by portable water quality detector, and TN, TP are detected by spectrophotometry with water samples sent to laboratory. The measured data range of Donghu Lake water samples at different times is shown in table 1.

<table>
<thead>
<tr>
<th>Sampling Time</th>
<th>Sampling Points</th>
<th>Chl-a (μg/L)</th>
<th>DO (mg/L)</th>
<th>T (℃)</th>
<th>TN (mg/L)</th>
<th>TP (mg/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20171115</td>
<td>29</td>
<td>6.91~50.65</td>
<td>8.82~10.52</td>
<td>15.0~16.0</td>
<td>0.53~1.46</td>
<td>0.09~0.25</td>
</tr>
<tr>
<td>20171217</td>
<td>42</td>
<td>5.25~157.72</td>
<td>10.03~12.31</td>
<td>6.0~8.5</td>
<td>0.43~7.23</td>
<td>0.02~0.47</td>
</tr>
<tr>
<td>20180326</td>
<td>43</td>
<td>3.57~163.95</td>
<td>2.75~6.85</td>
<td>16.5~21.0</td>
<td>0.36~6.89</td>
<td>0.01~0.45</td>
</tr>
<tr>
<td>20181026</td>
<td>43</td>
<td>17.46~140.93</td>
<td>6.57~11.22</td>
<td>18.8~21.3</td>
<td>0.44~5.42</td>
<td>0.04~0.32</td>
</tr>
</tbody>
</table>

The measured data range of Donghu Lake water samples at different times.

The artificial bee colony algorithm

The artificial bee colony (ABC) algorithm was proposed by Karaboga in 2005⁴², which is a global optimization algorithm based on swarm intelligence inspired by bee's behaviour of searching food sources⁴³. It has an ability to get out of a local minimum and can be efficiently used for multivariable, multimodal function optimization⁴⁴. And, it has fewer control parameters compared to other algorithms and successfully implemented for solving complex nonlinear optimization problems and engineering problems with high dimensionality⁴⁵. The ABC algorithm consists of three essential components: food sources, employed foragers and unemployed foragers, and the goal is to find the food source with the greatest profitability. Initially, half of the forager population is the employed foragers which is the same as the number of food sources, and the other half is unemployed foragers. Every food source is a possible solution to the problem. Employed foragers store about source information and share information with other foragers. There are two kinds of unemployed foragers, one is onlookers which wait in the nest and look for new food sources according to the information shared by the employed foragers, and the other is scouts which randomly search food source near the nest.

The basic steps of ABC algorithm are discussed below⁴⁶:

(a) Initialization phase

At beginning, SN number of D-dimensional food sources are randomly generated:

\[ x_{i,j} = x_{\min,j} + \text{rand}[0,1](x_{\max,j} - x_{\min,j}) \]  

(1)

Where, \( i = 1,2,\ldots,SN \), \( j = 1,2,\ldots,D \), \( X_i \) is the \( i \)-th food source, \( X_i = \{x_{i,1}, x_{i,2},\ldots,x_{i,D}\} \), \( x_{\min,j} \) and \( x_{\max,j} \) are the maximum and minimum values of \( x_{i,j} \), respectively. These food sources with certain fitness are randomly assigned to SN employed foragers.
(b) Employed foragers phase

Employed foragers try to search new food source around the current food sources:

\[ v_{i,j} = x_{i,j} + \text{rand}[-1,1](x_{i,j} - x_{k,j}) \]  

(2)

Where \( k = 1,2, ..., SN \) and \( k \neq i; v_{i,j} \) is the new food source. After the new food source is obtained, the greedy selection algorithm is used to compare the fitness of the new and old food source and select the better one.

(c) Onlookers phase

Employed foragers recruit onlookers by roulette gambling strategy according to the information of food sources. The probability of each food source being selected can be calculated by:

\[ P_i = \frac{\text{fit}(X_i)}{\sum_{j=1}^{SN} \text{fit}(X_j)} \]  

(3)

Where \( \text{fit}(X_i) \) is the fitness of the \( i \)-th food source. Then, a number is randomly generated from the interval of \([-1, 1]\). If the probability value of the food source is greater than the random number, the onlooker will generate a new food source from equation (2) and calculate the fitness of the new food source. If the fitness of the new food source is better than the old one, the onlooker will remember the new food source and abandon the old food source. Otherwise, it will keep the old food source. Trial is a parameter which is used to the number of times food sources have been abandoned.

(d) Scouts phase

When trial exceeds the limit that is the predetermined threshold, the food source would be abandoned and new food sources would be randomly generated as the initialization phase. And, the employed foragers switch to scouts to search new food sources. The best food source and the optimal fitness value would be obtained when the maximum number of explorations is reached.

**Support Vector Machine**

Support vector machine (SVM) was proposed by Cortes and Vapnik in 1995. It can be used for regression as well as classification. When used for regression modeling, it has advantages in establishing models to solve small sample, nonlinear and multidimensional problems. And, the penalty parameter \( (c) \) and the kernel parameter \( (g) \) determine the applicability and accuracy of the algorithm. The principle of SVM is discussed as follows:

Suppose there is a training sample \( D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\} \), the regression model \( f(x) = \omega^T x + b \) can meet the condition that the maximum deviation between \( f(x) \) and \( y \) is below \( \epsilon \) for any \( x \). Where, \( \omega \) and \( b \) are the
model parameters to be determined, $\epsilon$ is the tolerable deviation. Then, the optimization target of SVM can be formalized as

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{m} l_\epsilon(f(x_i) - y_i)$$  \hspace{1cm} (4)$$

Where, $\frac{1}{2} \|\omega\|$ is the measurement of function flatness, $C$ is the regularization constant, $l_\epsilon$ is $\epsilon$-insensitive loss function, and $l_\epsilon$ can be expressed as

$$l_\epsilon(z) = \begin{cases} 0, & \text{if } |z| \leq \epsilon \\ |z| - \epsilon, & \text{otherwise} \end{cases}$$  \hspace{1cm} (5)$$

By introducing relaxation variables $\xi_i$ and $\hat{\xi}_i$, the formula (4) is transformed into

$$\min_{\omega, b, \xi_i, \hat{\xi}_i} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{m} (\xi_i + \hat{\xi}_i)$$

subject to

$$f(x_i) - y_i \leq \epsilon + \xi_i$$

$$y_i - f(x_i) \leq \epsilon + \hat{\xi}_i$$

$$\xi_i \geq 0, \hat{\xi}_i \geq 0, i = 1, 2, ..., m$$  \hspace{1cm} (6)$$

By introducing lagrange multiplier $\mu_i \geq 0, \hat{\mu}_i \geq 0, \alpha \geq 0, \hat{\alpha} \geq 0$, lagrange function is obtained

$$L(\omega, b, \alpha, \hat{\alpha}, \xi_i, \hat{\xi}_i, \mu_i, \hat{\mu}_i) = \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{m} (\xi_i + \hat{\xi}_i) - \sum_{i=1}^{m} \mu_i \xi_i - \sum_{i=1}^{m} \hat{\mu}_i \hat{\xi}_i$$

$$+ \sum_{i=1}^{m} \alpha_i (f(x_i) - y_i - \epsilon - \xi_i) + \sum_{i=1}^{m} \hat{\alpha}_i (y_i - f(x_i) - \epsilon - \hat{\xi}_i)$$  \hspace{1cm} (7)$$

At the same time, satisfying the KKT condition, the regression model of SVM can be obtained as

$$f(x) = \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) \kappa(x_i^T x) + b$$  \hspace{1cm} (8)$$

Where, $\kappa(x_i^T x) = \Phi(x_i)^T \Phi(x_i)$, is the kernel function. The sample that can meet the formula $\hat{\alpha}_i - \alpha_i \neq 0$ is the support vector of SVM.

The selection of kernel function and related parameters in support vector machine has a great influence on the accuracy of the model.

**Modeling and Simulation**

In order to effectively use the inherent relationship between chlorophyll $a$ concentration and water quality indexes to simulate chlorophyll $a$ concentration, our work uses ABC to optimize the parameters $c$ and $g$ of SVM, and builds the ABC-SVM model of chlorophyll $a$ concentration. At the same time, SVM model is built as the control
group based on support vector machine. Modeling and simulation flow chart of ABC-SVM model is shown in figure 2.

![Modeling and simulation flow chart of ABC-SVM model.](image)

**Figure 2. Modeling and simulation flow chart of ABC-SVM model.**

The modeling and simulation process of ABC-SVM model and SVM model based on water quality indexes is introduced as follows:

(a) The data of 43 sampling points on March 26, 2018 is randomly divided into training dataset and testing dataset which include the data of 28 sampling points and the data of 15 sampling points respectively. Chlorophyll a concentration is set as the output layer and water quality indexes (DO, T, TN, TP) as the input layer of the model.
Mapminmax function is used in MATLAB software to normalize the training dataset and test dataset. The data normalization formula is as follows:

$$x^* = \frac{(x - x_{\text{min}})(y_{\text{max}} - y_{\text{min}})}{(x_{\text{max}} - x_{\text{min}})} + y_{\text{min}}$$  \hspace{1cm} (9)

Where, $x^*$ is the normalized output value, $x$ is the normalized input value, $x_{\text{min}}$ is the minimum value of the dataset to be normalized, $x_{\text{max}}$ is the maximum value of the dataset to be normalized, $y_{\text{min}}$ is the lower limit of normalized interval, $y_{\text{max}}$ is the upper limit of normalized interval.

(b) Based on the measured data of water quality indexes, the ABC-SVM model for simulating chlorophyll a concentration is built by using ABC to optimize SVM. At the same time, SVM model for simulating chlorophyll a concentration is built by using unoptimized SVM. When building chlorophyll a concentration model based on SVM, the sigmoid function is selected as the activation function, RBF is selected as the kernel function, and the penalty coefficient $c$ and kernel parameter $g$ can be obtained from model training process. ABC is used to optimize SVM to the parameters $c$ and $g$ in ABC-SVM modelling. The maximum number of iterations of ABC is set to 100, the limit number of experiments is set to 50, the number of employed foragers is set to 100, the number of investigated foragers is set to 10, and the number of observed foragers is set to 90. RMSE between simulated and measured chlorophyll a concentration is selected as fitness function. After ABC optimization, the optimal $c$ and $g$ can be obtained, and the ABC-SVM model for simulating chlorophyll a concentration can be obtained. Using the same data and basic conditions as ABC-SVM model, the SVM model is built after SVM training and testing.

(c) Modeling effect and model accuracy evaluation are conducted by comparing measured values and simulated values of chlorophyll a concentration of ABC-SVM modelling and SVM modelling, and analysing the error between measured values and simulated values of chlorophyll a concentration in the training process and the testing process of two models.

(d) Sensitivity analysis of chlorophyll a concentration to water quality indexes in ABC-SVM model simulation is carried out to determine the impact of each input index on chlorophyll a concentration simulation. The specific steps are as follows: First, selecting one of the four water quality indexes (DO, T, TN, TP) and change the values by a fixed multiple to generate new dataset. Then, the new dataset is used to simulate chlorophyll a concentration by using ABC-SVM model, and calculating RMSE between Chlorophyll a concentration simulated values and measured values. At last, RMSE is used as the evaluation index, sensitivity of chlorophyll a concentration to each index is analysed by comparing RMSE of ABC-SVM simulation under different change multiples of each index.
(e) The ABC-SVM model is used to simulate the chlorophyll a concentration in Donghu Lake at different times, the simulation results are analysed and the errors between simulated values and measured values are compared to study the universality at different times of the ABC-SVM model.

**Analysis and evaluation**

Determination coefficient ($R^2$), relative error (RE), absolute relative error (ARE) and root mean square error (RMSE) between the simulated values and the measured values are chosen as evaluation indexes to analyse the performance of the model. The equations of these indexes are defined as

$$R^2 = \left( \frac{\sum_{i=1}^{m} (y_m - \bar{y}_m)(y_s - \bar{y}_s)}{\sqrt{\sum_{i=1}^{m} (y_m - \bar{y}_m)^2(y_s - \bar{y}_s)^2}} \right)^2$$

(10)

$$RE = \frac{y_s - y_m}{y_m}$$

(11)

$$ARE = \frac{|y_s - y_m|}{y_m}$$

(12)

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_s - y_m)^2}$$

(13)

Where, $y_s$ is the simulated value, $y_m$ is the measured value, and $m$ is the length of the data series.

**Results and discussion**

**Comparison of modeling process between SVM model and ABC-SVM model**

When modeling, the model runs ten times and the average values of all simulated results are taken as the simulated values. The RMSE between the simulated values and the measured values is selected as the fitness function of the model, and the parameters of SVM are optimized by ABC using the training dataset. The ABC-SVM model is built to simulate the chlorophyll a concentration. The parameters of SVM model and the optimal parameters of ABC-SVM are shown in table 2. In SVM model, the penalty parameter $c$ is 9.19, the kernel parameter $g$ is 0.33. And, the optimized parameters of ABC-SVM are: the penalty parameter $c$ is 5.00, the kernel parameter $g$ is 0.63. That is, the two parameters of SVM model have been adjusted after ABC optimization.

<table>
<thead>
<tr>
<th>Model</th>
<th>$c$</th>
<th>$g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM model</td>
<td>9.19</td>
<td>0.33</td>
</tr>
<tr>
<td>ABC-SVM model</td>
<td>5.00</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Notes: $c$ is the penalty parameter, $g$ is the kernel parameter.
Figure 3 (a), figure 3 (b), figure 3 (c) and figure 3 (d) respectively show the comparison between simulated values and measured values of chlorophyll a concentration at each sampling point of SVM model training process, ABC-SVM model training process, SVM testing process and ABC-SVM testing process. It is shown that the fit degree between regression fitting line (Fit) and reference line (Y = T) is high both in SVM model and ABC-SVM model, and ARE of all the points for training in the two models does not exceed 1. Compared with SVM model, ABC-SVM model has fewer points that have large ARE between simulated value and measured value in the training process. In testing process, the fit degree between regression fitting line (Fit) and reference line (Y = T) is low and there are several points far from reference line (Y = T) for SVM model, however, ABC-SVM model has a good fit degree between regression fitting line (Fit) and reference line (Y = T), and all points are not far from the line reference line (Y = T). In addition, almost all the points for testing in SVM model have greater ARE than that in ABC-SVM model. What’s more, ARE of one point in SVM model testing process is even larger than 3. The higher the fit degree between regression fitting line (Fit) and reference line (Y = T), the closer the whole simulated values dataset is to the measured values dataset. Moreover, simulated values of the point with smaller ARE is closer to the measured values. The comparison of the modeling results from Figure 3 indicates that the models built by the two methods can be applied to the simulation of chlorophyll a concentration in Donghu Lake to a certain extent. Furthermore, the error between the measured values and the simulated values of ABC-SVM model is less than that of SVM model in both training process and testing process. And, the error between the simulated values and the measured values of ABC-SVM model is less than that of SVM model, whether in the training process or testing process. In other words, more suitable parameters are obtained through ABC optimization, and the error of modeling is reduced with SVM optimized by ABC.
Figure 3. The comparison of modeling results between SVM model and ABC-SVM model.

Simulation results and accuracy analysis of ABC-SVM model

The SVM model, ABC-SVM model and water quality indexes data are used to simulate chlorophyll a concentration. The accuracy of ABC-SVM model is evaluated by comparing the simulated values of the two models with the measured values, and comparing the error of the simulated values of the two models at each sampling point. $R^2$ and RMSE between the simulated value and the measured value as the evaluation indexes are select to analyse the error of the models.

Figure 4. The simulation results and error comparison of SVM model and ABC-SVM model.

Figure 4(a) shows the comparison between the simulated values of SVM model, the simulated values of ABC-SVM model and the measured values. RE represents the relative error between the simulated values of the model and the measured values. Figure 4 (b) compares RE of SVM model and RE of ABC-SVM model at each sampling
point. The comparison results show that the simulated value of ABC-SVM model is closer to the measured value
at each sampling point compared with the simulated values of SVM model, and, the RE of ABC-SVM model is
closer to 0 than RE of SVM model except for a few points. Then, $R^2$ and RMSE of simulated values and measured
values are compared both in the training process and testing process of SVM model and ABC-SVM model in
table 3. The results show that the $R^2$ of ABC-SVM model is greater than that of SVM model, and the RMSE of
ABC-SVM model is less than that of SVM model. $R^2$ in the testing process of ABC-SVM model is 0.96, which
is 0.04 higher than that of SVM model. And, RMSE in the testing process of ABC-SVM model is 10.40 $\mu$g/L, which is 4.12 $\mu$g/L lower than that of SVM model.

<p>| Table 3. Error analysis of SVM model and ABC-SVM model. |</p>
<table>
<thead>
<tr>
<th>Model</th>
<th>Training process</th>
<th>Testing process</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>RMSE($\mu$g/L)</td>
</tr>
<tr>
<td>SVM model</td>
<td>0.97</td>
<td>9.56</td>
</tr>
<tr>
<td>ABC-SVM model</td>
<td>0.98</td>
<td>8.62</td>
</tr>
</tbody>
</table>

Figure 4 shows the simulation performance of ABC-SVM model from the level of each sampling point, and the
data in the table 3 integrally and comprehensively reveals the accuracy of the model. The simulation comparing
results and error analysis results show that the simulated values of ABC-SVM model are closer to the measured
values than that of SVM model, and the simulation result error of ABC-SVM model is smaller than that of SVM
model. In a word, ABC has effectively improved the performance of SVM in chlorophyll a concentration
simulation. The ABC-SVM model based on water quality index has high accuracy and good simulation effect on
chlorophyll a concentration.

Sensitivity of chlorophyll a concentration to water quality indexes in ABC-SVM model simulation
A new dataset is generated by changing one of the four water quality indexes with a fixed multiple. And based on
the new dataset, chlorophyll a concentration is simulated by ABC-SVM model, and the simulated values are
compared with the measured values to study the sensitivity of chlorophyll a concentration to DO, T, TN and TP
in ABC-SVM model simulation. RMSE between the simulated values and measured values is chosen as evaluation
index to assess the sensitivity. Figure 5 shows the sensitivity analysis results.
One of the four water quality indexes respectively change -0.5, 0.25, 0.25, 0.5 multiple in turn to generate new dataset, and new dataset is used for ABC-SVM model to simulate chlorophyll a concentration. It can be seen from the sensitivity analysis results in figure 5 that RMSE of ABC-SVM model simulation results with each new dataset is greater than RMSE of simulation result with unchanged dataset. It means that the change of each four water quality indexes has an impact on the simulated values of chlorophyll a concentration in the simulation of ABC-SVM model. For the same water quality index, the greater the change, the greater the value of RMSE. In addition, RMSE increased obviously with the change of TN or TP. And, under the same change multiple, the change of RMSE value is the largest when TP changes, while the change of RMSE value is the smallest when T changes. In a word, when using ABC-SVM model for simulation, the change of TP has the greatest impact on the model results, and chlorophyll a concentration is more sensitive to TP than other water quality indexes.

The formation mechanism of lake eutrophication and algal bloom is complex. As a key index of lake eutrophication and algal bloom, chlorophyll a concentration in lake is related to many factors. When the water body is seriously eutrophic, the algae in the water will proliferate maliciously, and the respiration of algae and the decomposition of dead algae by microorganisms will consume large amounts of oxygen, as a result it will sharply reduce the DO in the water. In other words, the decrease of DO is the result of lake eutrophication. In addition, the sampling time interval is small, and the water surface temperature of Donghu Lake has little change, so there is little difference in the measured values of T at each sampling point. Therefore, the impact of DO, T on chlorophyll a concentration is not big. However, nitrogen, phosphorus and other nutrients are the main causes of lake eutrophication. So, the change of TN and TP in water body will greatly affect the growth of algae, and then affect the concentration of chlorophyll a in water. Eutrophic ecosystems can display P limitation, N limitation, and colimitation\textsuperscript{54}. Most scholars believe that the increase of nitrogen, phosphorus and other nutrients
concentration is the main reason for algae mass reproduction, and phosphorus is the key factor\(^5\). And sensitivity analysis result of ABC-SVM model shows that the change of TP has the greatest impact on the model results, and chlorophyll a concentration is more sensitive to TP than other water quality indexes, which is consistent with these conclusions.

**Universality of ABC-SVM model at different times**

The measured data of water quality indexes at four times are sent to ABC-SVM model to simulate the chlorophyll a concentration by using the ABC-SVM model, and the universality of ABC-SVM to simulate the chlorophyll a concentration of Donghu Lake in time dimension is analysed by comparing the simulated values and measured values at different times.

The figure 6 shows the comparison between the simulated values and measured values of chlorophyll a concentration of different time, and the results demonstrated that \(R^2\) between the simulated values and the measured values of chlorophyll a concentration at four different times can reach 0.89, 0.95, 0.93 and 0.92, respectively. Among them, \(R^2\) of four times exceedes 0.89, which indicated that the simulated values of chlorophyll a concentration were close to the measured values. At the same time, it can be found that ARE of all sampling points at four different times is generally not big, there is only one sampling point whose ARE between the simulated value and the measured value exceeded 1. Although there are some points far from reference line (Y = T) at four times, that is, some sampling points have obvious error between the simulated values and the measured values. However, the difference is not big between the simulated values and the measured values of chlorophyll a concentration at four different times. Furthermore, there are observation errors in the modeling data and systematic errors in the simulation results, and they all have an impact on the error between the simulated values and the measured values. In general, the ABC-SVM model has good universality and can be used to simulate the chlorophyll a concentration of Donghu Lake at different times.
In this study, we use ABC to optimize the parameters of SVM and build ABC-SVM model to simulate the chlorophyll a concentration in Donghu Lake based on the correlation between chlorophyll a concentration and water quality indexes (DO, T, TN, TP). The modeling results show that more suitable parameters are obtained through ABC optimization, and the error of modeling is reduced with SVM optimized by ABC. R² and RMSE in the testing process of ABC-SVM model is 0.96 and 10.40 μg/L, ABC optimization increases R² by 0.04 and reduces RMSE by 4.12 μg/L compared with SVM model. In other words, the ABC-SVM model based on water quality index has high accuracy and good simulation effect on chlorophyll a concentration. Sensitivity analysis result of ABC-SVM model shows that the change of TP has the greatest impact on the model results, and
chlorophyll a concentration is more sensitive to TP than other water quality indexes. When using the ABC-SVM model simulate the chlorophyll a concentration with water quality indexes measured data at four different times, R^2 between the simulated values and the measured values exceeds 0.89, and ARE of all sampling points at four different times is generally not big. Universality analysis results indicate that the ABC-SVM model has good universality and can be used to simulate the chlorophyll a concentration of Donghu Lake at different times.

**Data availability**

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

**References**


Acknowledgements

This work is supported by the National Key R&D Program of China [No.2017YFC0405901], and the National Natural Science Foundation of China [No.51579108]. The authors would like to thank Dr. Li Li for her invaluable suggestions and contributions.

Author information

Affiliations
Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Xiaodong Tang. Funding acquisition is performed by Mutao Huang. The first draft of the manuscript is written by Xiaodong Tang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Corresponding author

Correspondence to Mutao Huang.

Ethics declarations

Competing interests

The authors declare no competing interests.