Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2020.25621

189 (2020) 207–216
June

A multi-factor data-driven prediction model for cyanobacteria blooms in lakes and reservoirs

Lei Zheng[a], Bo Hu[b], Aizhong Ding[a,*]

[a]*College of Water Sciences, Beijing Normal University, Beijing 100875, China, Tel. +86 10 5880 5051; email: ading@bnu.edu.cn (A. Ding), Tel. +86 138 10821932; email: Zhengleilei@bnu.edu.cn (L. Zheng)*
[b]*School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China, Tel. +86 18210016058; email: crosscshb@163.com (B. Hu)*

Received 21 June 2019; Accepted 1 January 2020

ABSTRACT

The generation process of cyanobacteria blooms in lakes and reservoirs is complex. Many influencing factors with time-varying characteristics make predictions more difficult. This study proposed a systematic multifactor data-driven prediction model. Through this systematic method, a prediction model could be built with fewer key influencing factors for cyanobacteria bloom in lakes and reservoirs. First, we introduced a definition of the correlation coefficient and synthesized the consistency of the changing trend on a time series and the structural similarity of state characteristics to select the key factors influencing cyanobacteria blooms in lakes and reservoirs. We used an improved wavelet threshold method to filter the data noise involved in modeling. After the nonlinear time-series analysis method of phase-space reconstruction, we realized a clustering method according to a comprehensive metric scale of similarity. Then, we determined the partial neighbor information of current points in each time series to predict cyanobacteria blooms in lakes and reservoirs using a local linear regression prediction model. Compared with conventional prediction models and the classical long short-term memory model, the results of the trend consistency and the accuracy demonstrated that the proposed model was superior. This verified the validity of the systematized model construction method. This model can be used to predict cyanobacteria blooms in lakes and reservoirs and supports the treatment of the water environment.

Keywords: Lakes and reservoirs; Cyanobacteria bloom; Key influencing factors; Correlation coefficient; Prediction; Metric scale of similarity

1. Introduction

In recent years, rapid industrialization and urbanization with limited environmental protection have resulted in the eutrophication of water in lakes and reservoirs. Eutrophication is the source of cyanobacteria bloom formation, aggregation, and outbreak. At present, mechanism models and data-driven models are the two main methods used to research the cyanobacteria bloom prediction model. The mechanism model is used to simulate the ecological dynamics of cyanobacteria growth process based on the differential equations [1-3]. Multiple factors influence cyanobacteria bloom in lakes and reservoirs, including total phosphorus (TP), total nitrogen (TN), water temperature, and even wind speed, as well as other meteorological factors [4]. The multivariate model can be used to obtain sufficient information to understand the mechanism of cyanobacteria bloom more clearly and can improve the accuracy of the model. It also, however, increases monitoring and maintenance costs. In addition, in some politically sensitive areas, because of the inability to use some sensors, the application of the existing multivariate model is limited.

* Corresponding author.

Understanding how to use fewer variables to build a prediction model of cyanobacteria bloom in lakes and reservoirs has practical significance for specific areas of political or economic sensitivity.

Cyanobacteria blooms are characterized by random-like sensitive chaos [5]. Therefore, it is more difficult to calibrate the parameter of the mechanism-driven model with time-varying characteristics. The data-driven model is a method to mine the inherent laws hidden of a research object from data information, which stiffs to overcome the difficulty of the mechanism model of complex systems [6-10]. To address the sensitive and chaotic properties of cyanobacteria blooms in lakes and reservoirs, the existing multifactor chaotic prediction models can be used for reference in finding the delay time and embedding dimension [11,12]. Understanding how to construct the framework of a data-driven model from the perspective of systematization needs to be explored. This study provides a systematic modeling process, including the choice of modeling factors, data cleaning, and optimization and determination of the model parameters on a nonlinear time-varying object. The following contributions are made:

- We proposed the definition of a correlation coefficient to comprehensively consider the trend consistency of a time series and the structural similarity of state characteristics on cyanobacteria blooms in lakes and reservoirs to select the key factors during the formation of cyanobacteria blooms.
- After data reconstruction for the phase space of the time series, we improved the clustering algorithm by coordinating distance, similarity, horizontal migration, and amplitude expansion as a comprehensive metric scale of similarity to select and identify the local neighborhood of the prediction center.

This study is organized as follows: Section 2 introduces the theory and construction framework of the prediction model. Section 3 presents the experiments and comparisons of different models. Section 4 provides conclusions.

2. Theory and construction of the framework

2.1. Selection of the influencing factors for modeling

The characterization factors of cyanobacteria blooms in urban lakes and reservoirs are chlorophyll-a concentration and algae density, which are closely related. Although the online and real-time monitoring of chlorophyll-a concentration is convenient, the cost of online monitoring of algae density is high. Because chlorophyll-a concentration not only can characterize the algae stock [13] but also can reflect the physical, chemical, and biological indicators of water, it is used as a characterization factor for the formation process of cyanobacteria blooms in lakes and reservoirs.

Many factors influence the formation of a cyanobacteria bloom, including pH value, temperature, TN and TP, and dissolved oxygen (DO). The selection of modeling factors directly affects the accuracy of the model. Therefore, in this study, we proposed the definition of a correlation coefficient to screen the key factors to improve the final prediction accuracy by selecting the information with a strong correlation.

2.1.1. Consistency of trends in time-series change

For the chlorophyll-a concentration time series, $X_0 = \{x_0(t), t = 1, 2,\ldots, N\}$ as the characterization factor, and the other influence factor time series $X_h = \{x_h(t), h = 1, 2,\ldots, M\}$ acquired in the same period. We calculated the proximity of the curve slopes of the two-time series at each moment. If it was equal or smaller, the change trend of the two-time series was more consistent.

The consistency of the change trend is expressed as follows:

$$\mathrm{Corr}(h) = \mathrm{Corr}(X_0, X_h) = \frac{1}{N-1} \sum_{t=1}^{N-1} \frac{1}{1+\left|\alpha^{(1)}\left(y_0(t+1)\right)-\alpha^{(1)}\left(y_h(t+1)\right)\right|} \tag{1}$$

where $y_0(t) = \frac{x_0(t)}{x_0(1)}$, $y_h(t) = \frac{x_h(t)}{x_h(1)}$, $\alpha^{(1)}(y_0(t+1)) = y_0(t+1)–y_0(t)$, and $\alpha^{(1)}(y_h(t+1)) = y_h(t+1)–y_h(t)$. Corr($h$) is the consistency coefficient of the variation trend between the hth influencing factor and the characterization factor, which reflect the consistency of the variation trend between the time series of the characterization factor and the influencing factor.

2.1.2. Relevance of structural similarity of state characteristics

Autocorrelation is a unique property of time series, which is represented by the correlation of characteristic structures of time-series state characteristics at different times.

For the other influence factor time series $X_h = \{x_h(t), h = 1, 2,\ldots, M\}$, take t and $s \in N$, and the autocorrelation coefficient Auco(t, s) of the sequence is as follows:

$$\mathrm{Auco}(h) = \mathrm{Auco}(t,s) = \frac{E\left(x_h(t)-\mu\right)\left(x_h(s)-\mu\right)}{\mathrm{sqrt}\left(Dx_h(t)\cdot Dx_h(s)\right)} \tag{2}$$

where μ is the mean of the sequence, E is the mathematical expectation, and D is the variance. Auco(h) is the autocorrelation coefficient of the hth influencing factor, which reflects the structural similarity of state characteristics of the influencing factor.

The time series of the characterization factor can be calculated similarly. The autocorrelation coefficient is Auch for the time series of the characterization factor $X_0 = \{x_0(t), t = 1, 2,\ldots, N\}$.

2.1.3. Definition of a correlation coefficient

The greater the consistency of the changing trend between the influencing factors and the characterization factors, the more consistent the changing trend of the time series of the two factors will be. The smaller the correlation coefficient difference between the influencing factors and the characterization factor, the more similar the state characteristics will be. On the basis of the previous two indicators, the correlation coefficient is defined as follows:

$$\mathrm{Sico}(h)=\frac{\mathrm{Corr}(h)}{\left|\mathrm{Auco}(h)-\mathrm{Auch}\right|} \tag{3}$$

where Auch is the autocorrelation coefficient of the characterization factor chlorophyll-a. Sico(h) is the correlation coefficient between the hth influencing factor and the characterization factor. The larger Sico(h) is, the greater the correlation between the influencing factor and the characterization factor.

In this study, we selected the key factors that had the greatest correlation with the characterization factors for modeling research. We selected the factors corresponding to the maximum correlation coefficient as the key factors in the formation process of cyanobacterial blooms and constructed a multifactor time-series model.

2.2. Data filtering based on wavelet

Time series data with noise inevitably are disturbed by various factors, such as instruments and the environment, in the measurement process. It is essential to effectively filter and preprocess the time-series data for the precision of the prediction model.

Time-series data for the sensitive chaotic properties of cyanobacteria blooms in lakes and reservoirs have spikes and mutations. Therefore, it is difficult for traditional filtering methods to distinguish the intrinsic random-like signals and extrinsic high-frequency noise. A wavelet theoretically can distinguish noise from real signals by carefully selecting the threshold or threshold function, and retaining the effective components in the original data to the greatest extent [14,15]. This signal, however, may be distorted by using the universal threshold with the wavelet coefficients. To avoid this problem, the thresholds on different decomposition scales should be different to adapt to the noise distribution at each level.

Therefore, we proposed an improved threshold selection method. The formulation is represented as follows:

$$T=\frac{\sigma\sqrt{2\ln\left(n_l\right)}}{e^{l-1/2}} \tag{4}$$

where σ is the standard variance of the noise, σ = median(W^{HH})/0.6475, W^{HH} is the orthogonal wavelet coefficient of noise in a high-frequency sub-band and median means picking median, l is the decomposition scale, and n_l is the signal length of the corresponding scale after wavelet transform. With an increase in the decomposition scale l, the threshold decreases correspondingly. The new threshold T is adaptive and more in line with the distribution of noise at all levels.

The hard threshold method in the threshold function can preserve local features, such as edge of signal, but discontinuity after filtering will cause the filtering result to have a large variance. Although the soft threshold method is continuous at the threshold point and the filtering result is relatively smooth, it shrinks all of the coefficients that are larger than the threshold. This causes the filtering result to deviate significantly, which then affects the filtering effect. To overcome the shortcomings of the previous two methods, we proposed a new threshold function:

$$\hat{\omega}_{l,k}=\begin{cases}\omega_{l,k}-\zeta T+\dfrac{2\zeta T}{1+e^{\left(\left|\frac{T}{\omega_{l,k}}\right|^{u}-1\right)}+u} & \left|\omega_{l,k}\right|\geq T\\ 0 & \left|\omega_{l,k}\right|<T\end{cases} \tag{5}$$

where k is layers of decomposition, $\omega_{l,k}$ is wavelet coefficients, $\hat{\omega}_{l,k}$ is the wavelet coefficients after de-noising, ζ and u are tune parameters, $\zeta\in(0,1)$, the general value of u is 1, and T is a constant threshold.

2.3. Phase-space reconstruction of time series

For the multifactor time series, X_t = ($x_{0,t}$, $x_{1,t}$,…, $x_{M,t}$), t = 1,2,…, N, and the phase-space reconstruction is as follows:

$$\overline{X}_i=\begin{pmatrix}x_{0,i},x_{0,i-\tau_1},\ldots,x_{0,i-(m_0-1)\tau_1},\\ x_{1,i},x_{1,i-\tau_2},\ldots,x_{1,i-(m_1-1)\tau_2},\\ \vdots\\ x_{M,i},x_{M,i-\tau_M},\ldots,x_{M,i-(m_M-1)\tau_M}\end{pmatrix} \tag{6}$$

where $i=\max\limits_{0\leq v\leq M}\left(m_v-1\right)\tau_v+1,\ldots N$, τ_v and m_v are, respectively, delay time and the embedding dimension of the time series for factor v.

The selection of the embedding dimension and the delay time for a chaotic time series based on the phase-space reconstruction of multiple factors is significant. If the selected embedding dimension is too small, some points belonging to different parts will intersect in small neighborhoods of regions and result in a greater prediction error. If the selected embedding dimension is too large, the historical data needed for phase-space reconstruction will increase and the rounding error and noise will have a greater impact on the results, which would result in greater prediction error.

This study synthesized the C–C method, mutual information method, and minimum prediction error method to optimize the selection of an embedding dimension and delay time. The specific steps are as follows:

- We selected the delay time of each factor in a multiple factors time series according to the mutual information method.
- We selected the delay time of each factor in a multiple factors time series based on the C–C method.
- We matched the range of the embedding dimensions of a two-set time series from the low to the high end of the delay time, which was calculated before the phase-space reconstruction. Then, we determined the nearest neighborhood point of the prediction center based on the Euclidean distance method.
- We used the next evolutionary point in the nearest neighbor point of the local linear regression model as the prediction value and then calculated the average one-step prediction error square.

- We repeated steps 3 and 4, and iterated the process until we reached the maximum embedding dimension. We determined the optimal embedding dimension and delay time-based on the minimum square of error of the average one-step prediction.

2.4. Local neighborhood determination based on clustering analysis

We identified the typical evolutionary modes of the system by clustering and then extracted samples with that had a high similarity to the prediction points for approximation. Then, we improved the accuracy of the prediction while greatly reducing the computational load of modeling and simulation [16]. The traditional clustering method [17,18] used Euclidean distance to measure the correlation between phase points. For the sensitive chaotic properties of cyanobacteria blooms in lakes and reservoirs that represented fractal similarity, this study proposed a function to integrate four similar features at a scaled metric to quantify the similarity and select the local neighborhood of the prediction center by a clustering algorithm.

2.4.1. Conventional correlation measurement method

2.4.1.1. Distance method

The conventional distance method uses the modulus of vector difference to describe the similarity between points.

Let the difference modulus of the two reconstructed phase vectors $\overline{X}_i$ and $\overline{X}_j$ be $R\left(\overline{X}_i,\overline{X}_j\right)$; the smaller the value is, the closer the two phase points are. To map the value of $R\left(\overline{X}_i,\overline{X}_j\right)$ to the interval [0, 1], a function $Y\left(\overline{X}_i,\overline{X}_j\right)$ is set up:

$$Y\left(\overline{X}_i,\overline{X}_j\right)=\frac{1}{R\left(\overline{X}_i,\overline{X}_j\right)+1} \tag{7}$$

where $Y\left(\overline{X}_i,\overline{X}_j\right)\in\left[0,1\right]$.

2.4.1.2. Similarity measure method

Relevance analysis is a method used to analyze the similarity between the two reconstructed phase vectors. It is described by the angle between vectors. Let the angle between $\overline{X}_i$ and $\overline{X}_j$ be $\theta\left(\overline{X}_i,\overline{X}_j\right)=\arccos\frac{\overline{X}_i\cdot\overline{X}_j}{\left|\overline{X}_i\right|\cdot\left|\overline{X}_j\right|},\theta\left(\overline{X}_i,\overline{X}_j\right)\in\left(0,\pi\right)$.

To map the value of $\theta\left(\overline{X}_i,\overline{X}_j\right)$ to the interval [0, 1], a function $g\left(\overline{X}_i,\overline{X}_j\right)$ is set up:

$$g\left(\overline{X}_i,\overline{X}_j\right)=\frac{1}{1+\theta\left(\overline{X}_i,\overline{X}_j\right)} \tag{8}$$

where $g\left(\overline{X}_i,\overline{X}_j\right)\in\left[0,1\right]$.

2.4.2. Similarity transformation

The orderliness of chaotic systems is represented by the similarity of fractals with different sizes. To effectively identify the performance similarity of structural proportions, it is necessary to measure the amplitude and range of similarity fluctuation.

2.4.2.1. Horizontal migration

Horizontal migration is the amount of migration that fluctuates around mean values. It is critical for exploring and determining the specific fractal morphology of cyanobacterial blooms. In this study, we used the difference between the mean phase points to represent the horizontal migration. The horizontal migration expression of time-series phase points is as follows:

$$\alpha\left(\overline{X}_i,\overline{X}_j\right)=\exp\left(-\frac{\left(\mu\left(\overline{X}_i\right)-\mu\left(\overline{X}_j\right)\right)^2}{\sigma^2\left(\overline{X}_i\right)+\sigma^2\left(\overline{X}_j\right)}\right) \tag{9}$$

where $\alpha\left(\overline{X}_i,\overline{X}_j\right)\in\left[0,1\right]$ is the horizontal migration of phase vectors $\overline{X}_i$ and $\overline{X}_j$, $\mu\left(\overline{X}_i\right)$ and $\mu\left(\overline{X}_j\right)$ are average values, and $\sigma\left(\overline{X}_i\right)$ and $\sigma\left(\overline{X}_j\right)$ are mean-variance, respectively.

2.4.2.2. Amplitude expansion

The reconstructed phase point data appear to be similar on the surface, but in fact, their fluctuations are not the same. Amplitude expansion is used to scale the relative magnitude of shape between two-phase points. It is necessary to measure its scale of amplitude expansion. The expression of magnitude dilation is as follows:

$$\beta\left(\overline{X}_i,\overline{X}_j\right)=\exp\left(-\frac{\left|\sigma^2\left(\overline{X}_i\right)-\sigma^2\left(\overline{X}_j\right)\right|}{\sigma^2\left(\overline{X}_j\right)}\right) \tag{10}$$

where $\beta\left(\overline{X}_i,\overline{X}_j\right)\in\left[0,1\right]$ is the amplitude expansion of $\overline{X}_i$ and $\overline{X}_j$.

2.4.3. Comprehensive metrics scale

On the basis of the calculated distance, similarity, horizontal migration, and amplitude dilation of phase points, we proposed a comprehensive metric to measure the similarity between phase points. The expression is as follows:

$$f\left(\overline{X}_i,\overline{X}_j\right)=\omega_1Y\left(\overline{X}_i,\overline{X}_j\right)+\omega_2g\left(\overline{X}_i,\overline{X}_j\right)+\omega_3\alpha\left(\overline{X}_i,\overline{X}_j\right)+\omega_4\beta\left(\overline{X}_i,\overline{X}_j\right) \tag{11}$$

where $f\left(\overline{X}_i,\overline{X}_j\right)$ is the comprehensive metric of $\overline{X}_i$ and $\overline{X}_j$. The larger the value is, the more similar the phase points are. To satisfy the requirements, ω_1, ω_2, ω_3, and ω_4 are the corresponding weights:

$$\omega_1 + \omega_2 + \omega_3 + \omega_4 = 1 \quad (12)$$

By adjusting the weights, we adjusted the focus of distance, similarity, horizontal migration, and amplitude expansion on the measurement of the similarity between phase points. It was convenient to synthesize four features flexibly to measure the similarity of the phase points for the study of complex systems with fractal self-similarity characteristics. Here, $\omega_1 = \omega_2 = \omega_3 = \omega_4 = 0.25$.

2.5. First-order local linear regression prediction model

After determining the nearest local neighborhood of the phase points generated by cyanobacteria blooms in lakes and reservoirs, we constructed the chaotic first-order local recursive function as follows:

$$\overline{X}_{t+1} = A_t + B_t \overline{X}_t \quad (13)$$

where A_t, B_t are the coefficient matrix obtained by least-squares estimation recursion based on local adjacent domain data.

3. Example analyses

3.1. Data source

The data obtained were from cyanobacteria blooms monitored at Jinshu station, Taihu Lake, Jiangsu Province, China. The specific indicators and units are given in Table 1.

We monitored 2,142 groups of data in cyanobacteria blooms every 4 h, from January 1, 2011, to January 3, 2012. The small sampling interval was good for capturing the time-varying dynamic characteristics of the system. In addition, the sensitivity of chaos relative to the initial value of the system demonstrated that its prediction ability was limited, which led to the short-term prediction of the chaotic system and to long-term inaccuracy.

In this study, to reserve a certain regulation time for early warning decision-making processes, we predicted chlorophyll-a concentration 3 d in advance (that is, the next 18 sampling time values are predicted). From the perspective of data modeling in the information field, we needed to accumulate a complete period of data before modeling. We selected two periods with a certain fluctuation of chlorophyll-a concentration to test the feasibility of this model. Modeled with about 75% of the data and predicted the following 3 d, modeled with about 85% of the data and predicted the following 3 d.

3.2. Selection of key influencing factors

According to the correlation coefficient analysis, we obtained the variation consistency coefficient between the time series of characterization factors and each time series of influencing factors, as shown in Table 2.

We also calculated the autocorrelation coefficients of the factors influencing the formation of cyanobacteria blooms, as shown in Table 3.

We calculated the correlation coefficient between the characterization factors and the influencing factors to determine the key influencing factors of cyanobacteria bloom formation process, as shown in Table 4.

According to Table 4, the correlation coefficient between the time series of TN concentration and the characterization factors was the largest when comprehensively considering the consistency of the trend and the structural similarity of state characteristics in the time series. The key influencing factors affecting the chlorophyll-a concentration were TN and the TP concentration. After consulting the advice of experts in the field, we established a prediction model based on the time series of the TP and chlorophyll-a concentration.

3.3. Evaluation analysis with data cleaning and de-noising

Results of the wavelet threshold de-noising of the time series are shown in Table 5.

Table 1
Formation influence factor of cyanobacteria bloom

Name of factors	Temperature	pH	DO	TN	TP	Chlorophyll-a concentration
Unit	°C	Dimensionless	mg/L	mg/L	mg/L	mg/L

Table 2
Consistency of trends in time series

Influence factor	Temperature	pH	DO	TN	TP
Consistency of trends in time series	0.8739°C	0.9084	0.9194	0.9379	0.9022

Table 3
Autocorrelation coefficient

Influence factor	Temperature	pH	DO	TN	TP	Chlorophyll-a concentration
Autocorrelation coefficient	0.0968°C	0.0807	0.0918	0.0762	0.0759	0.0673

Table 4
Correlation coefficient

Influence factor	Temperature	pH	DO	TN	TP
Correlation coefficient	29.6237	67.7910	37.5265	105.3820	104.9070

Table 5
Evaluation and comparison of de-noising effect

Experimental data (methods)	SNR (dB)	Mean square error
Chlorophyll-a concentration time-series data (general threshold and soft threshold function)	14.8978	1.1484
Chlorophyll-a concentration time-series data (general threshold and improved threshold function)	16.4998	0.7941
Chlorophyll-a concentration time-series data (improved threshold and improved threshold function)	21.1500	0.2722
TP concentration time-series data (general threshold and soft threshold function)	12.3062	7.3041e–5
TP concentration time-series data (general threshold and improved threshold function)	9.0514	1.5454e–4
TP concentration time-series data (improved threshold and improved threshold function)	19.4075	1.4238e–5

From Table 5, both the SNR and the mean square error are better with the improved threshold and the threshold function.

3.4. Phase-space reconstruction of the multifactor time series

We determined the delay time in the multifactor time series according to the mutual information method. The delay time for chlorophyll-a concentration was $\tau_1 = 15$ and TP concentration was $\tau_2 = 15$. According to the C–C method, the delay time for chlorophyll-a concentration was $\tau_1 = 7$ and the TP concentration was $\tau_2 = 6$.

We set the embedding dimensions for time series as an integer from 2 to 10. According to the delay time range obtained by the C–C and mutual information methods, we calculated the average one-step prediction error square by matching the different embedded dimensions. When the average one-step prediction error square was the smallest, the corresponding optimal embedding dimension and delay time were as follows: the embedding dimension of chlorophyll-a concentration time series was $m_1 = 3$, and the delay time was $\tau_1 = 7$; the embedding dimension of TP concentration time series was $m_2 = 2$, and the delay time was $\tau_2 = 6$.

3.5. Prediction model

In the phase-space reconstruction of multiple factors based on the optimized embedding dimension and delay time, we used the metric scale introduced in Section 2.4 to improve the clustering algorithm of data with sensitive chaotic properties to determine the local neighborhood of the prediction center. The predicted results are shown in Fig. 1.

Fig. 1 shows that this model can predict the rising and falling trends of chlorophyll-a concentration, providing feasible lead time for early decision-making and regulation. To verify the validity of the multifactor prediction model proposed in this study, we established and tested a single-factor and multifactor prediction models, as well as a traditional hierarchical clustering method. We recently used the more popular long short-term memory (LSTM) model for comparison. The LSTM model consisted of three hidden layers with 20 memory cells in each layer. We used the sigmoid function and tanh function as activation functions. The window length was 6 and the training times were 100. In this study, we selected the average relative error as the evaluation index of the model. The average relative errors of different prediction models are shown in Table 6.

As shown in Table 6, compared with the single-factor model, the multifactor time-series model was better and provided more abundant modeling information to describe the evolution of the system. This showed that the proposed method based on similarity was more suitable for the time series of cyanobacteria blooms in lakes and reservoirs with sensitive chaotic properties.

LSTM is a kind of time-recurrent neural network, which is suitable for processing and predicting time-series information with relatively long intervals and delays. We limited the LSTM model by a single factor and its prediction accuracy was slightly lower. Although the LSTM model with multiple factors was superior to the traditional method, it was slightly lower than the model used in this study. This may have been due to the fact that fewer multicycle intensive data had accumulated for the cyanobacteria blooms in the lakes and reservoirs compared with the time-series information. This affected the application effect of the model.

To verify the validity and applicability of the proposed algorithm, we downloaded public datasets from the GLEON network (Where the Southeast Environmental Research Center at Florida International University operates a network of 331 fixed sampling sites distributed throughout the estuarine and coastal ecosystems of south Florida. The purpose of this network is to address concerns in regional water quality which cross and overlap separate political boundaries. Funding has come from different sources with individual programs being added as funding became available). The locations included (a) Tarpon Bay and (b) Oyster Bay. We sampled the data at equal intervals, with a sampling period of 1 month, and the data had accumulated for more than 10 y.

After analysis and optimization of the modeling data, we obtained the embedding dimension and delay time of

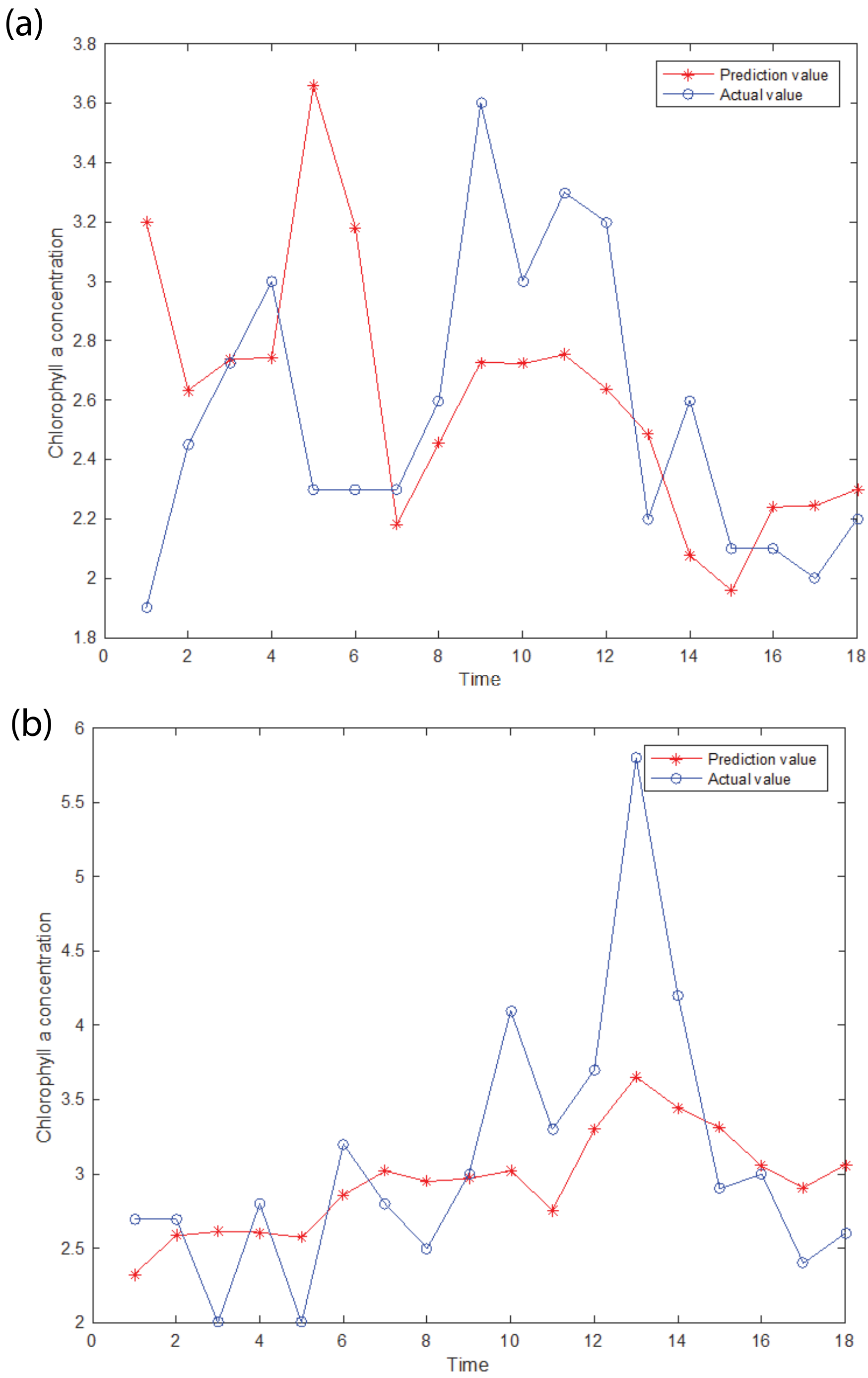

Fig. 1. Results of the prediction based on an improved clustering method with key influencing factors. Prediction results from (a) 1,640th and (b) 1,860th monitoring data.

the corresponding location, respectively: (a) Embedding dimension of the chlorophyll-a concentration time series is $m_1 = 3$, and the delay time is $\tau_1 = 5$; the embedding dimension of TP concentration time series is $m_2 = 3$, and the delay time is $\tau_2 = 4$. (b) The embedding dimension of the chlorophyll-a concentration time series is $m_1 = 3$, and the delay time is $\tau_1 = 5$; the embedding dimension of TP concentration time series is $m_2 = 6$, and the delay time is $\tau_2 = 4$.

Fig. 2 shows that the prediction trend was better than that for Taihu Lake in this paper. This may have been due to the fact that although the data volume accumulation for these two sites was small, the time span was long, which benefited the learning of the cyclical change mode. In this study, the data volume for Taihu Lake appeared to be relatively large, but it had only one year's periodic accumulation. This two bays accumulated long-term data with the longer sampling period, but the amount of accumulated data is small. It is not suitable to use LSTM algorithm which needs more data. The average relative errors of different prediction models are shown in Table 7. From the point of view of data science, the pattern with fewer samples was insufficient for learning, which was not conducive to improve its recognition ability.

4. Conclusion and future work

If a certain area has less accumulated historical empirical rules, we can use data-driven methods to build an early warning prediction model, which would save the cost of accumulated experience. In addition, we can reduce the variables involved in modeling to reduce the economy and time cost of sensor acquisition and maintenance. Typically, to predict cyanobacterial blooms in lakes and reservoirs, data-driven prediction models can be constructed theoretically according to the single-factor time series. Although this approach has practical applications, it is often difficult to obtain complete information for complex dynamic systems.

Therefore, the key factors needed to match the characterization factors for cyanobacteria blooms in lakes and reservoirs can be screened according to the algorithm proposed in this study. Furthermore, we proposed a systematical framework process to construct a multifactor data-driven prediction model to better study evolutionary characteristics. The main process is as follows:

- The factors involved in modeling directly affected the accuracy of prediction. Therefore, this study proposed a definition of the correlation coefficient. We determined the key factors influencing the formation process of cyanobacteria blooms in lakes and reservoirs based on the correlation of the consistency of the change trend and the structural similarity of state characteristics.
- For modeling data with sensitive chaotic properties and inevitable noise, this study improved the traditional

Table 6a
Prediction error of different models

Name of model	Average relative error
Single-factor prediction model based on traditional hierarchical clustering method	0.1906
Single-factor prediction model based on LSTM	0.2310
Two-factors prediction model based on LSTM with key influencing factors	0.2277
Two-factor prediction model based on improved clustering method with key influencing factors	0.1797

Table 6b
Prediction error of different models

Name of model	Average relative error
Single-factor prediction model based on the traditional hierarchical clustering method	0.2090
Single-factor prediction model based on LSTM	0.2310
Two-factor prediction model based on LSTM with key influencing factors	0.2303
Two-factors prediction model based on improved clustering method with key influencing factors	0.1585

Table 7a
Prediction error of different models

Name of model	Average relative error
BP neural network prediction model	0.5317
Single-factor prediction model based on traditional hierarchical clustering method	0.5127
Two-factor prediction model based on improved clustering method with key influencing factors	0.3545

Table 7b
Prediction error of different models

Name of model	Average relative error
BP neural network prediction model	0.4391
Single-factor prediction model based on traditional hierarchical clustering method	0.3747
Two-factor prediction model based on improved clustering method with key influencing factors	0.2725

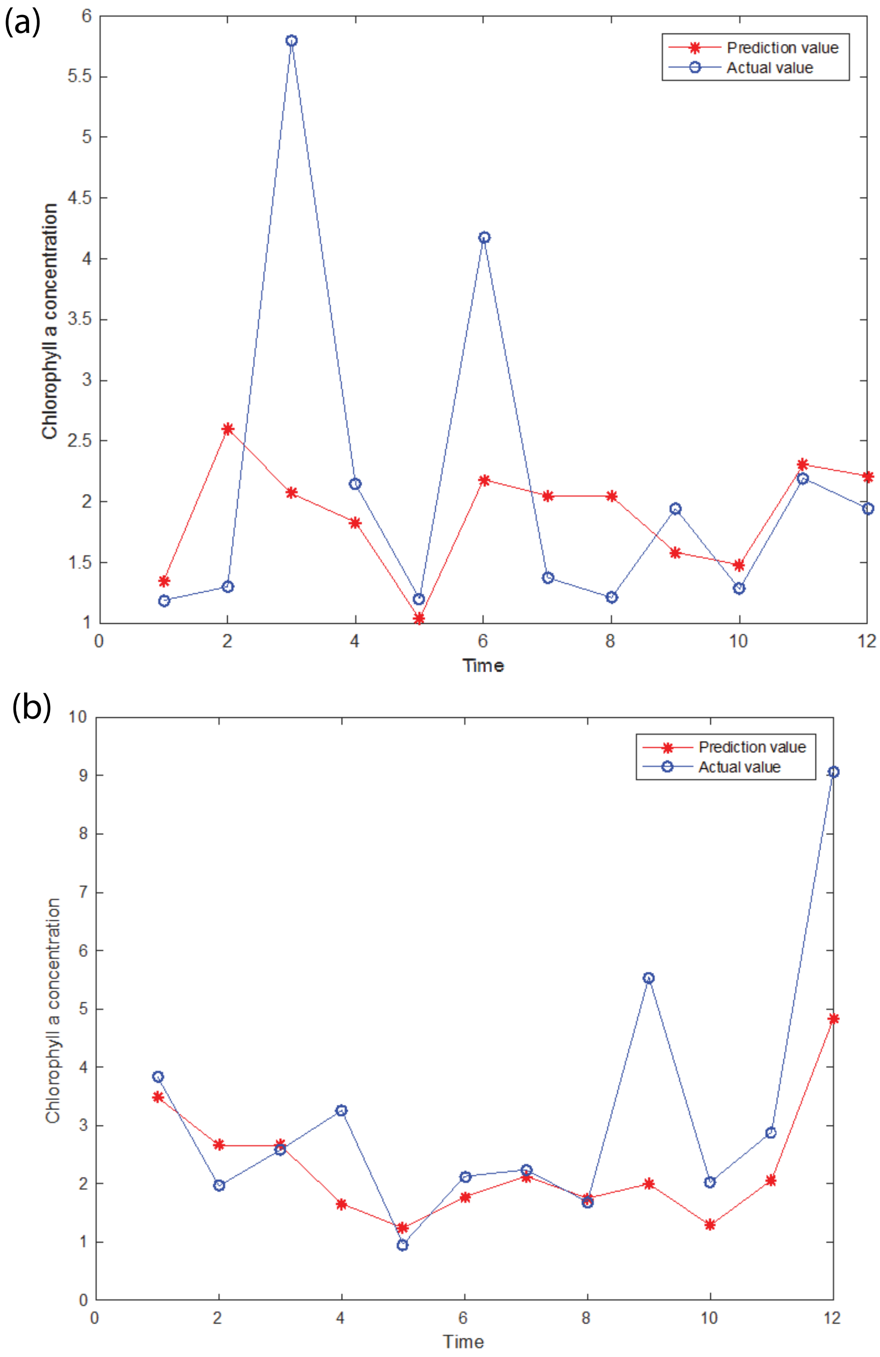

Fig. 2. Prediction results with key influencing factors for the GLEON data. Prediction results for (a) Tarpon Bay and (b) Oyster Bay.

wavelet threshold and threshold function. The result of this evaluations showed that it had good effectiveness and applicability for the pretreatment of data in cyanobacterial blooms in lakes and reservoirs.

- We determined the range of the embedding dimension and the delay time for each of the factors involved in modeling based on the phase-space reconstruction of single-factor time series. Then, we optimized the embedding dimension and delay time of the phase-space reconstruction of the multifactor time series using an iterative algorithm according to the C–C method, mutual information, and the minimum prediction error method.
- We used an improved hierarchical clustering method to determine the point sets in the phase-space reconstruction to prepare for the multifactor prediction of the subsequent modal division. This method was able to more flexibly adapt to the diversity and complexity of the trans-scale evolution model of the cyanobacteria blooms

in lakes and reservoirs. The proposed comprehensive characteristic scale function was also more flexible than the Euclidean distance to measure the intraclass distance.

- We used the first-order local linear regression algorithm to establish the local prediction model of multifactor chaotic time series after determining the local neighborhood of the prediction points.

For the sensitive characteristics of cyanobacteria blooms in lakes and reservoirs, we synthesized a new clustering measure scale. We explored a similarity identification of the cyanobacteria blooms in lakes and reservoirs. We studied the data from the perspective of data mining and found that the amount of data, the accumulated data period, and the sampling interval all had an important impact on model building. In the future, we will identify an appropriate sampling interval and time to capture information, combine the data-driven model with a better explanatory ability and predictive effect. The establishment of a prediction model for cyanobacterial blooms in lakes and reservoirs provides the basis for early warning decision-making.

Acknowledgments

We gratefully acknowledge Drs. Xianguo, Miansong, and Li Zhao for their considerable help for the acquisition of data and for providing historical experience. The research was supported financially by the National Water Pollution Control and Management Technology Major Projects (No. 2017ZX07103-004). We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

References

[1] J. Mao, J. Lee, K. Choi, The extended Kalman filter for forecast of algal bloom dynamics, J. Water Res., 6 (2009) 513–517.

[2] M. Rowe, E. Anderson, T. Wynne, R. Stumpf, D. Fanslow, K. Kijanka, H. Vanderploeg, J. Strickler, T. Davis, Vertical distribution of buoyant Microcystis blooms in a Lagrangian particle tracking model for short-term forecasts in Lake Erie, J. Geophys. Res., 7 (2016) 5296–5314.

[3] L. Wang, T. Zhang, X. Jin, J. Xu, X. Wang, H. Zhang, J. Yu, Q. Sun, Z. Zhao, L. Zheng. Multi-factor nonlinear time-series ecological modelling for algae bloom forecasting, Desal. Water Treat., 122 (2018) 91–99.

[4] J. Deng, H.W. Paerl, B. Qin, Y. Zhang, G. Zhu, E. Jeppesen, Y. Cai, H. Xu. Climatically-modulated decline in wind speed may strongly affect eutrophication in shallow lakes, Sci. Total Environ., 645 (2018) 1361–1370.

[5] J. Mcgowan, E. Deyle, H. Ye, M. Carter, Predicting coastal algal blooms in southern California, Ecology, 98 (2017) 1419–1433.

[6] D. Obenour, A. Gronewold, C. Stow, D. Scavia, Using a Bayesian hierarchical model to improve Lake Erie cyanobacteria bloom forecasts, Water Resour. Res., 50 (2014) 7847–7860.

[7] Y. Kim, H. Shin, J. Plummer, A wavelet-based autoregressive fuzzy model for forecasting algal blooms, Environ. Modell. Software, 62 (2014) 1–10.

[8] X. Bai, H. Zhang, X. Wang, L. Wang, J. Xu, J, Yu, The adaptive-clustering and error-correction method for forecasting cyanobacteria blooms in lakes and reservoirs, Adv. Math. Phys., 7 (2017) 1–7.

[9] J. Shin, S. Yoon, Y. Cha, Prediction of cyanobacteria blooms in the lower Han River (South Korea) using ensemble learning algorithms, Desal. Water Treat., 84 (2017) 31–39.

[10] G. Lee, J. Bae, S. Lee, M. Jang, H. Park, Monthly chlorophyll-a prediction using neuro-genetic algorithm for water quality management in lakes, Desal. Water Treat., 57 (2016) 26783–26791.

[11] M. Ghorbania, R. Khatibic, A. Mehrd, H. Asadi, Chaos-based multigene genetic programming: a new hybrid strategy for river flow forecasting, J. Hydrol., 562 (2018) 455–467.

[12] W. Pan, C. Wu, Z. Li, M. Li, Prediction of self-heating process of sulfide ore heap using trend and chaos prediction model, J. Cent. South Univ., 3 (2015) 901–907.

[13] T. Kutser, Quantitative detection of chlorophyll in cyanobacterial blooms by satellite remote sensing, Limnol. Oceanogr., 49 (2004) 2179–2189.

[14] L. Lu, W. Jin, X. Wang, Non-local means image denoising with a soft threshold, IEEE Signal Process Lett., 22 (2015) 833–837.

[15] S. Paris, P. Kornprobst, J. Tumblin, Bilateral filtering, Int. J. Numer. Methods Eng., 63 (2009) 1911–1938.

[16] Y. Chen, P. Luh, C. Guan, Y. Zhao, L. Michel, M. Coolbeth, P. Friendland, S. Rourke, Short-term load forecasting: similar day-based wavelet neural networks, IEEE Trans. Power Syst., 25 (2008) 322–330.

[17] R. Fang, J. Zhou, Probabilistic interval forecasting of short-term load on the basis of clustering algorithm and Chaos theory, Power. Syst. Technol., 34 (2010) 70–76.

[18] L. Dong, L. Wang, S. Khahro, S. Gao, X. Liao, Wind power day-ahead prediction with cluster analysis of NWP, Renewable Sustainable Energy Rev., 60 (2016) 1206–1212.