Environmental Science and Pollution Research (2024) 31:262–279
https://doi.org/10.1007/s11356-023-31148-6

RESEARCH ARTICLE

A novel hybrid model based on two-stage data processing and machine learning for forecasting chlorophyll-a concentration in reservoirs

Wenqing Yu[1] · Xingju Wang[1] · Xin Jiang[2] · Ranhang Zhao[1,3] · Shen Zhao[2,4]

Received: 5 September 2023 / Accepted: 17 November 2023 / Published online: 28 November 2023

Abstract

The accurate and efficient prediction of chlorophyll-a (Chl-a) concentration is crucial for the early detection of algal blooms in reservoirs. Nevertheless, predicting Chl-a concentration in multivariate time series poses a significant challenge due to the complex interrelationships within the aquatic environment and the discrete and non-stationary nature of online monitoring of water quality data. To address the aforementioned issue, this paper proposes a novel prediction model named SGMD-KPCA-BiLSTM (SKB) for predicting Chl-a concentration. The model combines two-stage data processing and machine learning (ML). To capture nonlinear relationships in multivariate time series data, the optimal data subset is determined by combining symplectic geometry mode decomposition (SGMD) and kernel principal component analysis (KPCA). This subset is then input into a bidirectional long short-term memory (BiLSTM) model, and the model's hyperparameters are optimized using the sparrow search algorithm (SSA) to improve the accuracy of predictions. The performance of the model was evaluated at Qiaodian Reservoir in Shandong, China. To assess its superiority, the evaluation criteria included the root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE), coefficient of determination (R^2), frequency histograms of the prediction error, and the Taylor diagram. The prediction performance of five single models, namely the back-propagation (BP) neural network, support vector regression (SVR), long short-term memory (LSTM), convolutional neural network with long short-term memory (CNN-LSTM), and BiLSTM, as well as three hybrid models, namely SGMD-LSTM, SGMD-KPCA-LSTM, and SGMD-BiLSTM, were compared against the SKB model. The results demonstrated that the SKB model performs best in predicting Chl-a concentration ($R^2 = 96.19\%$, RMSE = 1.05, MAE = 0.65, MAPE = 0.08). It significantly reduced the prediction error compared to other models for comparison. Furthermore, the multi-step predictive capabilities of the SKB model are also discussed. The analysis shows a decline in predictive performance with larger prediction time steps, and the SKB model exhibits slightly superior performance compared to the other model at corresponding prediction intervals. The model has significant advantages in terms of its ability to accurately predict the non-smooth and nonlinear Chl-a sequences observed by the online monitoring system. This study presents a potential solution for controlling and preventing reservoir eutrophication, as well as an innovative approach for predicting water quality.

Keywords Prediction · Chlorophyll-a · Eutrophication · Symplectic geometry mode decomposition · Kernel principal component analysis · Bidirectional long short-term memory

Responsible Editor: Marcus Schulz

Highlights
- A hybrid prediction model for Chl-a concentration based on two-stage data processing and machine learning was developed.
- The two-stage data processing method efficiently selects predictive data and captures nonlinear correlations in water quality monitoring data.
- Compared to the other common models, the new model shows superior performance in forecasting the concentration of Chl-a in waters ($R^2 = 96.19\%$).

Extended author information available on the last page of the article

Introduction

Reservoirs play an irreplaceable role in both the natural ecosystem and human social life (Glasgow et al. 2004). As industrial and agricultural modernization has advanced rapidly over the past few decades, there has been a significant increase in the influx of nutrients, such as nitrogen and phosphorus, into the reservoir through runoff. Furthermore, due to the long hydraulic retention time, eutrophication has emerged as the main factor contributing to the deterioration

of water quality in reservoirs (Shi et al. 2023). The algal bloom caused by eutrophication seriously harms the ecological environment, which in turn jeopardizes the security of the water supply (Li and Li 2023; Niu et al. 2021). Chlorophyll-a (Chl-a) is a significant indicator of water quality, especially for evaluating eutrophication. It is consistently found in aquatic algal cells, and its concentration serves as an indicator of the amount of algae in water bodies (Boyer et al. 2009; Rakocevic-Nedovic and Hollert 2005). Algal blooms can be easily predicted by studying the concentration of Chl-a in water bodies (Dzurume et al. 2022). Therefore, accurately predicting the concentration of Chl-a in reservoirs is crucial. However, due to the complexity of the intrinsic mechanism of the eutrophication ecological process, the uncertainty factors in the water environment, and the limitations of water quality monitoring technology, establishing a highly accurate and stable prediction model for eutrophication is a significant challenge.

Currently, scholarly research on predicting algal dynamics can be categorized into two distinct approaches: process driven and data driven (Kerimoglu et al. 2018; Zhu et al. 2023). Process-driven models involve physical, chemical, and biological processes in the growth of algae, resulting in a complex model structure with numerous parameters. These characteristics limit the use of mechanistic models and reduce their overall applicability. With the advent of the big data era and the development of artificial intelligence (AI) technology, data-driven models have become popular tools (Hejazi and Cai 2009). In addition to traditional water quality monitoring, the utilization of online monitoring systems, remote sensing, and other emerging technologies has increased the availability of data. The data-driven method has been widely used in predicting algae dynamics (Lee et al. 2003; Pepe et al. 2001). Data-driven methods can be further refined into empirical models and ML-based models (Alexakis et al. 2013). Empirical models, such as logistic regression (LR), generalized additive models (GAM), and autoregressive integrated moving average (ARIMA), have been widely used in water quality prediction (Carvalho et al. 2011; Mohebzadeh et al. 2020; Myronidis et al. 2018). However, these empirical models cannot accurately capture the nonlinear characteristics of fluctuations in water quality parameters, resulting in higher prediction errors (Xie et al. 2019). ML-based models can effectively capture the nonlinear characteristics of data, allowing them to overcome certain limitations of empirical models. For example, Cho et al. (2014) used an artificial neural network (ANN) to predict the Chl-a concentration in Juam Lake. They discovered that the concentration of Chl-a was primarily influenced by environmental factors such as total organic carbon (TOC), pH, and water temperature. Park et al. (2015) used support vector machines (SVM) and ANN to predict Chl-a concentration. The SVM model demonstrated superior predictive performance in estimating Chl-a concentration compared to the ANN model. Lee and Lee (2018) conducted a comparative analysis of multilayer perceptron (MLP), recurrent neural network (RNN), and long short-term memory (LSTM) models for predicting harmful algal blooms in four rivers located in South Korea. The findings of this investigation showed that deep learning models, including MLP, RNN, and LSTM, exhibited superior predictive capabilities compared to the conventional ordinary least square simple linear regression method. Wang and Xu (2020) proposed a novel spatio-temporal distribution model that utilizes LSTM to predict the future trend of Chl-a concentration. The validation results yielded a mean square error (MSE) of 0.7778 and a root mean square error (RMSE) of 1.201. Unfortunately, the applicability of the individual ML-based models used in these studies is limited in complex water environments. Additionally, the performance of the models is significantly affected by the training samples, resulting in considerable uncertainty regarding their performance.

ML-based prediction models have been widely used to predict Chl-a concentration. However, the dependability of the input variables limits the predictive accuracy of ML models. The discrete and non-stationary nature of online monitoring of water quality data related to algae may limit the model's ability to accurately capture the dynamic trends of algae. Hybrid models can synthesize the advantages of each algorithm. Through careful data processing, the hidden information within the data can be fully extracted, thereby improving the predictive capability of the model (Liu et al. 2023; Yu et al. 2020). Recently, many scholars have conducted extensive research on the "decomposition-prediction-reconstruction" method. Firstly, the sequence is decomposed using decomposition technology. Subsequently, the appropriate model is selected to predict each component. Finally, the reconstructed prediction results are obtained to enhance the prediction performance of a single model (Tong et al. 2019; Zhang et al. 2023b). Decomposition algorithms are an innovative strategy for preprocessing in the ML modeling process. The essence of decomposition is to convert non-stationary time series data into stationary data. The usefulness of extracting dynamic features from time series data has been demonstrated in previous studies (Lu and Ma 2020; Wang et al. 2023). These algorithms have been widely used to predict hydrological and meteorological parameters, including wind speed, rainfall, and floods (Antico et al. 2014; Gao et al. 2020; Zhang et al. 2017). Empirical mode decomposition (EMD), ensemble empirical mode decomposition (EEMD), and wavelet transform (WT) are frequently used decomposition algorithms (He et al. 2019). Nevertheless, EMD suffers from issues such as redundant and insufficient decomposition. As an improved method of EMD, EEMD can mitigate the impact of the modal aliasing phenomenon. However, this method can easily lead to

non-convergence of the function (Xie et al. 2019). WT uses multiple wavelet functions, which complicates the selection process (Hadi and Tombul 2018). To overcome these limitations, a novel adaptive time–frequency decomposition method called symplectic geometry modal decomposition (SGMD) has been introduced for prediction purposes. However, there are few applications of SGMD to water quality time series data, and the effectiveness of applying the "decomposition-prediction-reconstruction" methodology in water quality prediction deserves further investigation.

The primary objectives of this study are as follows: (1) To efficiently and accurately extract valuable information from large volumes of water quality monitoring data, considering the complexity and multiple correlations of the water environment. The appropriate input variables are selected to achieve accurate predictions of multivariate time series. (2) A new two-stage data processing method is proposed to address the nonlinearity and instability of online water quality monitoring data in multivariate time series. This method demonstrates the effectiveness of efficiently selecting a subset of prediction data while also accurately capturing the nonlinear relationship present in water quality monitoring data. (3) A new hybrid prediction model has been developed to overcome the limitations of using a single model. Furthermore, an intelligent optimization algorithm is applied to optimize the hyperparameters of the model, aiming to enhance the performance of the prediction model. In this paper, we construct a short-term prediction model that combines two-stage data processing and machine learning (ML) techniques. The model utilizes SGMD, kernel principal component analysis (KPCA), sparrow search algorithm (SSA), and bidirectional long short-term memory (BiLSTM). Furthermore, the data obtained from the real-time online monitoring system of the reservoir is used to validate the accuracy of the model in predicting the temporal variations of Chl-a concentrations within the reservoir.

Materials and methods

Study area

This study used the Qiaodian Reservoir as an example to assess the effectiveness of the method. Qiaodian Reservoir is situated in Jinan, Shandong Province, China, along the Xinzhuang River, which is a tributary of the Mouwen River. The location is situated at the global geographical coordinates of 117°51′34″ east longitude and 44°23′40″ north latitude (see Fig. 1). The dam was constructed in 1965 and underwent repairs in 2005 to reinforce its structure. The reservoir controls a watershed area of 85 km^2, with a total capacity of 27.99 million m^3. The reservoir provides an average annual water supply of 10 million m^3. It is a medium-sized reservoir used for water supply, flood control, agricultural irrigation, and power generation, among other purposes. This significant water source was recognized as an important national drinking water source in 2016. The water quality of the reservoir is relatively exceptional, making it a valuable source of drinking water in Jinan City.

The Qiaodian Reservoir is located in the temperate monsoon climate zone, which is characterized by four distinct seasons, significant annual temperature variations, and uneven distribution of precipitation throughout the year. The annual average temperature is 15.2 °C, based on data from the period 2005–2022. The warmest period of the year occurs between June and September, when the average highest temperature exceeds 35 °C. The average duration of sunlight in the reservoir over several years is 2296.71 h,

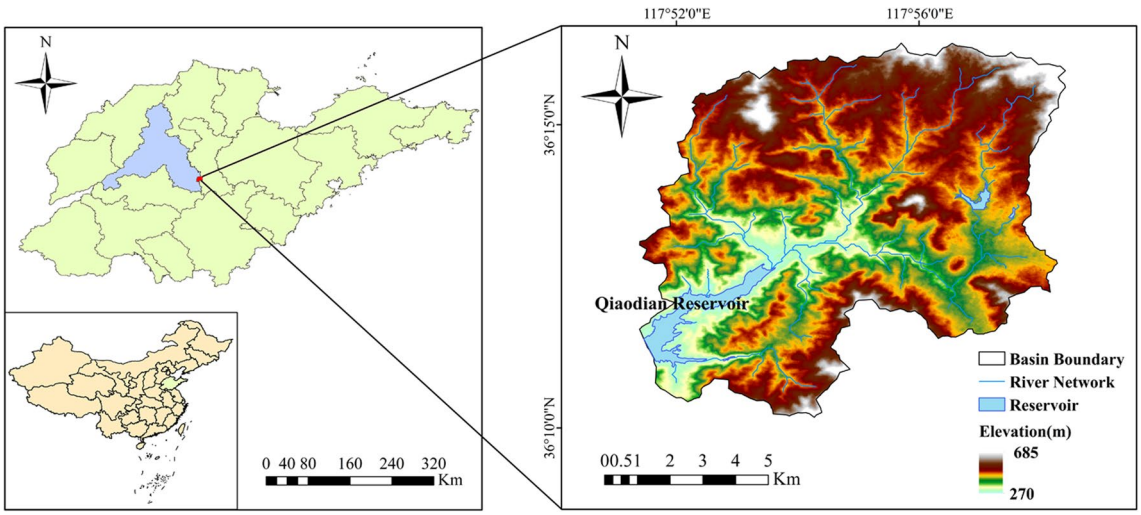

Fig. 1 Study area and location of Qiaodian Reservoir

and the average wind speed for many years is 2.6 m/s. The annual average precipitation in the reservoir basin is 683.3 mm, with the majority of rainfall occurring during the flood season (July–September).

In recent years, there has been an emergence of cyanobacterial blooms in the Qiaodian Reservoir. This poses a potential hazard to the water quality and safety of urban water supply systems. Even though several scientific prevention and control measures were implemented to reduce pollution and limit discharge, signs of localized cyanobacteria growth were still observed in the reservoir. The risk of eutrophication cannot be ignored.

Dataset collection and selection

The phenomenon of reservoir eutrophication arises from the synergistic impacts of multiple environmental conditions (Chen et al. 2011; Gentine et al. 2022). Appropriate meteorological conditions, sufficient nutrient levels, and proper hydraulic conditions all contribute to the increase in Chl-a content in water (Wu et al. 2014). Therefore, before predicting Chl-a concentration, it is important to identify and screen the main driving factors of Chl-a to eliminate any interference from irrelevant factors. To eliminate the interference of irrelevant factors and optimize the selection of input variables, the gray correlation analysis method and Pearson correlation analysis were employed to filter out extraneous variables.

The research data were collected from daily water quality monitoring data spanning from January 2019 to December 2022, which were obtained through the online monitoring system of the Qiaodian Reservoir. The automatic monitoring station is located 20 m away from the water intake, and the water point is positioned approximately 3 m below the water surface. The water quality indicators involved included water temperature (WT), pH, turbidity (TD), electrical conductivity (EC), permanganate index (COD_{Mn}), ammonia nitrogen (NH_3-N), and Chl-a. Furthermore, considering the interdependence between the growth and reproduction of algae and meteorological factors, the research data for this study included daily meteorological data from January 2019 to December 2022. This dataset includes variables such as air temperature (T), atmospheric pressure (P), wind speed (WS), sunshine hours (SUN), and precipitation (PRCP).

The magnitude of the gray correlation degree directly reflects the level of correlation between the two sequences. The strength of the association between the two variables is directly proportional to the magnitude of the correlation coefficient. The Chl-a concentration monitoring data was used as the reference sequence, while the other 11 groups of water quality monitoring data were used for comparison. According to the steps for calculating the degree of correlation, the gray correlation between Chl-a and the other 11 indexes was analyzed and calculated. The specific analysis steps are as follows:

Step 1: Determine the reference sequence and standardize it. Let the reference sequence $X_0=\{X_0(k)|k=1,2,\cdots n\}$ and the comparison sequence $X_i=\{X_i(k)|k=1,2,\cdots n\}$,$i=1,2,\cdots m$. Due to the use of different units in water quality monitoring data, it is easy to introduce errors in the analysis results. Therefore, the mean values of 12 groups of water quality monitoring data are standardized.

$$\begin{cases} x_i(k) = X_i(k)/X_i(l) \\ X_i(l) = \frac{1}{n}\sum_{k=1}^{n} X_i(k) \end{cases} \tag{1}$$

Step 2: Calculate the correlation coefficient $\xi_i(k)$.

$$\xi_i(k) = \frac{\min\limits_i \min\limits_k |x_0(k) - x_i(k)| + \rho \max\limits_i \max\limits_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \max\limits_i \max\limits_k |x_0(k) - x_i(k)|} \tag{2}$$

where ρ is the resolution coefficient within [0, 1], generally, $\rho = 0.1$.

Step 3: Calculate the correlation coefficient $r_i(k)$.

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \xi_i(k) \tag{3}$$

The correlation degree values of 11 groups of comparison sequences and the reference sequence Chl-a were sorted. The results are shown in Table 1. The variables that exhibited a strong positive correlation with Chl-a concentration were identified as follows: pH, electrical conductivity (EC), turbidity (TD), atmospheric pressure (P), permanganate index (COD_{Mn}),

Table 1 Result of gray relational degree

Serial number	Factors	Gray correlation degree
1	pH	0.8205
2	EC	0.8176
3	TD	0.8176
4	P	0.8155
5	COD_{Mn}	0.8149
6	WS	0.8149
7	SUN	0.8124
8	T	0.8083
9	WT	0.7987
10	NH_3-N	0.7986
11	PRCP	0.7751

wind speed (WS), sunshine hours (SUN), and air temperature (T).

The Pearson correlation analysis method was chosen to identify and eliminate duplicate components that contain overlapping information. Figure 2 shows the results of the univariate correlation analysis among the indicators. Except for NH_3-N and WS, all other variables showed a significant correlation with Chl-a concentration ($P < 0.05$). The correlation coefficients among the three environmental factors, T, WT, and P were high. These factors could be included or excluded, depending on the circumstances. It is important to acknowledge that nitrogen, a significant factor in algae growth, was not initially considered one of the primary factors associated with Chl-a concentration. The omission may be attributed to the stable and low concentration of NH_3-N observed throughout the monitoring period.

Based on the aforementioned analyses, the water quality prediction model incorporates T, WS, SUN, pH, EC, TD, COD_{Mn}, and Chl-a as variables. The data was preprocessed and then divided into training and test sets. The selected indicators were used as input variables to train the BiLSTM network, with the daily variation in Chl-a designated as the output variable.

Framework of the prediction model

The proposed Chl-a prediction system mainly consists of data processing and time series prediction, as shown in Fig. 3. It can be decomposed into seven steps. Step 1 involves data preparation. Step 2 and step 3 represent the two stages of data processing, respectively. At the end of step 3, the optimal data subset can be selected. The remaining four steps are structured as time series prediction modules. In step 4, the hyperparameters of the BiLSTM model are optimized using the SSA algorithm. Step 5 represents the training set as the input to the BiLSTM. Step 6 outputs predicted values, and step 7 evaluates the model's performance in making predictions. The specific steps are as follows:

Step1: Constructing the input dataset. The input variables were screened using Chl-a influence factor analysis. Each variable was checked for outliers, and missing values were interpolated to obtain a complete time series dataset.

Step2: Sequence decomposition using SGMD. The SGMD algorithm is applied to decompose the data processed in step 1 into multiple symplectic geometric modal components and residual components. It adaptively decomposes and reconstructs the single-component signal while preserving the original time series.

Step3: Data dimensionality reduction using KPCA. Since the number of dimensions of the input variables decomposed in step 2 is too large, KPCA dimensionality reduction is performed on the data. On the basis of ensuring the preservation of accurate information in the data, the correlation and redundancy of various time series data are eliminated. This is done to prevent the issue of overfitting in the prediction model and improve computational efficiency.

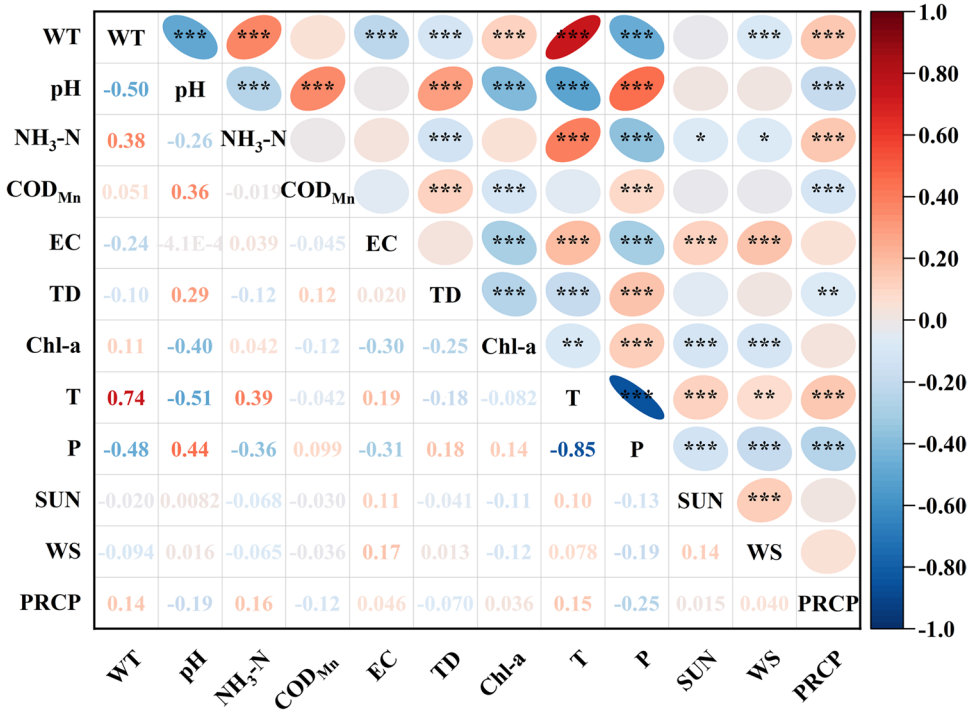

Fig. 2 Correlation analysis among the factors. Note: *correlation is significant at 0.05; **the correlation is significant at 0.01; ***the correlation is significant at 0.001

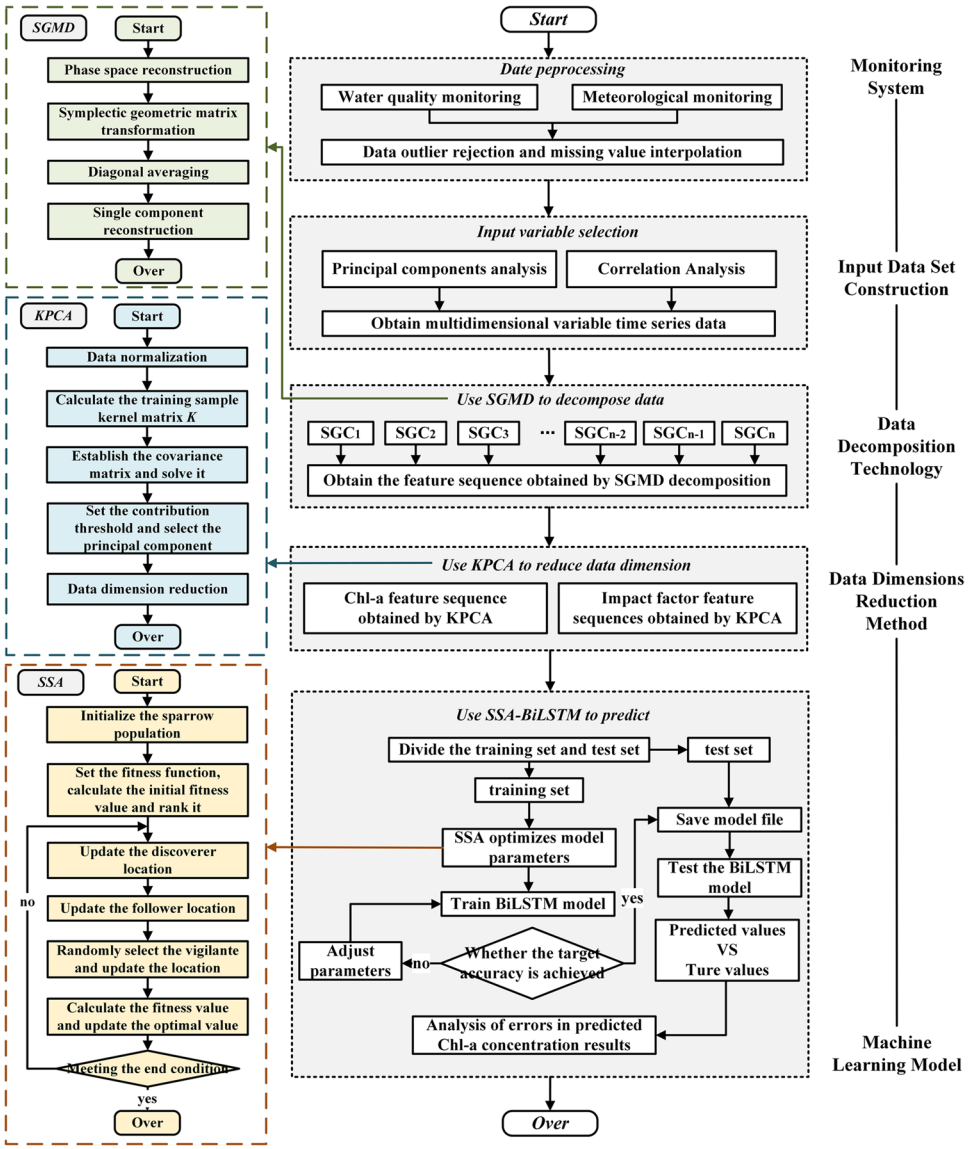

Fig. 3 A flowchart diagram of the process of model construction

Step4: BiLSTM network training. To mitigate significant data fluctuations during the training phase, the data is normalized and converted into a suitable format for training the BiLSTM network. This is done after reducing the dimensionality in step 3. The training set and test set are then divided.

Step5: Network parameter optimization using SSA. The SSA is used to optimize the hyperparameters of the BiLSTM model. The optimization parameters include the number of hidden layer nodes, the initial learning rate, and the regularization coefficient.

Step6: Sequence prediction. After training the network model in step 4, the test set is used for evaluation.

Step7: Validation of the prediction model. The predicted values obtained in step 6 are compared with the actual values to calculate errors and verify the model's predictive performance. The selected evaluation indicators include the root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE), and coefficient of determination (R^2). The formulas for each indicator are as follows:

$$E_{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i'\right)^2} \tag{4}$$

$$E_{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i - y_i'\right|}{y_i} \times 100\% \tag{5}$$

$$E_{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - y_i'\right| \tag{6}$$

$$R^2 = 1 - \frac{\sum_i \left(\hat{y}_i - y_i\right)^2}{\sum_{i=1} \left(\overline{y}_i - y_i\right)^2} \times 100\% \tag{7}$$

where N is the length of the predicted time sequence, y_i and y_i' are the actual and predicted values of the sequence at the ith moment, respectively.

Two-stage data processing method

A novel two-stage data processing method is proposed to address the nonlinearity and instability of online water quality monitoring data in multivariate time series. Specifically, the SGMD decomposition algorithm captures the nonlinearities in the series, improving the smoothness and predictability of the time series. Nonetheless, it increases the dimensionality of the input variables in the model. The KPCA method is used to reduce the dimensionality of the input variables. The model's computational efficiency and accuracy are improved, while also ensuring the validity and representativeness of the information.

Symplectic geometry mode decomposition

Symplectic geometric mode decomposition (SGMD) is a relatively new method for decomposing modes that can eliminate noise interference while preserving the characteristics of the original time series. It is suitable for analyzing nonlinear and unstable time series. Compared to empirical mode decomposition (EMD), this method can avoid modal aliasing and sensitivity to parameter selection (Pan et al. 2019). The steps for SGMD decomposition are as follows:

(1) Phase space reconstruction
Let the original signal time series be denoted as $x=(x_1, x_2, \ldots, x_n)$. According to the Takens embedding theorem, the trajectory matrix X is defined by Eq. (8).

$$X = \begin{bmatrix} x_1 & x_{1+\tau} & \cdots & x_{1+(d-1)\tau} \\ x_2 & x_{2+\tau} & \cdots & x_{2+(d-1)\tau} \\ \vdots & \vdots & \ddots & \vdots \\ x_m & x_{m+\tau} & \cdots & x_{m+(d-1)\tau} \end{bmatrix} \tag{8}$$

where d and τ represent the embedding dimension and the delay time, and $m=n$—$(d$—$1)\ \tau$.
(2) Symplectic geometric matrix transformation
Let the covariance symmetric matrix $A=X^TX$, the Hamiltonian matrix M is obtained using Eq. (9).

$$M = \begin{bmatrix} A^T & 0 \\ 0 & -A \end{bmatrix} \tag{9}$$

Let $F = M^2$, then F is also a Hamiltonian matrix, and the symplectic orthogonal matrix Q is constructed.

$$Q^T F Q = \begin{bmatrix} B & R \\ 0 & B^T \end{bmatrix} \tag{10}$$

where B and R represent the upper triangular matrix and the submatrix after matrix transformation, the eigenvalues of the matrix B are $\lambda_1, \lambda_2, \cdots, \lambda_d$.
Let the eigenvalues of matrix A be σ_i, $\sigma_i=\sqrt{\lambda_i}$(i= 1, 2, ⋯ ,d), and the corresponding eigenvectors of matrix A be Q_i(i= 1, 2, ⋯ ,d). The reconstructed trajectory matrix Z is constructed from a series of initial single-component matrices Z_i(i= 1, 2, ⋯ ,d), the matrix Z is obtained using Eq. (11).

$$Z = Z_1 + Z_2 + \cdots + Z_d \tag{11}$$

In Eq. (11), $Z_i = Q_i S_i,\ S_i = Q_i^T X^T$.
(3) Diagonal averaging
Since the reconstructed single-component matrix Z_i is an $m \times d$ matrix, it is necessary to transform the single-component matrix $Z_i(1 \le i \le d)$ into a time series of length n. The sum of d sets of time series of length n should equal the original time series signals. Let the elements of matrix Z_i be $z_{ij}(1 \le i \le d, 1 \le j \le m)$, if $m<d$, then $z_{ji}^*=z_{ij}$, otherwise, $z_{ij}^* = z_{ji}$. The formula for diagonal averaging is as follows:

$$y_k = \begin{cases} \frac{1}{k}\sum_{p=1}^{k} z_{p,k-p+1}^* & 1 \le k < d^* \\ \frac{1}{d}\sum_{p=1}^{d^*} z_{p,k-p+1}^* & d^* \le k < m^* \\ \frac{1}{n-k+1}\sum_{p=k-m^*+1}^{n-m^*+1} z_{p,k-p+1}^* & m^* \le k \le n \end{cases} \tag{12}$$

In Eq. (12),
$d^*=min(m, d)$, $m^* = max(m, d)$, $n = m+(d-1)\tau$.
The time series $Y_i(y_1, y_2, \cdots, y_n)$ corresponding to $Z_i(1 \le i \le d)$ is obtained from Eq. (12), and the matrix Y of $d \times n$ is obtained by averaging the individual reconstruction matrices diagonally.

$$Y = Y_1 + Y_2 + \cdots + Y_d \tag{13}$$

(4) Single-component reconstruction

The initial d symplectic geometric modal components obtained from the decomposition can be used to reconstruct components that exhibit a high degree of similarity, based on the similarity criterion.

Kernel principal component analysis

Kernel principal component analysis (KPCA) is a dimensionality reduction algorithm suitable for processing linearly inseparable data (Chen et al. 2019; Zhou and Peng 2020). By employing the kernel method to map all samples in the input space to a high-dimensional space, KPCA achieves linear separability of data. It then applies

principal component analysis (PCA) for linear dimensionality reduction in the high-dimensional space, aiming to preserve the nonlinear information of the data to the greatest extent possible. This approach offers several advantages over PCA, including the ability to obtain an accurate covariance matrix and effectively process nonlinear data. (Wang et al. 2021). The main steps of the KPCA algorithm are as follows (Zhang et al. 2023a):

(1) Construct the covariance matrix C.
Suppose the sample x_k undergoes a nonlinear transformation into $\varphi(x_k)$, which is then mapped to the high-dimensional feature space F. Subsequently, the covariance matrix C is constructed.

$$C = \frac{1}{n}\sum_{j=1}^{n}\varphi\left(x_j\right)\varphi\left(x_j\right)^T \tag{14}$$

(2) Calculate the eigenvalue λ and the eigenvector V.
The eigenvalues λ and the eigenvector V should satisfy $\lambda V = CV$, and by introducing the nonlinear function $\varphi(x_k)$, the eigenvector V can be represented linearly using $\varphi(x_k)$.

$$V = \sum_{i=1}^{m}\alpha_i\varphi\left(x_j\right) \tag{15}$$

(3) Calculate the kernel matrix K.
By introducing the kernel function $K_{ij}=\varphi(x_i)\varphi(x_j)$, the following formula can be obtained.

$$n\lambda\alpha - K\alpha = 0 \tag{16}$$

where α represents the eigenvector of the kernel matrix K.
(4) Select the principal component.

The projection of any sample onto the principal component $\varphi(x)$ in the feature space F can be expressed using the following equation.

$$V\varphi(x) = \sum_{i=1}^{m}\alpha_i\varphi\left(x_j\right)\varphi(x) = \sum_{i=1}^{m}\alpha_i K(x_i, x) \tag{17}$$

Select the principal component whose cumulative contribution exceeds the specified threshold to satisfy the following conditions.

$$\frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{m}\lambda_k} \geq d \tag{18}$$

where p represents the number of principal components that satisfy the condition, and d represents the specified threshold of cumulative contribution, typically with $0.8 \leq d \leq 0.95$.

If the assumption of $\sum_{i=1}^{n}\phi\left(x_k\right) = 0$ is not satisfied, replace K with $\widetilde{K}$:

$$\widetilde{K} = K - SK - KS + SKS \tag{19}$$

where S is an $n \times n$ order unit matrix with a coefficient of $1/n$.

Machine learning prediction model

Sparrow search algorithm

The sparrow search algorithm (SSA) has robust stability, rapid convergence, and effective global search capabilities (Yu et al. 2022).

(1) The formula for updating the discoverer's location

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^{t}\cdot\exp\left(-\frac{i}{\alpha\cdot i_{item,\max}}\right) & if\ \ R_2 < S_T \\ X_{i,j}^{t} + Q\cdot L & if\ \ R_2 \geq S_T \end{cases} \tag{20}$$

where t represents the current number of iterations, $i_{item,max}$ represents the maximum number of iterations, $X_{i,j}$ represents the position information of the ith sparrow in the jth dimension, α is a random number within [0, 1], R_2 is the early warning value within [0, 1], S_T is the safety value within [0.5, 1], Q is a random number obeying normal distribution, and L is a matrix of $1 \times d$ whose internal elements are all 1.
(2) The formula for updating the joiner's location

$$X_{i,j}^{t+1} = \begin{cases} Q\cdot\exp\left(\frac{X_{worst}-X_{i,j}^{t}}{i^2}\right) & if\ i < n/2 \\ X_p^{t+1} + \left|X_{i,j}^{t} - X_p^{t+1}\right|\cdot A^T\left(AA^T\right)^{-1}\cdot L & if\ otherwise \end{cases} \tag{21}$$

where X_p represents the current best position occupied by the discoverer, X_{worst} represents the current global worst position, and A is a matrix of $1 \times d$ whose elements are either -1 or 1.
(3) Assuming that 10–20% of the sparrows in the flock are aware of the danger, Eq. (22) indicates the location of these sparrows.

$$X_{i,j}^{t+1} = \begin{cases} X_{best}^{t} + \beta\left|X_{i,j}^{t} - X_{best}^{t}\right| & f_i > f_g \\ X_{i,j}^{t} + K\left(\frac{\left|X_{i,j}^{t} - X_{worst}^{t}\right|}{(f_i - f_w)+\varepsilon}\right) & f_i = f_g \end{cases} \tag{22}$$

where X_{best} represents the current best global position, β is a random number obeying a normal distribution with a mean value of 0 and a variance of 1, K is a random number within [0, 1], f_i is the fitness value of an individual

sparrow, f_g and f_w are the global optimal position and worst position, respectively, and ε is a constant.

Bidirectional long short-term memory network

The long short-term memory (LSTM) is a type of recurrent neural network (RNN) that incorporates a distinct gating mechanism and memory units, improving upon the conventional RNN architecture. By employing selective forgetting and selective memory connections, the LSTM effectively addresses problems related to long-term dependency, gradient descent, and gradient vanishing (Qin et al. 2019). Currently, the neural network in question has gained popularity and has shown superior performance (Cen et al. 2022; Shin et al. 2020). The architecture of LSTM comprises three fundamental components: the forgetting gate, the input gate, and the output gate.

(1) Forgetting gate. The sigmoid function is utilized to determine whether to retain or discard information at the output of the previous time step and at the input of the current time step. The calculation equation is as follows:

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{23}$$

where f_t is the output value of the forgetting gate, W_f is the weight matrix of the forgetting gate, σ is the sigmoid function, b_f is the bias term of the forgetting gate, h_{t-1} is the output value from the previous time step, and x_t is the input value at the current time step.

(2) Input gate. The information to be updated is determined using the sigmoid and tanh functions. The calculation equations are as follows:

After being filtered by the sigmoid function:

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{24}$$

After being filtered by the tanh function:

$$C_t' = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{25}$$

The formula for updating the unit status is as follows:

$$C_t = f_t C_{t-1} + i_t C_t' \tag{26}$$

where i_t is the output value of the input gate, W_i is the weight matrix of the input gate, b_i is the bias term of the input gate, C_t' is the current cell state information, W_c is the weight matrix of C_t', b_c is the bias term of C_t', C_t is the unit state at the current time step, and $C_{t\text{-}1}$ is the unit state at the previous time step.

(3) Output gate. Determine the information available in the current moment unit state C_t for the current moment output h_t.

The input gate information is determined using the sigmoid function, and the updated information is processed using the tanh function. These two results are then multiplied to calculate the current output value at the specified moment. The calculation equation is as follows:

$$o_t = \sigma\left(W_0\left[h_{t-1}, x_t\right] + b_0\right) \tag{27}$$

$$h_t = o_t \tanh(C_t) \tag{28}$$

where o_t is the output value of the output gate, W_o is the weight matrix of the output gate, b_o is the bias term of the output gate, and h_t is the output value at the current moment.

Bidirectional long short-term memory (BiLSTM) is a combination of forward LSTM and backward LSTM that can capture long-term dependencies while simultaneously processing information in both directions (Latifoglu 2022). The BiLSTM combination mechanism can effectively extract data features and fully leverage the temporal correlation between them. It has a strong capability to capture sequence correlations and make nonlinear predictions, providing a significant advantage in time series prediction (Ozdogan-Sarikoc et al. 2023). The main structure of the BiLSTM neural network model is shown in Fig. 4.

Results and discussion

Two-stage data processing

Decomposition of data by SGMD

The time series data for each input variable is decomposed using SGMD. Since using a small value of K for the decomposition layer of SGMD can result in under-decomposition of the data and negatively impact prediction accuracy, and using a large value can lead to repeated modes and introduce noise, it is important to test different values of K. After debugging, it has been determined that $K = 5$ is the optimal value. Taking the decomposition results of the Chl-a sequence data as an example, as shown in Fig. 5, the original Chl-a concentration sequence is decomposed into five components (SGC) spanning from high frequency to low frequency. As the volume of data increases, the high-frequency subsequence can capture the detailed signal and noise of the corresponding time, while the low-frequency subsequence can reveal hidden periodic oscillation trends. This is advantageous for the model because it helps accurately identify the internal transformation patterns of the data sequence, thereby improving the accuracy of predictions.

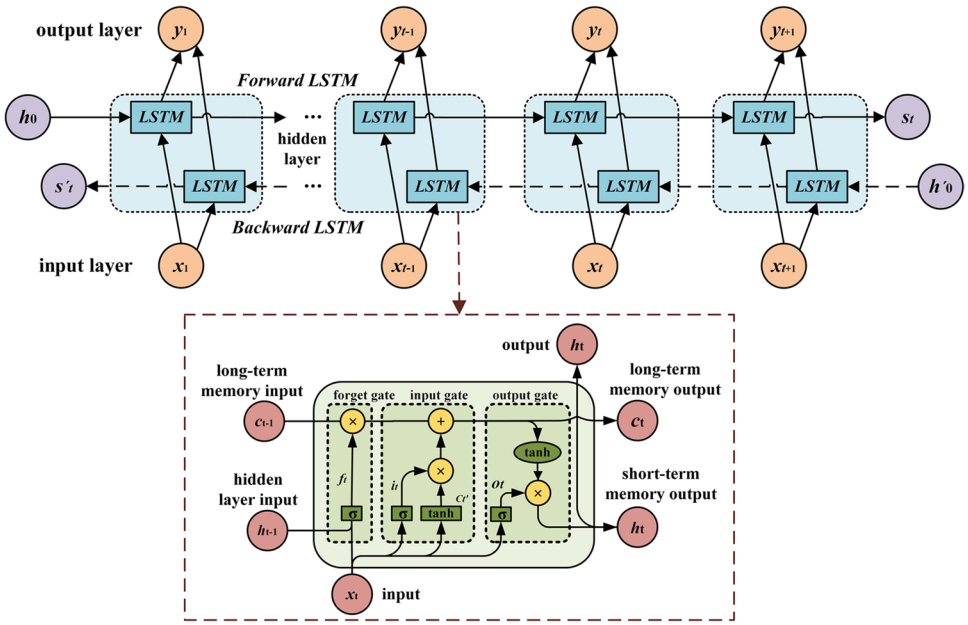

Fig. 4 Structure of BiLSTM network

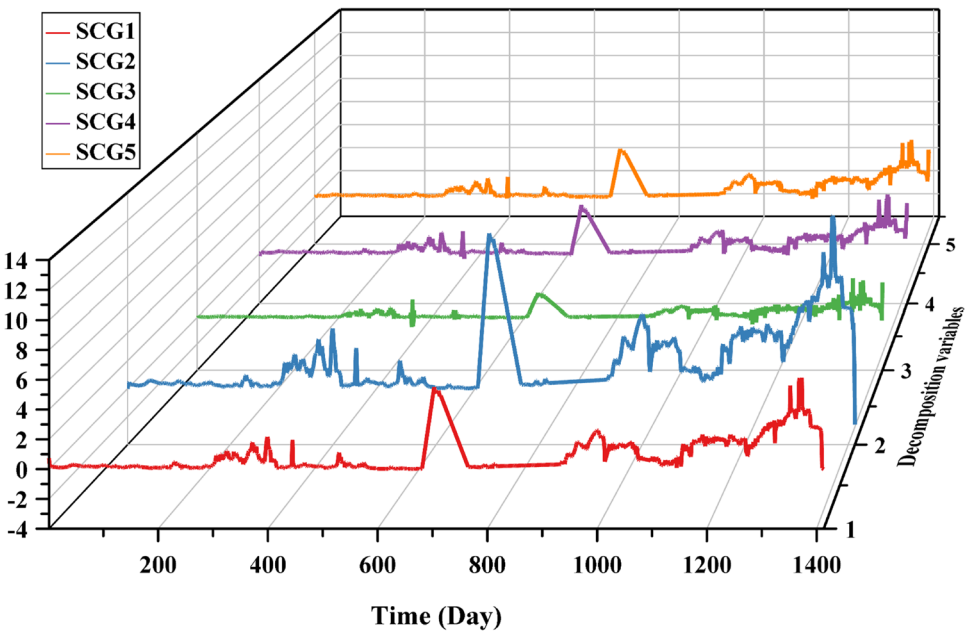

Fig. 5 Chl-a concentration decomposition results from 2019 to 2022 by SGMD

Dimensionality reduction of data by KPCA

Set the threshold for the cumulative contribution rate at 0.95. When the cumulative contribution of principal components exceeds the threshold, it indicates that these components already capture 95% of the original data. Therefore, these principal components can be extracted as the input parameters required to construct the model. When using KPCA to process the original data, Fig. 6 illustrates the cumulative contribution rate of each principal component and the corresponding change in the contribution rate. From Fig. 6, it can be observed that the contribution rate of principal component 1 exceeds 55%, indicating a significant portion of the cumulative contribution rate. By the time we reach principal component 6, the cumulative contribution rate has already reached 96.01%. The contribution rate of the subsequent principal components is negligible. Therefore, the first six principal components are selected as the input parameters for the prediction model, reducing the input data to six dimensions.

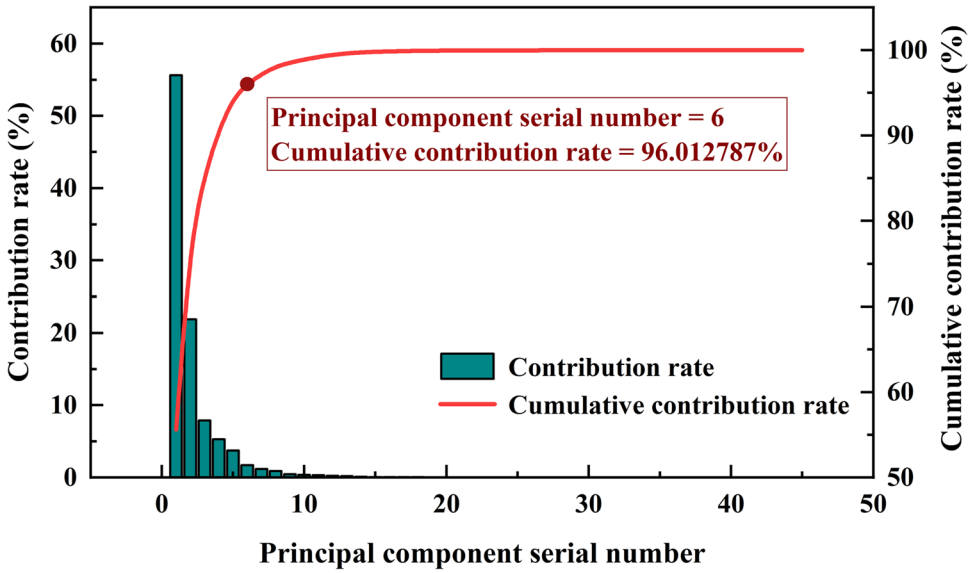

Fig. 6 Cumulative contribution rate of principal components and the change of contribution rate

Predictions obtained by BiLSTM

The feature sequences obtained by applying SGMD for data decomposition and KPCA for data dimensionality reduction are used as inputs for training and prediction in the BiLSTM network. The input and output data are normalized to eliminate the influence of dimensions and individual data samples. The training set consists of 60% of the total data, while the test set consists of the remaining 40%. The input layer dimension of the prediction model is 8, and the output layer dimension is 1. The input time steps correspond to the duration of the historical data sequence used for predictive purposes, and the prediction time step is 1. Furthermore, the selection of BiLSTM network parameters is crucial as it directly affects the accuracy of the model's predictions. SSA is used to optimize the three hyperparameters of the BiLSTM model, which include the number of hidden layer nodes (NHN), the initial learning rate (α), and the L2 regularization coefficient. In the SSA optimization parameters, the sparrow population size is set to 30, the maximum number of iterations is set to 10, the ratio of discoverers to joiners is set at 1:4, and the warning threshold is set to 0.8. The upper and lower limits of the three hyperparameter settings are shown in Table 2. The remaining primary parameters of the BiLSTM structure are selected as indicated in Table 3.

Figure 7 demonstrates the predictive effect of Chl-a concentration using the SKB model. From the linear fit plot of the predicted and true values on the test set, the R^2 value is 0.96, and the correlation coefficient is close to 1. This demonstrates that the predicted Chl-a concentration value of the SKB model closely aligns with the true value and exhibits a strong goodness of fit. To further test the accuracy of the prediction results, we calculated frequency histograms of the errors. The closer the histogram is to a normal distribution, the more stable the prediction result will be. As shown in Fig. 7, the predicted results demonstrate a symmetrical distribution on both sides of the central point. The histogram closely resembles a normal distribution, indicating that the established SKB model produces reliable results.

Table 2 The search range of SSA optimization algorithm for BiLSTM hyperparameters

	Number of hidden layer nodes	Initial learning rate	L2 regularization coefficient
Upper limit	100	0.002	1×10^{-2}
Lower limit	10	0.0001	1×10^{-10}

Model evaluation and comparison

Comparison of different prediction models

To validate the effectiveness and superiority of the SKB model, a total of eight alternative models were selected for comparison. These included five single models (BP, SVR, LSTM, CNN-LSTM, and BiLSTM) and three hybrid models (SGMD-LSTM, SGMD-KPCA-LSTM, and SGMD-BiLSTM). The test set was used to verify the model's prediction results, and several metrics were employed to evaluate and compare the model's

Table 3 BiLSTM structure parameters selection

Parameters	Hidden Layers	Optimizer	Batch_size	Epochs	Loss function	Activation function
Value	2	Adam	18	250	MSE	ReLU

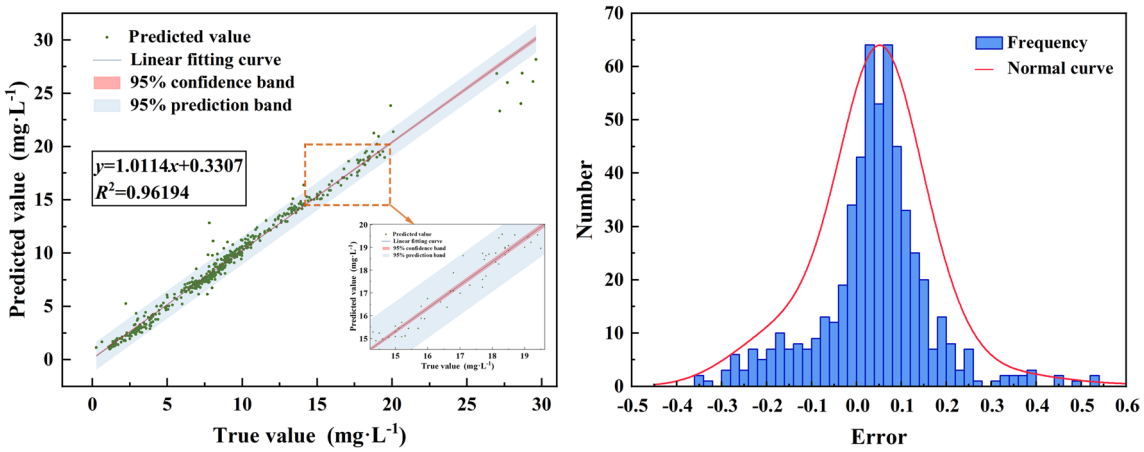

Fig. 7 Prediction effect for Chl-a concentrations of the SKB model

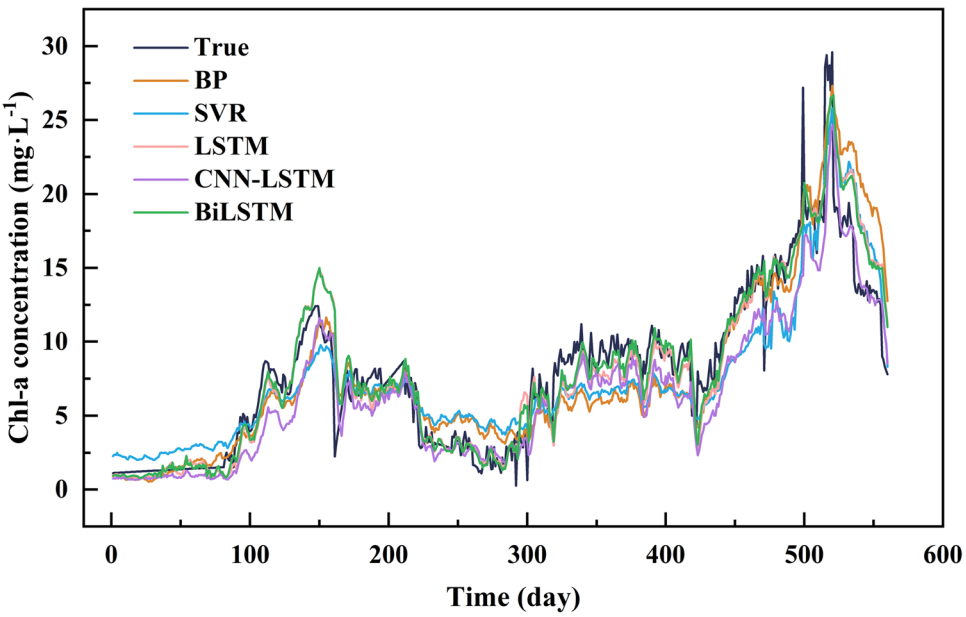

Fig. 8 Comparison of prediction results of single models

performance. These metrics include the root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2).

The comparison results of the prediction effects of the individual models are shown in Fig. 8. Table 4 presents the results of the calculation of four evaluation indicators. The LSTM and BiLSTM models demonstrated superior predictive performance compared to the other individual models. This suggests that the LSTM and BiLSTM models are more effective for predicting time series data. Furthermore, the BiLSTM model has the capability to capture information from both preceding and subsequent contexts, thereby improving its prediction accuracy compared to other models.

Table 4 Error evaluation of single models

Models	Evaluation indicators			
	RMSE	MAE	MAPE (%)	R^2
BP	2.5302	1.8808	20.7661	0.78012
SVR	2.402	2.0284	21.8717	0.80184
LSTM	1.9026	1.288	14.4392	0.87568
CNN-LSTM	2.2844	1.7398	20.1968	0.82076
BiLSTM	1.7104	1.1254	12.7041	0.89953

Next, LSTM and BiLSTM models are selected to be combined with two-stage data processing for model fusion. Four hybrid models, namely SGMD-LSTM, SGMD-KPCA-LSTM, SGMD-BiLSTM, and SKB, were developed. The hyperparameters of each network model were optimized using the SSA algorithm. The relationship between the predicted value and the true value in each model is shown in Fig. 9 below. Table 5 presents the quantitative results for RMSE, MAE, MAPE, and R^2.

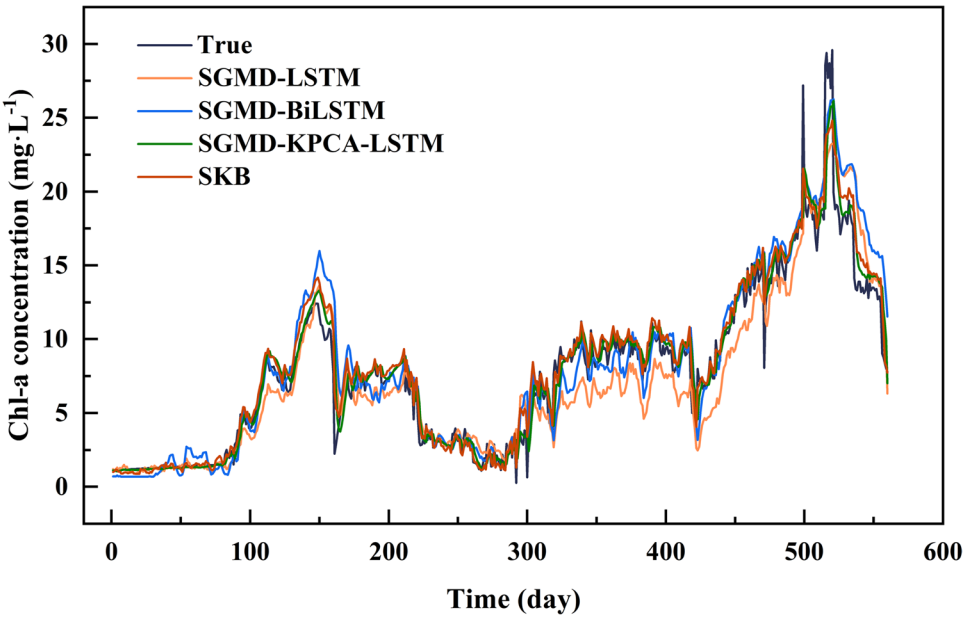

Fig. 9 Comparison of prediction results of hybrid models

Table 5 Error evaluation of hybrid models

Models	Evaluation indicators			
	RMSE	MAE	MAPE (%)	R^2
SGMD-LSTM	2.0045	1.4873	12.7955	0.84491
SGMD-BiLSTM	1.9341	1.3235	12.7617	0.86304
SGMD-KPCA-LSTM	1.3585	0.81151	10.6895	0.93662
SKB	1.0527	0.65194	7.9052	0.96194

The prediction of Chl-a concentration showed significant improvement when a two-stage data processing approach was incorporated into a single network model. From the numerical results presented in Table 4 and Table 5, it is evident that the SGMD-KPCA-LSTM model achieved a reduction in RMSE, MAE, and MAPE of 28.60%, 36.99%, and 25.97%, respectively, compared to the single LSTM model. In comparison to the SGMD-LSTM model, the RMSE, MAE, and MAPE exhibited reductions of 32.23%, 45.37%, and 16.46%, respectively. The R^2 of the SGMD-KPCA-LSTM model showed improvements of 6.96% and 10.85% in the respective cases. The SKB model demonstrated a significant improvement in performance metrics compared to the single BiLSTM model. Specifically, the SKB model achieved a reduction of 38.45% in RMSE, 42.07% in MAPE, and 37.77% in MAE. In comparison to the SGMD-BiLSTM model, the RMSE, MAE, and MAPE exhibited reductions of 45.57%, 50.74%, and 38.06%, respectively. The R^2 for the SKB model increased by 6.94% and 11.46%, respectively.

To further evaluate the predictive accuracy of the Chl-a concentration prediction model, the Taylor diagram is employed to visually summarize the agreement between the predicted and observed values. This graphical representation includes measures of R^2, RMSE, and standard deviation, enabling a comprehensive evaluation of the model's predictive performance, as shown in Fig. 10. Based on the three indicators mentioned above, the reference point is determined, and the position of each model in the figure is obtained. Among these models, the SKB model is the closest to the reference point and performs relatively well. The predictive performance is slightly better than that of the SGMD-KPCA-LSTM model. Notably, the prediction results of the SGMD-LSTM and SGMD-BiLSTM models are unsatisfactory. This suggests that using SGMD alone for data decomposition can effectively extract feature information from the sequence, but it also introduces data redundancy. Therefore, it is essential to employ KPCA to reduce the dimensionality of the data.

The SKB model predicted Chl-a concentration significantly better than the other models being compared. The values of RMSE, MAE, and MAPE were 1.0527, 0.65194, and 0.08, respectively. The evaluation indexes showed that the SKB model outperformed the other eight models. The comparison of the RMSE and MAE values for each model indicates that the SKB model has lower prediction error and higher prediction accuracy. Additionally, the comparison of the MAPE values suggests that the SKB model is more stable. Based on the above analysis, it is evident that the predicted values of the SKB model closely align with the actual values of Chl-a concentration, demonstrating the effectiveness of the model. When compared to directly inputting the data into a single model, the two-stage processing of the data using SGMD and KPCA can

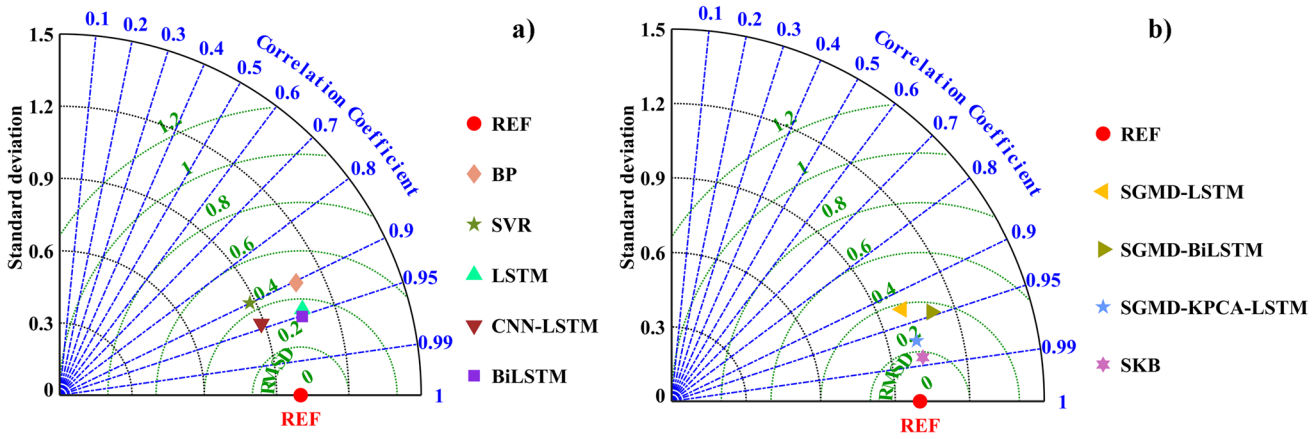

Fig. 10 The Taylor diagram of different models. **a** Single models; **b** hybrid models

enhance the data-driven model's ability to capture changing trends and improve its predictive performance. This finding is consistent with several studies that have used hybrid models. For example, Zamani et al. (2023) demonstrated that a hybrid model incorporating the PNFF prediction model outperformed other single ML algorithms

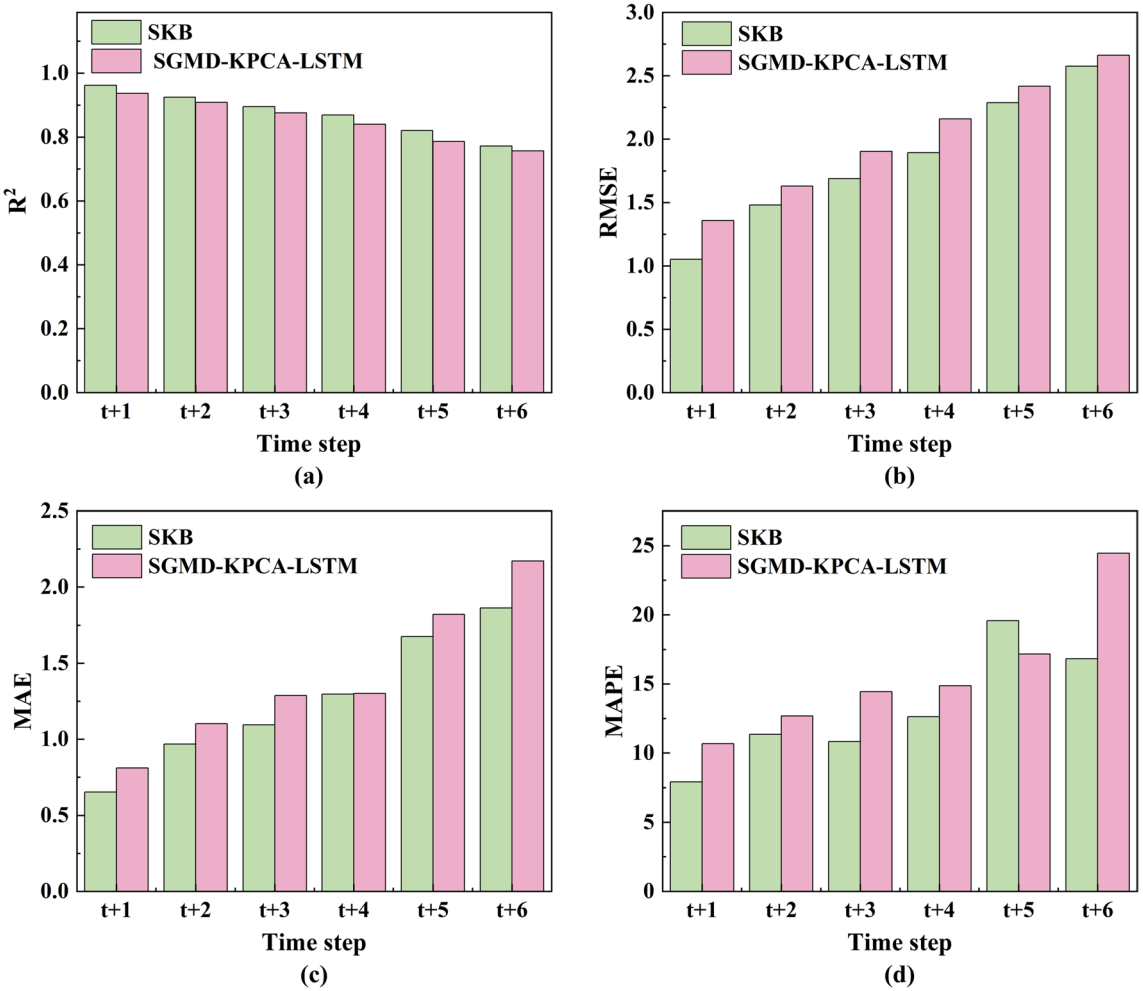

Fig. 11 Performance of the prediction model for Chl-a concentration at different prediction steps

in predicting Chl-a concentration. Zhang et al. (2023b) found that variational mode decomposition (VMD) can effectively reduce the non-smoothness of water quality data. The PV-GRU model proposed in the study significantly improved the accuracy of predicting Chl-a content in reservoirs. Moreover, this study considers the impact of the decomposition algorithm on the redundancy and correlation of sequences. KPCA is used to reduce the dimensionality of the input variables, thereby improving the computational efficiency and accuracy of the model. This is done while ensuring the validity and representativeness of the information. To improve the accuracy of algal bloom predictions, it is essential to utilize combined models, implement multi-level processing of water quality monitoring data, leverage the inherent features of the data, and integrate other intelligent models.

Comparison of different prediction steps

To evaluate the multi-step predictive performance of the SKB model, different prediction steps (1, 2, 3, 4, 5, and 6) were selected for prediction, and the SGMD-KPCA-LSTM model was chosen for comparison. The prediction results of the two models at different prediction time steps are shown in Fig. 11.

From the prediction results at different time steps, it is evident that the values of RMSE, MAE, and MAPE increase for each model as the prediction time step increases. Conversely, the R^2 value gradually decreases, indicating a decline in predictive performance as the number of prediction time steps increases. Additionally, the SKB model demonstrates slightly superior performance compared to the SGMD-KPCA-LSTM model in predicting step sizes. However, both models demonstrate a significant decrease in predictive performance as the prediction step size increases. Therefore, it is essential to investigate methods for maintaining high predictive accuracy when employing a large step size for prediction.

Conclusions

In this study, a combined prediction model named SKB was developed using the SGMD, KPCA, and BiLSTM algorithms. The model was developed based on online monitoring of reservoir water quality data. The SKB model was then used to make short-term predictions of Chl-a concentration. The aforementioned findings can be summarized succinctly as follows:

(1). Considering the inherent characteristics of severe nonlinearity and non-stationarity observed in online monitoring data related to water quality, the utilization of a two-stage data processing approach can effectively overcome the limitations of BiLSTM in handling nonlinear sequences. The utilization of this strategy improves the predictive capabilities of the SKB model.

(2). The historical Chl-a concentration data can be utilized to train the combined prediction model, and the SSA intelligent optimization algorithm is employed to optimize the hyperparameters of the BiLSTM model. The predicted results were significantly better than those of pure data-driven models, such as BP, SVR, LSTM, and CNN-LSTM. The prediction accuracy can reach 96.19%. In conclusion, the SKB model proposed in this paper effectively captures the dynamic change trend of high-frequency algae monitoring data and accurately predicts short-term Chl-a concentration. This provides valuable insights for developing strategies to manage algal blooms.

(3). In the prediction process, both Pearson correlation analysis and gray correlation analysis were employed to identify the main factors influencing the concentration of Chl-a. This laid the foundation for future initiatives aimed at preventing and controlling reservoir eutrophication.

The objective of this study is to develop a prediction model that combines a two-stage data processing approach, ML, and optimization algorithms. This model will be used to predict water quality indicators by considering the interactions among environmental variables. The proposed combined prediction model only utilizes the time series data of monitored water quality. It can efficiently and accurately extract valuable information from the data, making it scalable and applicable to other tasks related to predicting multivariate time series. However, this study still has some limitations that can offer suggestions for future research directions. Chl-a is present in a variety of algae, each exhibiting distinct physiological characteristics. Further studies should aim to improve the applicability of models by effectively managing data and fully extracting characteristics of the water environment from each monitoring station. Furthermore, a limitation of using ML models is their lack of interpretability. Interpretable analyses of the model can be conducted using interpretable ML techniques to improve the model's credibility. In addition, incorporating spatial characteristics into the model to predict the concentration of Chl-a was challenging. By considering the spatial characteristics of reservoirs, such as thermal stratification, it will be possible to make more accurate predictions of Chl-a concentration.

Acknowledgements We thank the anonymous editors and reviewers for their constructive comments and advice. This work was financially supported by the National Natural Science Foundation of China (No. 42377077), the Shandong Province Water Conservancy Research and Technology Promotion Program (No. SDSLKY201902), the Jinan Water Science and Technology Program (No. JNSWKJ202202), and the Research Fund of Water Research Institute of Shandong Province (No. SDSKYZX202102).

Author contribution Wenqing Yu: conceptualization, investigation, methodology, visualization, writing—original draft, writing—review and editing; Xingju Wang: conceptualization, formal analysis, project administration, writing—review and editing; Xin Jiang: formal analysis, resources, project administration, writing—review and editing; Ranhang Zhao: supervision, funding acquisition, writing—review and editing; Shen Zhao: data curation, visualization, software. All authors have read and agreed to the published version of the manuscript.

Funding This work was financially supported by the National Natural Science Foundation of China (No. 42377077), the Shandong Province Water Conservancy Research and Technology Promotion Program (No. SDSLKY201902), the Jinan Water Science and Technology Program (No. JNSWKJ202202), and the Research Fund of Water Research Institute of Shandong Province (No. SDSKYZX202102).

Data availability Not applicable.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent to publish Not applicable.

Competing interests The authors declare no competing interests.

References

Alexakis D, Kagalou I, Tsakiris G (2013) Assessment of pressures and impacts on surface water bodies of the Mediterranean. Case study: Pamvotis Lake, Greece. Environ Earth Sci 70:687–698. https://doi.org/10.1007/s12665-012-2152-7

Antico A, Schlotthauer G, Torres ME (2014) Analysis of hydroclimatic variability and trends using a novel empiricalmode decomposition: application to the Parana River Basin. J Geophys Res-Atmos 119:1218–1233. https://doi.org/10.1002/2013jd020420

Boyer JN, Kelble CR, Ortner PB, Rudnick DT (2009) Phytoplankton bloom status: chlorophyll a biomass as an indicator of water quality condition in the southern estuaries of Florida, USA. Ecol Indic 9:S56–S67. https://doi.org/10.1016/j.ecolind.2008.11.013

Carvalho L, Miller CA, Scott EM, Codd GA, Davies PS, Tyler AN (2011) Cyanobacterial blooms: statistical models describing risk factors for national-scale lake assessment and lake management. Sci Total Environ 409:5353–5358. https://doi.org/10.1016/j.scitotenv.2011.09.030

Cen HB, Jiang JH, Han GQ, Lin XY, Liu Y, Jia XY et al (2022) Applying deep learning in the prediction of chlorophyll-a in the East China Sea. Remote Sens 14:16. https://doi.org/10.3390/rs14215461

Chen MJ, Li J, Dai X, Sun Y, Chen FZ (2011) Effect of phosphorus and temperature on chlorophyll a contents and cell sizes of Scenedesmus obliquus and Microcystis aeruginosa. Limnology 12:187–192. https://doi.org/10.1007/s10201-010-0336-y

Chen JL, Jing HJ, Chang YH, Liu Q (2019) Gated recurrent unit based recurrent neural network for remaining useful life prediction of nonlinear deterioration process. Reliab Eng Syst Safe 185:372–382. https://doi.org/10.1016/j.ress.2019.01.006

Cho S, Lim B, Jung J, Kim S, Chae H, Park J et al (2014) Factors affecting algal blooms in a man-made lake and prediction using an artificial neural network. Measurement 53:224–233. https://doi.org/10.1016/j.measurement.2014.03.044

Dzurume T, Dube T, Shoko C (2022) Remotely sensed data for estimating chlorophyll-a concentration in wetlands located in the Limpopo Transboundary River Basin, South Africa. Phys Chem Earth 127:103193. https://doi.org/10.1016/j.pce.2022.103193

Gao BX, Huang XQ, Shi JS, Tai YH, Zhang J (2020) Hourly forecasting of solar irradiance based on CEEMDAN and multi-strategy CNN-LSTM neural networks. Renew Energ 162:1665–1683. https://doi.org/10.1016/j.renene.2020.09.141

Gentine JA, Conard WM, O'Reilly KE, Cooper MJ, Fiorino GE, Harrison AM et al (2022) Environmental predictors of phytoplankton chlorophyll-a in Great Lakes coastal wetlands. J Great Lakes Res 48:927–934. https://doi.org/10.1016/j.jglr.2022.04.015

Glasgow HB, Burkholder JM, Reed RE, Lewitus AJ, Kleinman JE (2004) Real-time remote monitoring of water quality: a review of current applications, and advancements in sensor, telemetry, and computing technologies. J Exp Mar Biol Ecol 300:409–448. https://doi.org/10.1016/j.jembe.2004.02.022

Hadi SJ, Tombul M (2018) Streamflow forecasting using four wavelet transformation combinations approaches with data-driven models: a comparative study. Water Resour Manag 32:4661–4679. https://doi.org/10.1007/s11269-018-2077-3

He XX, Luo JG, Zuo GG, Xie JC (2019) Daily runoff forecasting using a hybrid model based on variational mode decomposition and deep neural networks. Water Resour Manag 33:1571–1590. https://doi.org/10.1007/s11269-019-2183-x

Hejazi MI, Cai XM (2009) Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm. Adv Water Resour 32:582–593. https://doi.org/10.1016/j.advwatres.2009.01.009

Kerimoglu O, Grosse F, Kreus M, Lenhart HJ, van Beusekom JEE (2018) A model-based projection of historical state of a coastal ecosystem: relevance of phytoplankton stoichiometry. Sci Total Environ 639:1311–1323. https://doi.org/10.1016/j.scitotenv.2018.05.215

Latifoglu L (2022) A novel combined model for prediction of daily precipitation data using instantaneous frequency feature and bidirectional long short time memory networks. Environ Sci Pollut R 29:42899–42912. https://doi.org/10.1007/s11356-022-18874-z

Lee S, Lee D (2018) Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models. Int J Env Res Pub He 15:1322. https://doi.org/10.3390/ijerph15071322

Lee JHW, Huang Y, Dickman M, Jayawardena AW (2003) Neural network modelling of coastal algal blooms. Ecol Model 159:179–201. https://doi.org/10.1016/s0304-3800(02)00281-8

Li Y, Li R (2023) Predicting ammonia nitrogen in surface water by a new attention-based deep learning hybrid model. Environ Res 216:114723. https://doi.org/10.1016/j.envres.2022.114723

Liu W, Liu T, Liu Z, Luo H, Pei H (2023) A novel deep learning ensemble model based on two-stage feature selection and intelligent optimization for water quality prediction. Environ Res 224:115560. https://doi.org/10.1016/j.envres.2023.115560

Lu HF, Ma X (2020) Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere 249:12. https://doi.org/10.1016/j.chemosphere.2020.126169

Mohebzadeh H, Yeom J, Lee T (2020) Spatial downscaling of MODIS chlorophyll-a with genetic programming in South Korea. Remote Sens 12:19. https://doi.org/10.3390/rs12091412

Myronidis D, Ioannou K, Fotakis D, Dörflinger G (2018) Streamflow and hydrological drought trend analysis and forecasting in Cyprus. Water Resour Manag 32:1759–1776. https://doi.org/10.1007/s11269-018-1902-z

Niu Y, Liu CL, Lu XL, Zhu LX, Sun QW, Wang SF (2021) Phytoplankton blooms and its influencing environmental factors in the southern Yellow Sea. Reg Stud Mar Sci 47:8. https://doi.org/10.1016/j.rsma.2021.101916

Ozdogan-Sarikoc G, Sarikoc M, Celik M, Dadaser-Celik F (2023) Reservoir volume forecasting using artificial intelligence-based models: artificial neural networks, support vector regression, and long short-term memory. J Hydrol 616:11. https://doi.org/10.1016/j.jhydrol.2022.128766

Pan HY, Yang Y, Li X, Zheng JD, Cheng JS (2019) Symplectic geometry mode decomposition and its application to rotating machinery compound fault diagnosis. Mech Syst Signal Pr 114:189–211. https://doi.org/10.1016/j.ymssp.2018.05.019

Park Y, Cho KH, Park J, Cha SM, Kim JH (2015) Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci Total Environ 502:31–41. https://doi.org/10.1016/j.scitotenv.2014.09.005

Pepe M, Giardino C, Borsani G, Cardoso AC, Chiaudani G, Premazzi G et al (2001) Relationship between apparent optical properties and photosynthetic pigments in the sub-alpine Lake Iseo. Sci Total Environ 268:31–45. https://doi.org/10.1016/s0048-9697(00)00691-4

Qin Y, Li K, Liang ZH, Lee B, Zhang FY, Gu YC et al (2019) Hybrid forecasting model based on long short term memory network and deep learning neural network for wind signal. Appl Energ 236:262–272. https://doi.org/10.1016/j.apenergy.2018.11.063

Rakocevic-Nedovic J, Hollert H (2005) Phytoplankton community and chlorophyll a as trophic state indices of Lake Skadar (Montenegro, Balkan). Environ Sci Pollut R 12:146–152. https://doi.org/10.1065/espr2005.04.241

Shi XH, Yu HF, Zhao SN, Sun B, Liu Y, Huo JB et al (2023) Impacts of environmental factors on Chlorophyll-a in lakes in cold and arid regions: a 10-year study of Wuliangsuhai Lake, China. Ecol Indic 148:12. https://doi.org/10.1016/j.ecolind.2023.110133

Shin Y, Kim T, Hong S, Lee S, Lee E, Hong S et al (2020) Prediction of chlorophyll-a concentrations in the Nakdong River using machine learning methods. Water 12:18. https://doi.org/10.3390/w12061822

Tong XN, Wang XZ, Li ZK, Yang PP, Zhao M, Xu KQ (2019) Trend analysis and modeling of nutrient concentrations in a preliminary eutrophic lake in China. Environ Monit Assess 191:12. https://doi.org/10.1007/s10661-019-7394-3

Wang XF, Xu LY (2020) Unsteady multi-element time series analysis and prediction based on spatial-temporal attention and error forecast fusion. Future Internet 12:13. https://doi.org/10.3390/fi12020034

Wang X, Zhang YQ, Yu B, Salhi A, Chen RX, Wang L et al (2021) Prediction of protein-protein interaction sites through eXtreme gradient boosting with kernel principal component analysis. Comput Biol Med 134:13. https://doi.org/10.1016/j.compbiomed.2021.104516

Wang K, Fan X, Yang X, Zhou Z (2023) An AQI decomposition ensemble model based on SSA-LSTM using improved AMSSA-VMD decomposition reconstruction technique. Environ Res 232:116365. https://doi.org/10.1016/j.envres.2023.116365

Wu Q, Xia XH, Li XH, Mou XL (2014) Impacts of meteorological variations on urban lake water quality: a sensitivity analysis for 12 urban lakes with different trophic states. Aquat Sci 76:339–351. https://doi.org/10.1007/s00027-014-0339-6

Xie T, Zhang G, Hou JW, Xie JC, Lv M, Liu FC (2019) Hybrid forecasting model for non-stationary daily runoff series: a case study in the Han River Basin, China. J Hydrol 577:15. https://doi.org/10.1016/j.jhydrol.2019.123915

Yu ZY, Yang K, Luo Y, Shang CX (2020) Spatial-temporal process simulation and prediction of chlorophyll-a concentration in Dianchi Lake based on wavelet analysis and long-short term memory network. J Hydrol 582:10. https://doi.org/10.1016/j.jhydrol.2019.124488

Yu DK, Qiao XG, Wang XY (2022) Light intensity optimization of optical fiber stress sensor based on SSA-LSTM model. Front Energy Res 10:9. https://doi.org/10.3389/fenrg.2022.972437

Zamani MG, Nikoo MR, Niknazar F, Al-Rawas G, Al-Wardy M, Gandomi AH (2023) A multi-model data fusion methodology for reservoir water quality based on machine learning algorithms and Bayesian maximum entropy. J Clean Prod 416:18. https://doi.org/10.1016/j.jclepro.2023.137885

Zhang C, Zhang WN, Huang YX, Gao XP (2017) Analysing the correlations of long-term seasonal water quality parameters, suspended solids and total dissolved solids in a shallow reservoir with meteorological factors. Environ Sci Pollut R 24:6746–6756. https://doi.org/10.1007/s11356-017-8402-1

Zhang K, Zhang K, Bao R (2023a) Prediction of gas explosion pressures: a machine learning algorithm based on KPCA and an optimized LSSVM. J Loss Prevent Proc 83:14. https://doi.org/10.1016/j.jlp.2023.105082

Zhang XH, Chen XH, Zheng GC, Cao GL (2023b) Improved prediction of chlorophyll-a concentrations in reservoirs by GRU neural network based on particle swarm algorithm optimized variational modal decomposition. Environ Res 221:9. https://doi.org/10.1016/j.envres.2023.115259

Zhou T, Peng YB (2020) Kernel principal component analysis-based Gaussian process regression modelling for high-dimensional reliability analysis. Comput Struct 241:22. https://doi.org/10.1016/j.compstruc.2020.106358

Zhu XT, Guo HW, Huang JJ, Tian S, Zhang ZJ (2023) A hybrid decomposition and Machine learning model for forecasting Chlorophyll-a and total nitrogen concentration in coastal waters. J Hydrol 619:19. https://doi.org/10.1016/j.jhydrol.2023.129207

Authors and Affiliations

Wenqing Yu[1] · Xingju Wang[1] · Xin Jiang[2] · Ranhang Zhao[1,3] · Shen Zhao[2,4]

✉ Ranhang Zhao
ranhang-z@sdu.edu.cn

[1] Department of Civil Engineering and Water Conservancy, Shandong University, Jinan 250061, China

[2] Water Resources Research Institute of Shandong Province, Jinan 250014, China

[3] Qianfoshan Campus of Shandong University, No. 17923, Jingshi Road, Lixia District, Jinan City 250014, Shandong Province, China

[4] School of Water Conservancy and Environment, University of Jinan, Jinan 250022, China