Permanganate Index Variations and Factors in Hongze Lake from Landsat-8 Images Based on Machine Learning

Yan Lv, Hongwei Guo, Shuanggen Jin, Lu Wang, Haiyi Bian, and Haijian Liu

Abstract

The permanganate index (COD_{Mn}), defined as a comprehensive index to measure the degree of surface water pollution by organic matter and reducing inorganic matter, plays an important role in indicating water pollution and evaluating aquatic ecological health. However, remote sensing monitoring of water quality is presently focused mainly on phytoplankton, suspended particulate matter, and yellow substance, while there is still great uncertainty in the retrieval of COD_{Mn}. In this study, the Landsat-8 surface reflectance data set from Google Earth Engine and in situ COD_{Mn} measurements were matched. The support vector regression (SVR) machine learning model was calibrated using the matchups. With the SVR model, this study estimates the COD_{Mn} in Hongze Lake, presents the historical spatiotemporal COD_{Mn} distributions, and discusses the affecting factors of the change trend of the COD_{Mn} in Hongze Lake. The results showed that the SVR model adequately estimated COD_{Mn}, with a sum squared error of 1.49 mg^2/L^2, a coefficient of determination (R^2) of 0.95, and a root mean square error of 0.15 mg/L. CODMn in Hongze Lake was high in general and showed a decreasing trend in the past decade. Huai River, Xinsu River, and Huaihongxin River were still the main sources of oxygen-consuming pollutants in Hongze Lake. The wetland natural reserve near Yugou had a significant effect on reducing COD_{Mn}. This study provides not only a scientific reference for the management of COD_{Mn} in Hongze Lake, but also a feasible scheme for remote sensing monitoring of COD_{Mn} in inland water.

Introduction

With global warming and the intensification human activities, eutrophication has become a worldwide environmental problem. The deterioration of water quality seriously damages the stability of aquatic ecosystem and threatens the safety of domestic and production water. On the global scale, water quality monitoring of rivers and lakes is usually operated by measuring water temperature, pH, dissolved oxygen, chemical oxygen demand (COD_{Mn}), biochemical oxygen demand, ammonia nitrogen, volatile phenol, cyanide, arsenic, copper, lead, zinc, cadmium, mercury, hexavalent chromium, total nitrogen (TN), total phosphorus (TP), fluoride, transparency, chlorophyll a (Chl *a*), and other indicators on-site for water quality evaluation and management (Liu 1985).

At present, the water quality indicators in China are generally determined by sampling on sites in accordance with the national environmental protection standards (GB 3838-2002; Ministry of Ecology and Environment of the People's Republic of China, 2002). COD_{Mn} effectively indicates the pollution of oxidizable substances in water, and the accurate detection of its spatial distribution is of great significance for aquaculture, aquatic environmental health, and ecological early warning. Therefore, COD_{Mn} is an important indicator for water quality (Li *et al.* 2017). Shang *et al.* (2016) studied the temporal and spatial variation of water environmental factors and found that the main environmental influencing factor of benthic functional feeding groups was total nitrogen in spring and summer, and COD_{Mn} was the main environmental influencing factor in autumn. At the same time, COD_{Mn} was also the main environmental factor of temporal and spatial variation of zooplankton (Shang *et al.* 2016; Shang *et al.* 2021). Rao (2015) measured and obtained high-precision and highly sensitive COD_{Mn} in surface water based on spectrophotometry. A comparative study was conducted on the COD_{Mn} load in the Three Gorges Reservoir area of the Yangtze River between the wet and dry seasons (Huang *et al.* 2021) using in situ measurements. The higher the COD_{Mn} was, the more serious the pollution by organic matter and oxidizable inorganic matter was in water, found by mapping the COD_{Mn} distribution in the Tokyo Bay (Kawabe and Kawabe 1997). Jun *et al.* (2017) found that the main pollution factors in the Hailar River were COD_{Mn} and COD by studying the water quality of the river. All the above studies adopted traditional water quality monitoring methods, namely, field water sample collection and laboratory water quality parameters measurement. Such methods are time consuming and laborious. Meanwhile, these methods are difficult for large-scale water quality monitoring due to the limitation of manpower, material resources, weather, and hydrological conditions.

With the development of remote sensing technology, it is possible to realize water quality monitoring at a large-scale using satellite remote sensing. Compared with the traditional methods, remote sensing provides large-scale, quick-access, and dynamic water quality monitoring. Although the Landsat series of satellites were originally designed for land monitoring, their high spatial resolution (30 m) provides a unique opportunity for monitoring inland water environments with strong spatial heterogeneity and thus have become one of the mostly widely used multi-spectral remote sensing data sources for inland water quality monitoring. Tan *et al.* (2015) constructed empirical models of Chl *a* and

Yan Lv is with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China, and the Faculty of Electronic Information Engineering, Huaiyin Institute of Technology, Huaian, China.

Hongwei Guo is with the School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo, China.

Shuanggen Jin is with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China; the School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo, China; and the Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai, China (sgjin@shao.ac.cn, sg.jin@yahoo.com).

Lu Wang is with the Jiangsu Hongze Lake Management Office, Huaian, China.

Haiyi Bian is with the Faculty of Electronic Information Engineering, Huaiyin Institute of Technology, Huaian, China.

Haijian Liu is with the Faculty of Humanities, Huaiyin Institute of Technology, Huaian, China.

Contributed by Zhenfeng Shao, June 1, 2022 (sent for review July 28, 2022; reviewed by Md. Enamul Huq, Bin Hu).

Photogrammetric Engineering & Remote Sensing
Vol. 88, No. 12, December 2022, pp. 791–802.
0099-1112/22/791–802

doi: 10.14358/PERS.22-00091R2

total suspended solids concentrations using the *in situ* measured water quality data and spectral data synchronously collected by a hand-held spectrometer in the Wabash River and its tributaries in Indiana. Specifically, they estimated Chl *a* concentration using the ratio of the reflection peak in the red-edge band (704 nm) to the absorption valley (677 nm) (coefficient of determination [R^2] = 0.95). The concentration of the total suspended solids was estimated by the logarithm model established using the reflectance of 704 nm and 752 nm (R^2 = 0.83) (Tan *et al.* 2015). Yang (2020) used *Landsat-8* images to invert the Chl *a*, COD, total nitrogen, and total phosphorus of the rivers around Hefei City, Anhui, China, and further evaluated the eutrophication of the rivers. By using 30 Landsat images from 1984 to 2015, Ye *et al.* (2019) mapped the interannual variation of the maximum turbidity zone in Hangzhou Bay, China. Keith *et al.* (2018) studied the algal blooms in the Jordan Lake Reservoir with *Landsat-8* images. Lei *et al.* (2020) used Geostationary Ocean Color Imager images to invert the suspended particulate matter concentration in Hongze Lake. Xiong *et al.* (2019) conducted an inversion study on total phosphorus in Hongze Lake by using the images of a moderate-resolution imaging spectroradiometer. He *et al.* (2021) used *Landsat-8* images to invert the total nitrogen, total phosphorus, and Chl *a* in the main stream of the Yangtze River and analyzed the dynamics of the three water quality parameters from 2014 to 2020. Yang *et al.* (2013) mapped the spatial distribution of concentrations of Chl *a*, total nitrogen, total phosphorus, and other nutrients in the Chaohu Lake, China, and analyzed the seasonal variation characteristics of the water quality parameters, which showed that the biomass of algae in Chaohu Lake was first decreased and then increased. The difference in algal biomass between the east and the west of Chaohu Lake was obvious. The concentration of nutrients and Chl *a* in the west of the lake was higher than that in the east due to the influence of upstream catchment (Jiang *et al.* 2010; Yang *et al.* 2013). However, the existing water quality parameters of remote sensing studies were mainly focused on Chl *a*, suspended matter, and yellow substance. There are only a few studies on TN and TP, while the existing studies have rarely studied COD_{Mn}. Therefore, quantifying the spatial and temporal distribution characteristics of COD_{Mn} in the lake is of great importance for the water management department to timely identify water quality disaster.

The remote sensing inversion of water quality based on different concentrations of optically active constituents (OACs) has different inherent optical properties (IOPs), and IOPs determine the apparent optical properties (AOPs) of a water body. Through simulating the relationship between the OACs and the IOPs and AOPs captured by satellite sensors, the concentration of OACs in water can be inverted by remote sensing images. Meanwhile, indirect remote sensing inversion of non-OACs can be realized by using the correlation between OACs and non-OACs. The common methods for remote sensing water quality inversion include mainly the analytical method, empirical method/semiempirical method, and machine learning (Yang 2020; Guo *et al.* 2021). The analytical method is based on bio-optical models, which have strict physical significance and high inversion accuracy. However, in the analytical method, a large number of model parameters need to be input in the modeling process (Yang 2020). Although the empirical method/semiempirical method has better prediction accuracy and faster modeling speed, the models need to be calibrated by a large amount of measured data and are seriously limited by region and time, so the model generalization is poor (Shen *et al.* 2020; Yang 2020). In recent years, with the development of artificial intelligence, more and more studies have applied machine learning for water quality remote sensing monitoring, achieved adequate model performance, and obtained reliable results (Zhang *et al.* 2009; Mountrakis *et al.* 2011; Guo *et al.* 2011; Guo *et al.* 2011; Xu *et al.* 2013; Jindal *et al.* 2014; Tomar and Agarwal 2015; Jing *et al.* 2015; Xu *et al.* 2019; Chen *et al.* 2020; Chen *et al.* 2021; Mohammad 2021). Based on machine learning, Sagan *et al.* (2020) studied the blue-green algae phycocyanin, Chl *a*, dissolved oxygen, specific conductivity, fluorescent dissolved organic matter, turbidity, and pollution/sediments from 2016 to 2018, utilizing over 200 sets of water quality data in eight lakes and rivers in the midwestern United States. The concentrations of TP and TN were predicted by the artificial neural network (ANN) model and the linear regression model based on the OLI data in the Geshlagh reservoir (Vakili and Amanollahi 2020). The support vector regression (SVR) model outperformed random forest and Cubist for coastal water quality estimation, yielding a calibration R^2 of 0.91 and a coefficient of variation (CV) root mean square error (RMSE) of 1.74 mg/m3 (40.7%) for Chl *a* and a calibration R^2 of 0.98 and CV RMSE of 11.42 g/m^3 (63.1%) (Kim *et al.* 2014). The SVM has a good performance for information prediction, especially for small samples (Duan *et al.* 2015; Liang and Yang 2016; Hou *et al.* 2021; Zhou *et al.* 2022).

In this study, the spatial distribution of COD_{Mn} in Hongze Lake is investigated by field sampling. Then *Landsat-8* multispectral image data are obtained from Google Earth Engine (GEE) and matched up with the in situ measured water quality data. The matchups are used to train, compare, and verify the machine learning model for the remote sensing inversion of COD_{Mn} in Hongze Lake. Subsequently, the monthly spatial distributions of COD_{Mn} in 2015 are mapped for analyzing the intra-annual COD_{Mn} variation and the influencing factors. Finally, the spatial and temporal distributions of COD_{Mn} in Hongze Lake from 2013 to 2021 are estimated and the potential influencing factors of the interannual COD_{Mn} variation are systematically analyzed.

Study Area

Hongze Lake is the fourth-largest freshwater lake in China, located between 118°10′–118°52′E and 33°06′–33°40′N. The lake covers an area of 6,853 km^2 and is shallow (with an average depth of 1.35 m and a maximum depth of 4.75 m) (Li *et al.* 2021) (Figure 1). The lake has a transitional climate between the north subtropical zone and the south temperate zone, with an average annual temperature of 16.3°C and an average annual precipitation of 925.5 mm. The rivers entering the lake include Huai River, Xinbian River, Laobian River, Xinsui River, Laosui River, Xuhong River, Huaihongxin River, and Andong River (Qiao *et al.* 2016), among which the Huai River accounts for more than 70% of the total lake inflow (Xun *et al.* 2003). Hongze Lake is not only an important water source for agricultural and industrial activities in northern Jiangsu Province but also an important water transmission line and regulation and storage lake for the eastern route of the South-to-North Water Diversion Project (Gao and Jiang 2012). The safety of its water quality is crucial for the diversion project and the sustainable economic development along the river, the lake, and even the whole Huai River basin (Guo *et al.* 2021). In recent years, affected by silt and lake reclamation, the area of Hongze Lake has been shrinking, and the water quality problem has been prominent (Xun *et al.* 2003; Wu 2018; Fu and Yue 2019; Cai *et al.* 2020). According to the reply to proposal No. 0435 of the fourth session of the twelfth Jiangsu provincial political consultative congress (Suggestions on adjusting and improving the water environmental quality evaluation system of Hongze Lake and other aquifer shallow lakes), Jiangsu Provincial Department of Ecology and Environment 2021, China (http://www.jiangsu.gov.cn/art/2021/6/15/art_59167_9849985.html), the current COD_{Mn} in Hongze Lake is Class III, with the annual average concentration increase of 9.8% year on year. In December 2021, it was mentioned in the Environmental Monthly Bulletin of the Huaian Ecological Environment Bureau that the COD_{Mn} of Hongze Lake was 3.8 mg/L, up 7.5% year on year and down 15.4% from the previous month (Yang 2021).

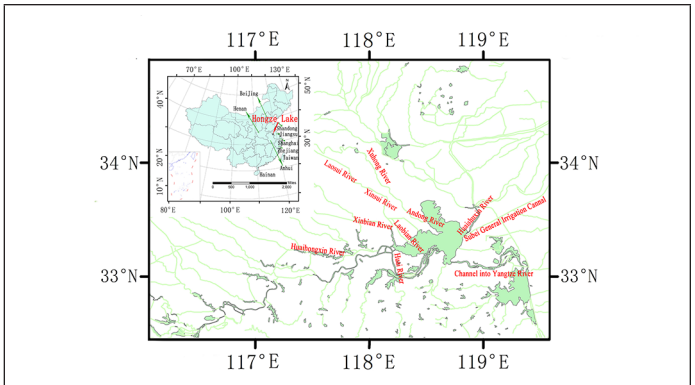

Figure 1. Distribution of the outflow and inflow rivers in Hongze Lake.

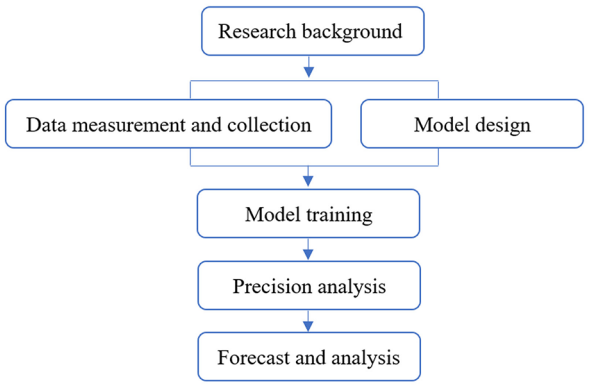

Figure 2. Framework of this study.

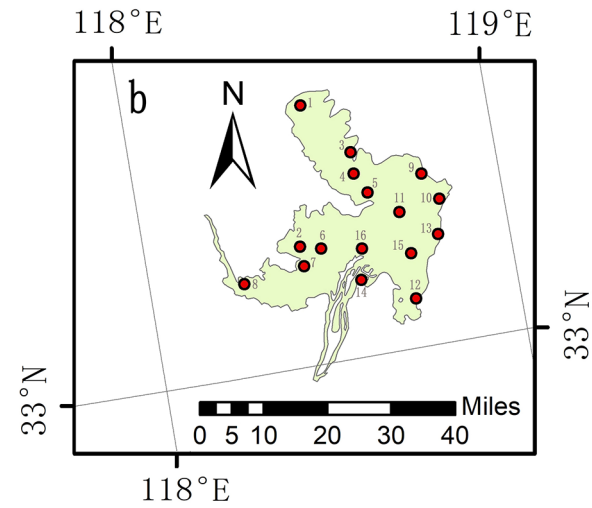

Figure 3. Location of sampling points in Hongze Lake.

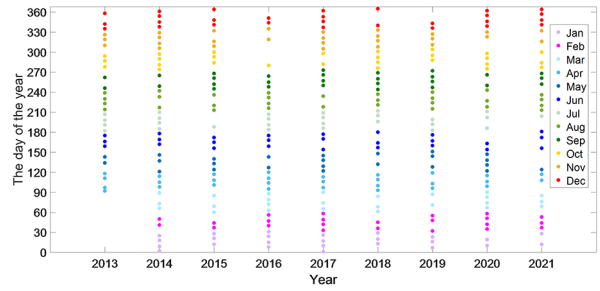

Figure 4. Statistics of satellite transit in the study area from 2013 to 2021.

Table 1. *Landsat-8* image information.

Basic Information		Bands	Bandwidth (μm)	Spatial Resolution (m)
Time horizon	11 April 2013–18 October 2021	Ultrablue	0.435–0.451	30
		Blue	0.452–0.512	30
		Green	0.533–0.590	30
		Red	0.636–0.673	30
Data provider	USGS	Near infrared	0.851–0.879	30
Google Earth Engine ID	LANDSAT/LC08/C02/T1_L2	Shortwave infrared 1	1.566–1.651	30
		Shortwave infrared 2	2.107–2.294	30

Data and Methods

In this study, *Landsat-8* multispectral image data and in situ measured water quality data were collected and then matched up. The matchups were used to build machine learning models and then the permanganate index variations and factors in Hongze Lake were investigated.

In Situ Water Quality Measurements

A total of 16 sampling points were evenly set up over the lake considering lake morphology, hydrodynamics, human activities, and other factors (Figure 3). The water samples were collected in the middle of each month in 2015 with the consistency of the transit time of *Landsat-8*. After being collected, the water samples were quickly stored in amber glass bottles to avoid the sunlight and sent to the laboratory for testing within six hours. The COD_{Mn} testing method followed the standard of the determination of COD_{Mn} of water quality (GB 11892-89, Ministry of Environmental Protection, China, 1990), and the main processing steps were as follows:

1. Sample conservation and delivery: Adding sulfuric acid to the sample to make the pH = 1 to 2 and control the test time within six hours. If the test time exceeds six hours, the test should be kept in a dark place at 0°C to 5°C for no more than two days.
2. Determination of actual samples: Adding a known amount of sulfuric acid (5 ± 0.5 ml) and potassium permanganate solution (10 ml) into the test tube containing the water sample, shake, and place the test tube in a boiling water bath for 30 minutes until the solution is fully reacted. The excess sodium oxalate (10 ml) is added until the solution becomes colorless, and then dropped with the potassium permanganate calibration solution while it is hot until it just appears pink with holding for 30 seconds. The volume of potassium permanganate solution consumed is recorded. Blank test: replacing the sample with 100 ml of water, repeating the steps described above, and recording the volume of potassium permanganate solution consumed.
3. Calculation of concentration: According to Equation 1, the COD_{Mn} concentration of the water sample was finally calculated, represented by mg/L of oxygen, acting as a comprehensive index of the organic pollution degree of the water body:

$$I_{Mn} = \frac{[(10+V_1)\frac{10}{V_2} - 10] * C * 8 * 1000}{100} \quad (1)$$

where V_1 is the volume of potassium permanganate solution consumed during sample titration (ml), V_2 is the volume of potassium permanganate solution consumed during calibration (ml), and C is sodium oxalate standard solution, that is, 0.01 mol/L.

Acquisition and Preprocessing of Remote Sensing Images

Landsat satellite data were used in remote sensing water quality monitoring of inland water (He *et al.* 2021). In this study, the surface reflectance (SR) data set of *Landsat-8* Operational Land Imager (OLI) images from 2013 to 2021 was selected as the data source. *Landsat-8* OLI images consist of five visible and near-infrared bands and two short-wave infrared bands that provide sufficient spectral information to monitor ocean color. The high spatial (30 m) and temporal (16 days) resolutions of the *Landsat-8* OLI satellite support periodic and cost-effective water quality monitoring. The SR data sets of *Landsat-8* OLI were obtained from the GEE platform, which is an immensely powerful and free tool for processing satellite images. The *Landsat-8* OLI SR products in GEE have been atmospherically corrected following the Land Surface Reflectance Code. The satellite transit in the study area is shown in Figure 4. The basis of satellite–earth synchronous screening was that the measured data corresponding to satellite images with cloud coverage of more than 20% or scattered cloud distribution were not used. Through eliminating the outliers of 128 satellite–earth synchronization data, 80 groups of satellite–earth synchronization data were selected for statistical analysis of COD_{Mn} (Figure 5).

Data Analysis

First, by the method of mean value, the overall situation and the monthly variation of the permanganate index in the lake were obtained. According to the match between the *Landsat-8* multispectral image

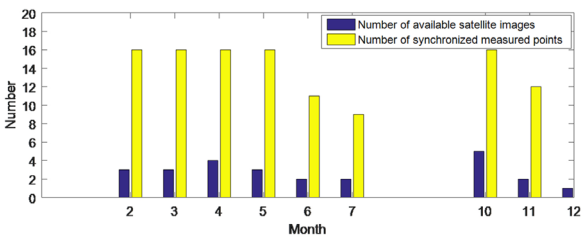

Figure 5. Number of available satellite images and measured data.

data and the in situ measured water quality data, the statistical analysis was carried out month by month and point by point. Finally, data cleaning was performed.

Model Development and Validation

Support Vector Regression

In this study, a machine learning algorithm, namely, the SVR model, was employed to retrieve COD_{Mn} concentrations in Hongze Lake. The SVR algorithm was first proposed by Cortes and Vapnik (1995), mapping x in low-dimensional space to a higher-dimensional characteristic space $\varphi(x)$ to identify a linear regression hyperplane in high-dimensional space that best fits the data (Xu *et al.* 2013; Yu *et al.* 2015). Rooted in statistical learning theory and the structural risk minimization principle, the SVR algorithm was proved to have good performance for handling nonlinear problems (Zhang *et al.* 2009; Andrew 2001). The linear function in the high-dimensional feature space can be expressed as

$$y = w\varphi(x) + b \tag{2}$$

where y is the output, $w\varphi(x)$ is the inner product of the feature space, and $\varphi(x)$ is a nonlinear mapping function. The weight vector w and the bias constant b can be obtained by minimizing the risk function

$$\min\left(\frac{w^2}{2} + C\sum_{i=1}^{N} L_{\mu}\left(x_i, y_i, f\right)\right) \tag{3}$$

where w is the Euclidean norm calculation. $L_{\varepsilon}(xi, yi, f)$ was calculated by the following formula:

$$L_{\mu}(xi, yi, f) = \begin{cases} \left|y_i - f\left(x_i\right)\right| - \mu, \left|y_i - f\left(x_i\right)\right| \geq \mu \\ 0, \text{otherwise} \end{cases} \tag{4}$$

where C is the penalty factor and ε is the deviation between the predicted values and the *in-situ* values. To transform the solution of Equation 3 by introducing the relaxation variable ξ_i and ξ_i^* into

$$\min\left(\frac{w^2}{2} + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)\right) \tag{5}$$

The constraint condition was set as follows:

$$\begin{gathered} y_i - [(w \times x_i) + \mathrm{b}] \leq \varepsilon + \xi_i \\ [(w \times x_i) + \mathrm{b}] - y_i \leq \varepsilon + \xi_i^* \\ \xi_i^*, \xi_i \geq 0 \end{gathered} \tag{6}$$

The Lagrange function was then established to solve the dual problem of the original problem. Based on Equations 5 and 6, the regression function of the optimal hyperplane was finally identified as

$$f(\mathrm{x}) = \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right) K\left(x_i, x\right) + b \tag{7}$$

where $K(x_i, x)$ is the kernel function. The formula can be expressed as

$$K(x_i, x) = \varphi x_i, x = \varphi(x_i)^T \varphi(x) \tag{8}$$

The common kernel functions include the sigmoid kernel, linear kernel, radial basis function kernel, and polynomial kernel. The calculation formula of each kernel function is

$$\begin{gathered} \text{linear: } K(x_i, x_j) = x_i^T, x_j \\ \text{polynomial: } K(x_i, x_j) = (\gamma x_i^T, x_j + r)^d, \gamma \succ 0 \\ \text{radial basis function: } K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2), \gamma \succ 0 \\ \text{sigmoid: } K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \end{gathered} \tag{9}$$

where γ, r, and d represent the kernel function parameters, respectively. The radial basis function was chosen as the kernel function because it can map to infinite dimensions, the decision boundaries are much more diverse, and it has only one parameter. But the linear kernel is only for linearly separable problems, the polynomial kernel refers to the kernel function expressed in polynomial form. It is a nonstandard kernel function suitable for orthogonal normalized data, and the sigmoid kernel is always used for making a multilayer perceptron neural network.

In this study, the SVR algorithm was implemented by the Libsvm program package of Matlab2016. The modeling parameters are shown in Table 2.

Table 2. Modeling parameters of support vector regression.

Modeling Parameters	C	γ
COD_{Mn}	1.4142	5.65690

Grid Search and k-Fold Cross Validation

According to the SVR principle, penalty coefficient C and kernel function parameter γ play an important role in the performance of the model, and adjusting the two parameters separately may not make the model achieve optimal global effect. Therefore, the grid search method is introduced in this study. Its advantage is that only the range of each hyperparameter needs to be set, and the algorithm will automatically find the optimal solution in the designated hyperparameter grid range. Since grid search is a discretized search for hyperparameters, C and γ are searched in the grid with an exponential range of 2 in this article. Meanwhile, as a semiempirical algorithm, the model performance is closely related to the sample numbers of the training model. For the same group of hyperparameter (C, γ) combinations, when the training sample size changes, the fitting performance of the model will also change.

A k-fold cross-validation method was employed to evaluate the model generalization ability under different hyperparameter combinations (Emhamed and Shrivastava 2021). The implementation process of the cross validation is first to randomly divide the original data set into k parts, among which $k - 1$ is used as the training set and the remaining one is used as the test set to verify the model's prediction ability. The final predicted values of the SVR algorithm were obtained by averaging the above k times results (Fu and Yue 2019). A schematic diagram of k-fold cross validation is shown in Figure 6.

Evaluation Indicators of the Model

The results of the SVR model were assessed using the sum of squared error (SSE), the coefficient of determination (R^2), and the RMSE. SSE is the sum of the squares of the errors between the predicted values and the in situ data. An SEE value closer to 0 indicates that model performance is better. R^2 is used to reflect the proportion of a predicted value explained by independent variables through the model. An R^2 closer to 1 indicates that model performance is better. RMSE represents the root mean square of the error between the predicted values and the *in situ* data. An RMSE closer to 0 means that the model has better performance. The calculation formula of each index is as follows:

$$\mathrm{SSE} = \sum_{\mathrm{i}=1}^{n} w_i (y_i - y_i^*)^2 \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i^* - y_i)^2}{\sum_{i=1}^{n} w_i (\overline{y_i} - y_i)^2} \quad (11)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} w_i (y_i - y_i^*)^2}{n}} \quad (12)$$

In the formula, y_i, y_i^*, and $\overline{y_i}$ are the *in situ* values, predicted value, and mean value of COD_{Mn}, respectively, and w_i and n represent the weight and the number of training samples, respectively.

The technical flowchart of COD_{Mn} retrieval in Hongze Lake based on *Landsat-8* images is shown in Figure 7.

Comparison of Different Machine Learning Modelings

In this study, a back-propagation artificial neural network (BPANN) was employed to compare the performance with the SVR algorithm (Xu *et al.* 2013). Structurally, BPANN is a typical multi-layered forward neural network with one input layer, several hidden layers (either one layer or multiple layers), and one output layer. Full connection is adopted between layers, and there is no mutual connection between neurons in the same layer. It has been proved theoretically that a three-layer network with a hidden layer can approximate any nonlinear function (Yu *et al.* 2015). Neurons in the hidden layer adopted a mostly S-type no-linear transfer function; a Relu activation function was added in each hidden layer. The output layer included two parts: the internal structure and the output result. The BPANN model was built in Matlab, and the structure of the model is shown in Figure 8.

BPNN, also known as BP neural network, is an entry classic neural network model, which can be divided into forward and backward propagation (Yu *et al.* 2015). The idea is to learn a certain number of samples (input and expected output); the sample of each input is sent to the network input layer neurons after being calculated by the hidden layer and the output layer, and each neuron outputs the corresponding predicted value by output the layer (Yu *et al.* 2015). In order to further evaluated the fitting performance of the SVR regression model, this study used a feedforward neural network in the BPANN toolbox of Matlab to conduct a modeling comparison between the training set and the test set. The reflectivity of seven bands was taken as the input layer, the measured value of COD_{Mn} was taken as the output layer, and the hidden layer was set to 1. The BPANN model training results were linearly fitted, and the optimal BPANN model was selected according to the comparison of the R^2, RMSE, and SSE of the fitting function. The optimal modeling results were obtained by setting the number of neurons for testing.

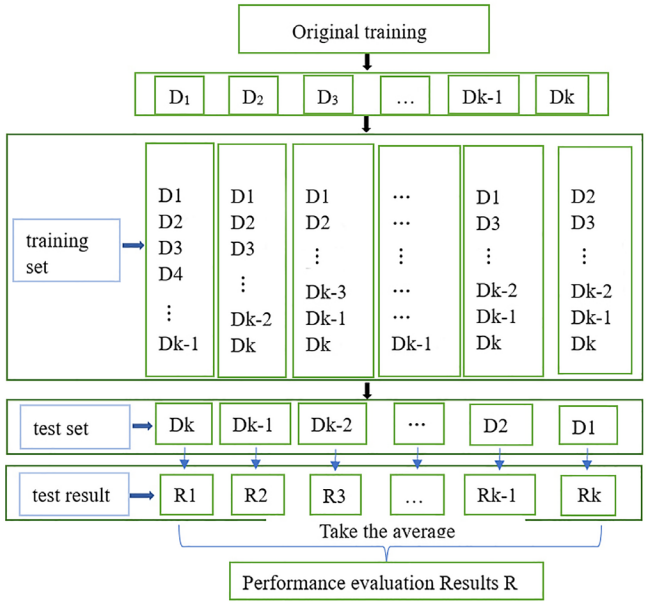

Figure 6. Schematic diagram of *k*-fold cross validation.

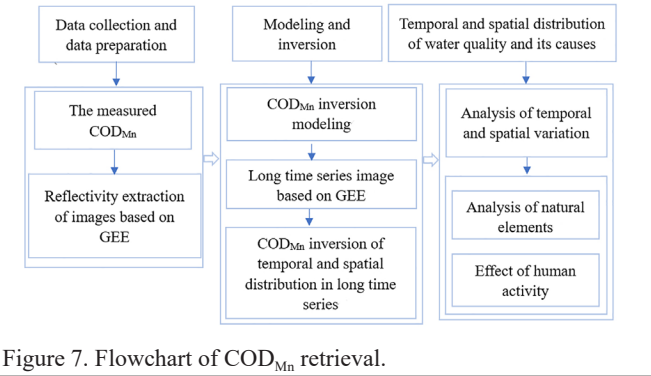

Figure 7. Flowchart of COD_{Mn} retrieval.

Results and Analysis

Analysis of Measured Data

Statistical analysis of COD_{Mn} was conducted at 16 measurement points (Figures 9–11). In general, the distribution of the COD_{Mn} concentrations was uniform. All the sites except Suqian North (the fourth point) have high COD_{Mn} concentrations. COD_{Mn} concentrations were relatively stable in each month of the year. The high values occurred in January and July, and lower values occurred in May and June.

Based on the preliminary data analysis, 128 groups of data were statistically analyzed by SVR, and the data were cleaned according to the deviation between the predicted value and the measured value. Finally, 80 groups of data were selected for SVR statistical regression modeling to improve the modeling accuracy. Analysis of the unselected data shows that the spectral values of all seven bands were on the high side, and the image query found that a small amount of thin cloud existed in the image location corresponding to the measured points. Through the analysis of COD_{Mn} variation with the spectral band, it was found that some unselected points were inconsistent with selected points, which might have been caused by the collection error of measured data. The results show that SVR is suitable for data cleaning.

Optimization of Model Parameters

A grid optimization method was applied to seek the optimal combination of regularization constant and kernel function parameter; the results are shown in Table 3. The model performance was much better in threefold, sixfold, and sevenfold, cross validation than in other folds. Therefore, sixfold cross validation was finally employed for testing the model performance, where the hyperparameters C and γ in the SVR model were obtained as 1.4142 and 5.6569, respectively. The data set was divided

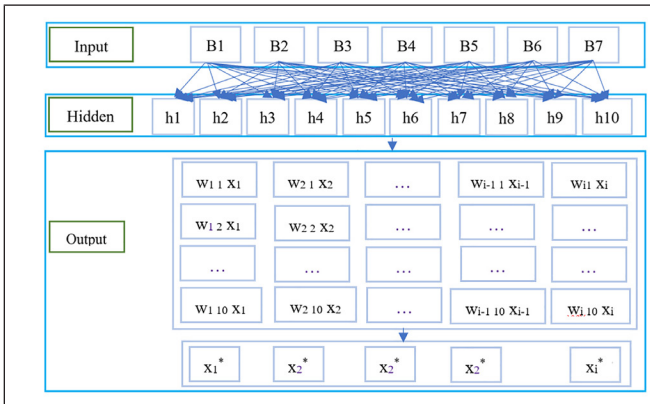

Figure 8. Structure of the BPANN model; the number represents the number of neurons in each layer.

Table 3. Evaluation of support vector regression modeling for different training sets.[a]

k-Folds	SSE	R^2	RMSE	C	γ	*k*-Folds	SSE	R^2	RMSE	C	γ
2	10.49	0.25	0.39	1024	0.002	**7**	**1.10**	**0.97**	**0.13**	**1**	**8**
3	**2.09**	**0.94**	**0.18**	**2**	**4**	8	11.20	0.22	0.41	512	9.7656e-04
4	11	0.23	0.40	724.0773	9.7656e-04	9	5.18	0.84	0.28	2.8284	2
5	8.24	0.71	0.35	1.4142	1.4142	10	5.99	0.81	0.30	2	2
6	**1.49**	**0.95**	**0.15**	**1.4142**	**5.6569**						

SSE = sum of squared error; RMSE = root mean square error.
[a]Bold indicates the setting of the optimal modeling fold for parameters in the test set evaluation.

Table 4. Validation set evaluation of cross-validation attempts to model.[a]

k-Folds	SSE	R^2	RMSE	C	γ
3	0.07	0.88	0.09	2	4
6	**0.06**	**0.88**	**0.09**	**1.4142**	**5.6569**
7	0.13	0.73	0.13	1	8

SSE = sum of squared error; RMSE = root mean square error.
[a]Bold indicates the setting of the optimal modeling fold for parameters in the test set evaluation.

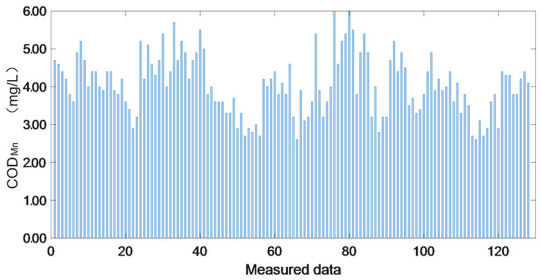

Figure 9. Measured data statistics of COD_{Mn}.

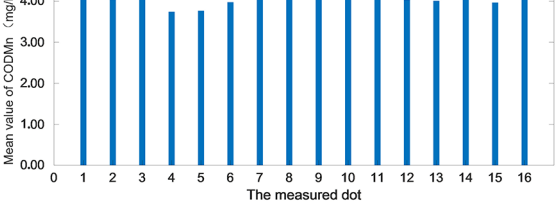

Figure 10. Statistics of COD_{Mn} at each station.

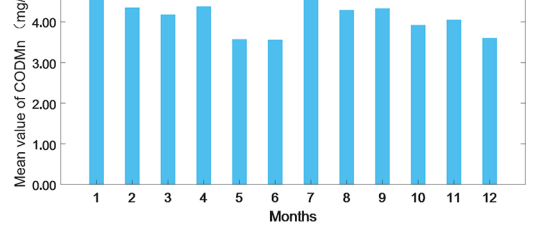

Figure 11. Statistics of COD_{Mn} at each month.

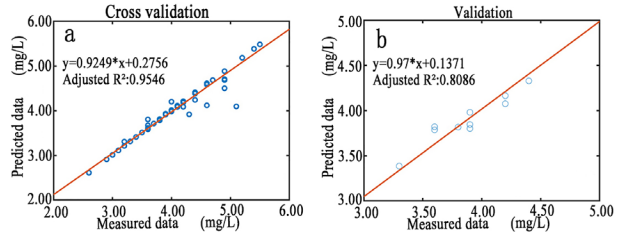

Figure 12. Support vector regression modeling of COD_{Mn} and spectral data on the training set (a) and the test set (b).

into six parts, five of which were trained and one verified in turn; the average value of the results repeated six times was used as the performance indicator of the model under the current sixfold cross validation.

The data set was divided into two parts; 80% were trained, and 20% were verified in turn, and the average of the results of the six runnings was used as the performance indicator of the model under the current sixfold cross validation. Sixfold cross validation was finally employed for testing the model performance, where the hyperparameters C and γ in the SVR model were obtained as 1.4142, and 5.6569, respectively.

Model Performance Evaluation

In the training stage of the SVR model, 70 training data sets were selected. The linear fitting coefficient of measured values and predicted values was $R^2 = 0.95$, the RMSE was 0.15 mg/L, and the SEE was 1.49 mg^2/L^2. Linear regression validation (Equation 12) was used to test the model on the remaining 10 groups of data. The measured value and predicted value had a good linear relationship, with slope a = 0.97, validation determination coefficient $R^2 = 0.88$, RMSE = 0.09 mg/L, and SEE = 0.11mg^2/L^2. The SVR model training and validation results showed that the regression model has good generalization ability, and it was feasible to use the spectral data for COD_{Mn} inversion,

$$y_i^* = ay_i + b \tag{13}$$

where y_i and y_i^* are the measured and predicted values of COD_{Mn}, respectively; a is the slope; and b is the regression constant.

BPANN modeling was also conducted by using satellite–earth synchronization data. After multi-testing, the model was optimized when the number of hidden-layer neurons was 10. The SSE, R^2, and RMSE of the fitting function between the measured and predicted values are shown in Table 5. R^2 performed well, while SSE was generally poor. The linear regression verification analysis of BPANN modeling is shown in Figure 13, in which the slope of linear regression of the training set and the validation set differed greatly. According to Table 5 and Figure 13, the modeling performance of BPANN was poor. Compared with BPANN, the SVR regression model had obvious advantages.

Table 5. Evaluation metrics of the model on the test set.

Test Serial Number	SSE	R^2	RMSE
8	5.41	0.54	0.74
13	5.31	0.53	0.73
21	8.21	0.82	0.91
27	6.35	0.63	0.80
38	5.96	0.60	0.77
43	9.66	0.97	0.98
56	8.93	0.89	0.95

SSE = sum of squared error; RMSE = root mean square error.

COD_{Mn} Estimation from *Landsat-8*

Based on the GEE remote sensing cloud platform, the *Landsat-8* image quality of Hongze Lake was analyzed, and the images of February, March, April, May, August, September, October, and December 2015 were selected to analyze the spatial variation rule of annual COD_{Mn}. By analyzing the image quality of each month in the past ten years, we found that there were more clouds in summer and winter, so year-over-year and month-over-month analysis could not be carried out. The

image quality of spring and autumn (April and October) was better. Therefore, April and October of each year since 2013 were selected as representatives to study the long-term spatial distribution change rule of COD_{Mn}. The established SVR model was used to reconstruct the temporal and spatial distribution of COD_{Mn} in Hongze Lake (Figures 14 and 15). COD_{Mn} in January, June, July, and November 2015, as well as April 2017, April 2019, and April 2021, could not be effectively estimated due to high cloud cover in the images. Since the climatic conditions of

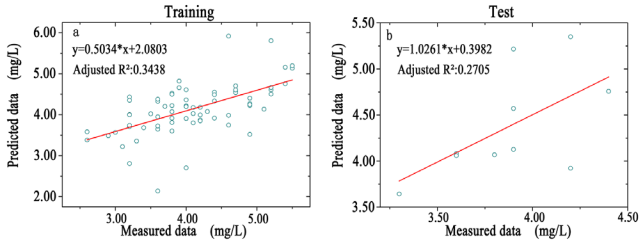

Figure 13. Back-propagation artificial neural network modeling of COD_{Mn} and spectral data on (a) the training set and (b) the test set.

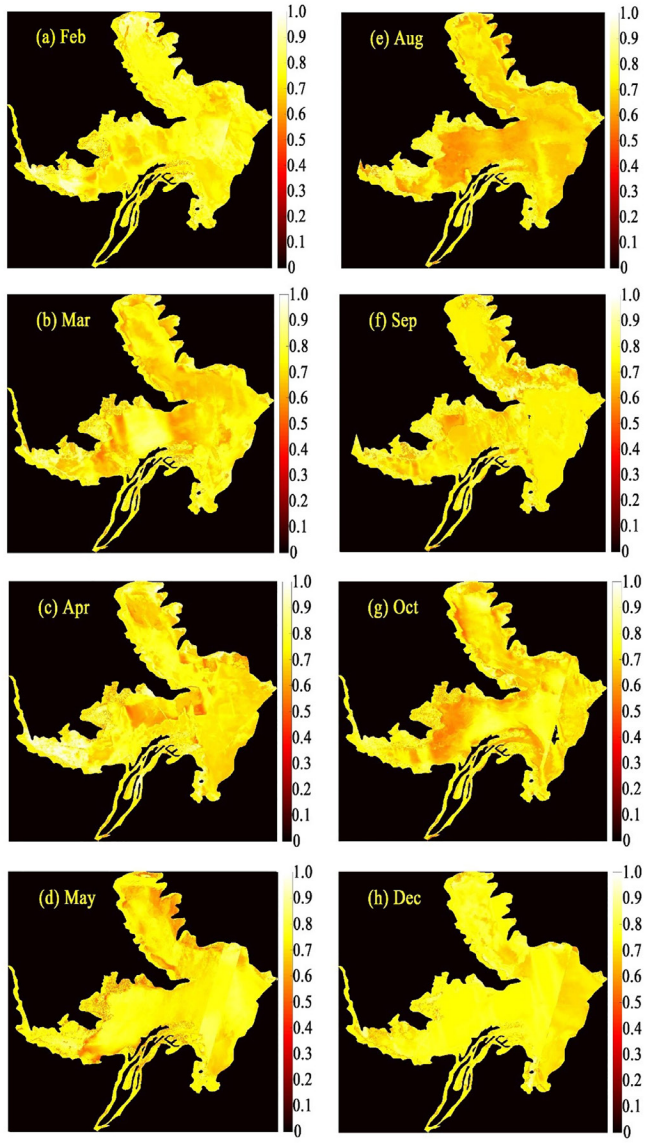

Figure 14. Monthly spatial distribution of COD_{Mn} in Hongze Lake in 2015 (a–h).

Hongze Lake in March and May were similar to those in April in terms of rainfall, temperature, and other conditions, there was little difference in the influence of human activities near the lake from March to May. In order to facilitate the long-term time-series study, on the premise of analyzing the image data quality in March and May, the images in March of the current year were selected as the replacement in 2017 and 2021, and the image in May of the current year was selected for replacement in 2019. In addition, in order to improve the efficiency of model inversion, the prediction initial assignment was set within the range of 2.5 to 5.5 mg/L according to the measured COD_{Mn} data in 2015. The predicted values were normalized to facilitate the comparison and analysis of different concentration distributions and proportions.

According to Figure 14, COD_{Mn} in 2015 was generally high, with the annual COD_{Mn} range of 3 to 6 mg/L in the lake. COD_{Mn} near Yugou (No. 2) in the western part of the lake was generally low throughout the year. Compared with February and December, COD_{Mn} from March to October was especially low. This might be related to the ecological interception of nearby wetlands, dilution of the lake, shallow water depth, and relatively developed aquatic vegetation in coastal waters (Xun *et al.* 2003). Among the inversion results for each month, December and February were relatively high, March and April were slightly higher than September and October, and April and May were obviously higher than other months, which may be related to the low water level of Hongze Lake from April to July (Xun *et al.* 2003). The distribution of COD_{Mn} in February was relatively high, which was consistent with the distribution of the measured data. In December, the COD_{Mn} has generally a certain predicted phenomenon of high relative to the measured data, which might be related to the spectral information of a little thin-ice interference in some regions (Xun *et al.* 2003). COD_{Mn} was relatively low in August, which might be related to more precipitation. The spatial distribution of COD_{Mn} in Chengzi Lake (Nos. 1, 3, and 4) was consistent in each month, which might be related to the relatively closed lake and low flow rate. The COD_{Mn} of Lihewa (No. 8) and Linhuai (No. 7) was higher in every month, which might be related to the organic pollutants discharged into the lake at the inflow of the lake. Suqian North (No. 4), Suqian South (No. 6) and Chenghe (No. 5) were located in the west-central part of the lake, which was consistent with the overall variation law. The COD_{Mn} was relatively low, which might be related to the open water and relatively high lake bottom elevation, significant wind effect, and short water changing cycle caused by wind-generated current (Xun *et al.* 2003). The COD_{Mn} was relatively high in Hanqiao (No. 9) and Xishunhe (No. 10). In Huaian North (No. 11), Huaian South (No. 15), Huaian East (No. 13), Huaian West (No. 16), Jiangba (No. 12), and Xishunhe (No. 10) near the artificial levee of Hongze Lake, the COD_{Mn} was generally lower than that of central and western Hongze Lake and with little change throughout the year. It might be related to the relatively stable annual discharge of domestic sewage from the nearby towns. According to the measured data, COD_{Mn} in Laozi Mountain (No. 14) was low in February, April, and October, but the predicted distribution was high throughout the year, which might be related to the acquisition of the surrounding land disturbance spectral information.

According to Figure 15, COD_{Mn} in Hongze Lake was high in general and has shown a decreasing trend in the past decade. This might be related to the management and control of catering vessels in the lake, wetland restoration, special regulation of river channels into the lake, sewage outlets into the lake, and agricultural non–point source pollution since 2013. Near the artificial levee of Hongze Lake, the decrease trend of COD_{Mn} was more obvious, which might be related to the effective promotion of relocation and reinforcement along the levee. In terms of time, COD_{Mn} was relatively high in spring. Comparative analysis of COD_{Mn} spatial distribution in the whole lake in two months showed that, except for 2017 and 2021, COD_{Mn} in April was generally higher than October, which was consistent with the measured results and might be related to the low water level from April to July (Xun *et al.* 2003). Since 2019, COD_{Mn} decreased significantly, especially in Chengzi Lake (Nos. 1 to 3), Yugou (No. 2), and Suqian North (No. 6). Among these, the cycle of Chengzi Lake was relatively slow, and

this area was surrounded by the development zone of Suqian City. The decreasing of COD_{Mn} fully demonstrated that the control of industrial wastewater pollution had achieved remarkable results (Li *et al.* 2021). The spatial heterogeneity of organic pollution content in Hongze Lake was small and at the inflow of the lake was obviously higher, which might be related to the injection of organic pollutants and other factors. As the main inlet of the lake, the COD_{Mn} of Huai River continued to be high, which was consistent with the research results that Huai River, as the main inflow source (Wang *et al.* 2013), was the main contributor to the pollution load of Hongze Lake, and its pollution into the lake accounted for more than 51% of the total pollution (Vapnik 1995; Ge and Wang 2008).

The COD_{Mn} in Yanwei (No. 1) and Gaohu (No. 3) was high in April but low in October 2016 and 2019. This water was relatively closed, with a slow flow rate and long water exchange cycle, which might be related to the discharge of pollution from the surrounding aquaculture. COD_{Mn} was relatively high in Suqian North (No. 4), Suqian South (No. 6), Chenghe (No. 5), and Hanqiao (No. 9) and was significantly higher in April than October, which might be related to the inflow of rivers into the lake, agricultural production and aquaculture waste discharge, and urban domestic sewage discharge. The COD_{Mn} of Xishunhe (No. 10) and Huaian North (No. 11) was relatively high, but it was lower in April than October because the area was close to Zhouyu Village and was affected by rural sewage and living habits (Ye *et al.* 2011).

The COD_{Mn} of Huaian East (No. 13), Huaian South (No. 15), Huaian West (No. 16), and Jiangba (No. 12) was relatively low, which might be due to its being located in the main water crossing area of Hongze Lake and water exchange being faster. In addition, the pollution discharge adjacent to the artificial levee of Hongze Lake was relatively well controlled. The COD_{Mn} of Yugou (No. 2) and Linhuai

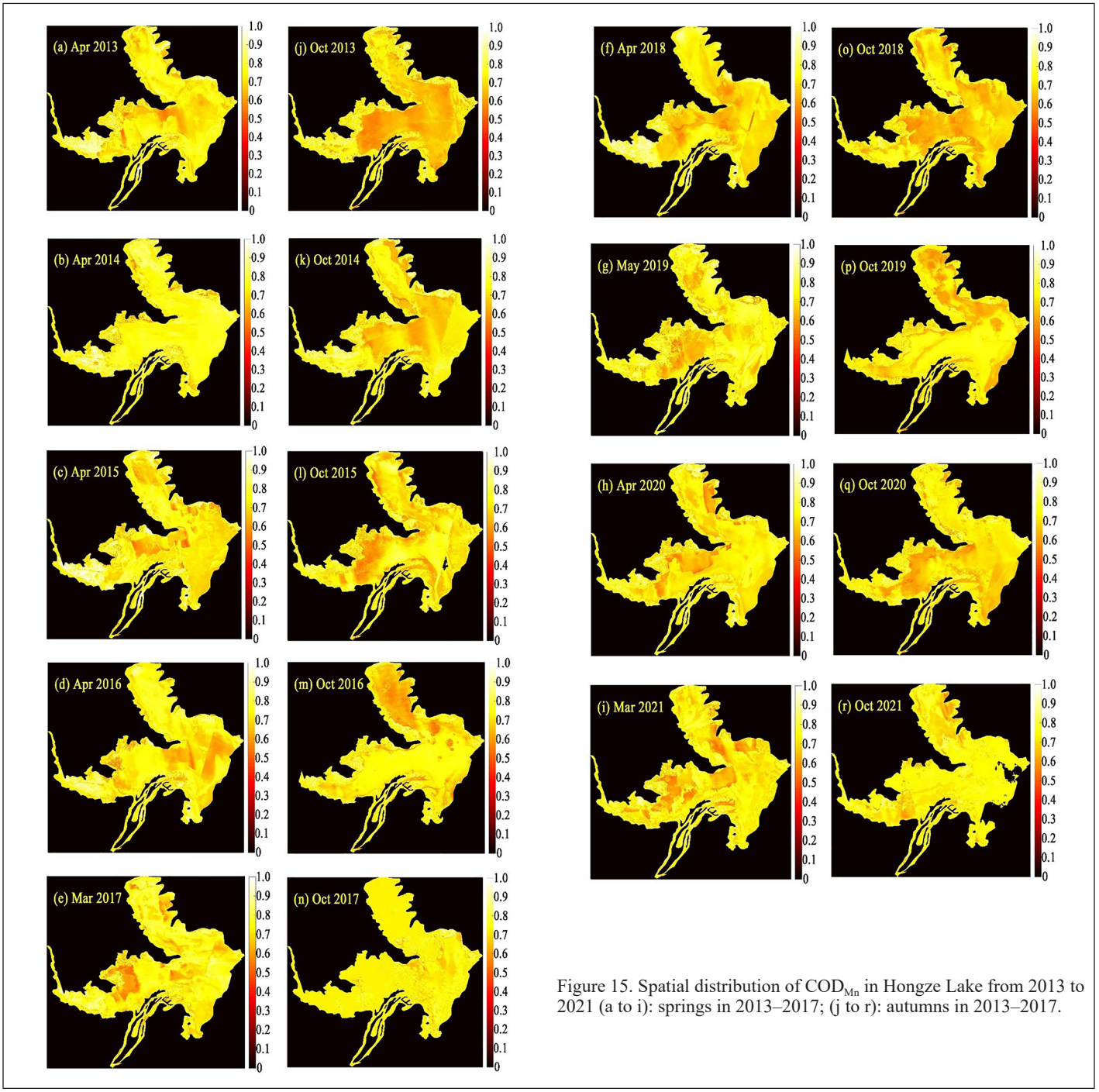

Figure 15. Spatial distribution of COD_{Mn} in Hongze Lake from 2013 to 2021 (a to i): springs in 2013–2017; (j to r): autumns in 2013–2017.

(No. 7) was located near the Hongze Lake Wetland National Nature Reserve. Compared with the adjacent Lihewa (No. 8), the COD_{Mn} of Yugou (No. 2) and Linhuai (No. 7) was significantly lower, which might be related to the comprehensive improvement of the rural environment. According to the measured values, the COD_{Mn} of Laozi Mountain (No. 14) was relatively low, which might be related to the good local natural environment and high green space coverage. However, the comparison between the predicted and measured values in this region showed that the predicted values were overestimated, which might be related to the influence of adjacent land pixels.

Spatial Distribution and Influencing Factors of COD_{Mn} in Hongze Lake

According to the surface water environmental quality standard of China (GB 3838-2002; Cui *et al.* 2021; Li *et al.* 2021), the surface water can be divided into five water quality classes according to their functions: Class I applies mainly to source water and national nature reserves. Class II applies mainly to the centralized drinking water surface water source level 1 protection zone, the habitat of rare aquatic organisms, the spawning grounds of fish and shrimp, the feeding grounds of young fish, and so on. Class III applies mainly to the centralized drinking water surface water source secondary protection area, fish and shrimp wintering grounds, migration channels, aquaculture areas, and other fishery waters and swimming areas. Class IV applies mainly to general industrial water areas and recreational water areas with no direct contact with the human body. Class V applies mainly to agricultural water use areas and general landscape water areas. Among these, the upper limit of COD_{Mn} in a Class II water body is 4 mg/L, and that in a Class III water body is 6 mg/L. The COD_{Mn} in the whole of Hongze Lake was generally between 3.0 and 6.0 mg/L, and the high value of COD_{Mn} had a wide distribution, showing non–point source characteristics. From the comparison of different months and years, COD_{Mn} in the lake was generally high, and the area of Class III water in each month was large, while COD_{Mn} in August was relatively low. From 2013 to 2021, the water quality was classified as Classes II to III, especially in winter. In terms of spatial distribution, the inlet area was significantly higher than the center area. COD_{Mn} is an index to characterize the degree of organic pollution and reductive inorganic substances in water and can be used to measure the content of oxygen-consuming substances in water. The higher COD_{Mn} is, the less dissolved oxygen in water is. Hypoxia seriously affects the growth of fish and crabs in the water and even leads to the death of a large number of fish and crabs. For example, from 25 to 26 August 2018, the water pollution incident of Hongze Lake caused by sewage being carried by upstream flood discharge resulted in the whole river turning black and a large number of fish and crabs dying. The overall distribution of COD_{Mn} along the lakeshore was relatively uniform, showing a slightly higher distribution of bay than non-bay areas and no obvious over-distribution of point sources. According to the high value present the characteristics of the non–point source distribution and the influencing factors for the spatial distribution of COD_{Mn} in Hongze Lake, it can be concluded that in recent years, Hongze Lake's surrounding industrial wastewater emissions regulation was productive. Due to that "water pollution prevention and control law," "a three-year work plan for Hongze Lake governance and protection" and other policies were promoted, and since 2019, COD_{Mn} in Chengzi Lake has been significantly reduced (Nos. 1 to 3), which showed that industrial wastewater pollution has been effectively controlled in recent years (Cui *et al.* 2021).

Exogenous pollution in the inflow of Hongze Lake was still the main reason for the relatively high COD_{Mn}. The COD_{Mn} values of Huai River, Xinsu River, and Huaihongxin River were above 4.4 mg/L. The mean value of COD_{Mn} measured near the Linhuai station was 4.5 mg/L. Hongze Lake was a waterborne lake, and the inflow of Huai River took up 70% of the total inflow of the lake (Xun *et al.* 2003). The Huai River pass through Huainan and Bengbu, Xinsu River pass through Xuzhou, Huaibei and Suzhou, and Huaihongxin River pass through Fuyang, respectively. The discharge of industrial sewage and domestic water from the cities causes serious pollution in the rivers. Strengthening the management of wastewater discharge in the upper reaches of rivers into the lake still played an important role in reducing the exogenous input of oxygen-consuming substances in Hongze Lake. In addition, compared with the adjacent Linhuai (No. 7) and Lihewa (No. 8), the COD_{Mn} of Yugou was obviously low throughout the year, with an average of 4.0 mg/L, indicating that the orderly construction of wetland nature reserve was conducive to the retention of organic pollutants and the improvement of water quality.

The main sources of pollution in the lake were agricultural production and domestic sewage. According to the survey, there were relatively few large enterprises around Hongze Lake, but there were phenomena such as land reclamation, illegal construction of docks, illegal cultivation, illegal construction of factories, fishermen's villages, catering boats, and so on. Due to the long shoreline and the large arbitrariness in the sale, transfer, and alteration of ships, it was difficult to make clear the facts and the situation, which also increased the difficulty of control. In addition, a considerable amount of garbage and pollutants was produced by the scattered domestic boats and their fishermen, transport ships, and the people in production and living nearby water bodies, and they were not collected and treated in a centralized manner. The biological and environmental pollution caused to the Hongze Lake system cannot be underestimated.

Some fishermen illegally catch aquatic products at night during the closed fishing season or in closed fishing areas or illegally hunt in remote waters, bringing uncontrollable human factors to the environmental protection of the river system. At present, the area of Seine farming in Hongze Lake is more than 200 km^2, and the output is as high as two times the fishing amount, of which more than 90% is from crab farming. Excessive feeding and discharge sedimentation produced a lot of endogenous pollution. In addition, the phenomenon of fencing and land reclamation is common in the south, west, and north of the lake, especially in the south of Sihong County and the north of Xuyi County. As a result, the free water surface reduced, which seriously affected the growth of vegetation along the lake and also reduced the interception and degradation of organic pollutants. The circle fencing phenomenon was particularly prominent near Linhuai (No. 7) and Laozi Mountain (in No. 14). In the eastern part of Hongze Lake, there was less circle fencing, but the population density was high. Under the strong interference of human activities, COD_{Mn} was also high. The discharge of pesticide and fertilizer, fishery tail water, and domestic sewage were the main causes of the endogenous pollution of Hongze Lake. Therefore, orderly promotion of wetland protection, returning farmland to the lake, ecological agriculture, illegal supervision, fishermen ashore, and other measures could help to reduce the pollution of Hongze Lake.

Discussion

In this study, the *Landsat-8* images were used to investigate the spatial and temporal distributions of COD_{Mn} in Hongze Lake, and the influencing factors were analyzed. With the help of the established inversion SVR model, the distributions of the monthly COD_{Mn} in 2015 and the interannual COD_{Mn} during the past decade in the whole lake were investigated. The approach proposed in this study provides reference for remote sensing monitoring of inland water quality but also has some limitations.

Approach Effectiveness

The evaluating metrics of the SVR model established in this study indicated that the modeling effect was good and outperformed the BPANN, with $R^2 = 0.95$, RMSE = 0.15 mg/L, and SSE = 1.49 mg^2/L^2. Xu *et al.* (2013) analyzed from the perspective of principle and believed that SVR conducts model training by seeking a way of minimizing structural risk, aiming at controlling the overall error, while the BPANN aims at continuously fitting the local truth value, without controlling the overall error, leading to poor model generalization in model prediction. The conclusion in this study further explained the predictive advantage of the SVR model in Hongze Lake.

The distribution of COD_{Mn} obtained from the *Landsat-8* images was consistent with the results obtained from traditional monitoring, and only a few points showed relatively large deviation, such as Laozi Mountain, which suffered from the influence of adjacent land pixels.

The distribution of COD_{Mn} and the spatiotemporal variation at nonsampling sites were in sync with the characteristics of local natural factors and human activities. The interannual variation of COD_{Mn} was consistent with the spatiotemporal variation of results during 2012–2018 (Li *et al.* 2021). The seasonal variation of COD_{Mn} was consistent with the variation in the inflow rivers of Hongze Lake proposed by (Cui *et al.* 2021). These results indicated that the proposed approach was effective in the investigation of COD_{Mn} distribution in Hongze Lake. Table 6 shows the comparison with other techniques.

Table 6. Final comparison of the model on test set.

Methods	R^2	RMSE (mg/L)	SSE (mg^2/L^2)	Index
Proposed (SVR)	0.95	0.15	1.49	COD_{Mn}
Kim et al. (2014) (SVR)	0.91	1.74		Chl *a*
	0.98	11.42		SPM
Vakili and Amanollahi (2020) (ANN)	0.64	0.04	0.03	TP
	0.86	0.06	0.05	TN
Xu et al. (2013) (RF)	0.95	2.784	—	Chl *a*

RMSE = root mean square error; SSE = sum of squared error; SVR = support vector regression; ANN = artificial neural network; RF = random forest.

Approach Limitations

In data preparation, after the screening of data, such as image cloud coverage without exceeding 20%, no scattered cloud blocking of the sampling points, and abnormal point elimination, only 80 groups of data met the requirements, and the sample size was relatively insufficient. The in situ measured data were obtained from 16 sampling points in the whole lake in monthly batches. The amount of modeling data in a single batch was insufficient, while the measured data in different batches had certain differences in imaging conditions, affecting the accuracy of the modeling. In the future research, more sample data in the same batch should be collected to study the applicable boundary of the model. In the process of the SVR modeling, the grid search was used to optimize the hyperparameters of the model, which lacks of physical basis, so it is impossible to carry out specific physical explanation. But from the evaluations of SVR modeling for the training set and the test set (R^2 = 0.95, RMSE = 0.15, SSE = 1.49; R^2 = 0.88, RMSE = 0.09, SSE = 0.06, respectively), the result is very beneficial for the management of COD_{Mn} in Hongze Lake and provides a feasible scheme for remote sensing monitoring of COD_{Mn} in inland water.

Due to the low temporal resolution of *Landsat-8* images and the influence of clouds, the monthly dynamic monitoring could not be satisfied. In this study, when studying the monthly distribution of COD_{Mn} in 2015, the images of some months were missing, and the detailed spatial distribution of COD_{Mn} in each month of the year could not be mapped. Similarly, due to the influence of clouds and other factors, although April and October were selected as representatives for the distribution of COD_{Mn} in the past decade, there were still missing data in some months, which would inevitably lead to certain errors if the images of the adjacent months were selected. Sentinel-2 images have higher temporal resolution and more bands than those of the *Landsat-8*, so it is suggested that the Sentinel-2 images could be used to investigate the distribution of COD_{Mn} in the future.

During the analysis of the influencing factors related to COD_{Mn} distribution, a comprehensive survey was carried out mainly in the lakefront and the lake area. But the surveys along the river were not included, especially the characteristics and laws of non–point source pollution. In the future research, land use in the western watersheds of Hongze Lake, such as the watersheds of Huai River, Xinsu River, Xinbian River, and others, should be investigated to find out the correlation between land use and the dynamics of COD_{Mn} distribution.

Conclusion

With the help of the GEE remote sensing cloud computing platform, the spatial and temporal distributions of COD_{Mn} in Hongze Lake were modeled by the SVR model, established on the basis of *Landsat-8* images and the synchronal in situ water quality measurements. Through analysis of the monthly spatiotemporal variation of COD_{Mn} in 2015 and the interannual spatiotemporal variation in the past decade, the potential influencing factors of COD_{Mn} distribution were explored. The main conclusions are as follows:

1. The SVR model accurately fits the spectral data of *Landsat-8* images and the measured COD_{Mn}, with R^2 = 0.95, RMSE = 0.15 mg/L, and SSE = 1.49 mg^2/L^2. The slope associated with the linear regression between the validating set and the measured data was 0.97, R^2 reached 0.88, and RMSE and SSE were 0.09 mg/L and 0.11 mg^2/L^2, respectively. Meanwhile, the model comparison showed that the SVR model performed better than the BPANN model for the COD_{Mn} remote sensing inversion in Hongze Lake.
2. Huai River, Xinsu River, and Huaihongxin River were still the main sources of oxygen-consuming pollutants in Hongze Lake. The wetland natural reserve near Yugou had a significant effect on reducing COD_{Mn}. According to the annual and interannual variations of COD_{Mn}, the COD_{Mn} in the whole lake was generally high and thus long-term control of endogenous pollution and fishermen's villages and wetland ecological restoration should be carried out from the perspective of overall ecological factor recovery.
3. The COD_{Mn} in the lake was high, and the COD_{Mn} at the eastern inlet of the lake was higher, indicating that the rivers into the lake were the main source of organic pollutants. The agricultural non–point source pollution in the basin and the overall management of river channels were necessary measures for the exogenous management of Hongze Lake. However, COD_{Mn} in August was significantly lower, which might be due to the large precipitation with a significant dilution effect on pollutants. In the coastal urban agglomeration areas within the jurisdiction of Siyang County, Hongze District, Xuyi County, and Sihong County, the COD_{Mn} had little change from year to year and did not change significantly throughout the year. It indicated that the main sources of organic pollutants were urban industrial wastewater and domestic water. It is necessary to further improve the treatment and management system of industrial and domestic wastewater, strengthen the promotion of rainwater and sewage diversion projects, and improve the efficiency of sewage treatment.
4. According to the change rule of COD_{Mn} in the past decade, the measures related to Hongze Lake governance in recent years have achieved certain results, but the results are not significant. It is necessary to further refine and promote the measures based on local conditions and ensure long-term effect. The water level of Hongze Lake is controlled by humans. The water level is higher in spring and lower in summer, which is contrary to the seasonal water level variation in dry, abundant, and normal periods in the Huai River basin. In the comprehensive treatment, it is necessary to combine the characteristics of water level control management, water resources development, utilization scenarios, spatial distribution of water quality pollution, and other factors to formulate a governance mode of integrated "one place, one policy, overall planning, and coordination" management, control, and governance.

Due to the limitation of field sampling and the low temporal resolution of the *Landsat-8* images, this study still has the problems of insufficient COD_{Mn} model training samples and a single sampling area. In the future research, higher temporal-resolution multispectral images and more in situ COD_{Mn} measurements are expected to further improve the model performance. In addition, multi-region sampling can also validate the model and improve the generalization performance of the SVR model.

Acknowledgments

This work was supported by the Strategic Priority Research Program Project of the Chinese Academy of Sciences (grant no. XDA23040100) and the Jiangsu Marine Science and Technology Innovation Project (Grant No.: JSZRHYKJ202202, and JSZRHYKJ202002). We thank the USGS for providing *Landsat-8* data and the Jiangsu Hongze Lake Management Office for providing the in situ measurements.

References

Andrew, A. M. 2001. An introduction to support vector machines and other kernel-based learning methods by Nello Christianini and John Shawe-Taylor, Cambridge University Press, Cambridge, 2000, 189pp, Robotica, 18(06): 687-689.

Cai, Y., Z. Zhang, R. Tang, Y. Chen, W. Huang, Z Gong. 2020. Health evaluation and protection of Hongze Lake ecosystem. *Jiangsu Water Conservancy* 7:2–7.

Chen, S., M. Ren, W. Sun. 2021. Combining two-stage decomposition based on machine learning methods for annual runoff forecasting. *Journal of Hydrology* 603:126945.

Chen, Y., S. Liu, K. Wang, S. Song, P. Liang, F. Chen. 2020. Remote sensing inversion of water quality and nutrient status evaluation of Caohai Sea based on Landsat satellite images. *Journal of Hydroecology* 5(41):25–27.

Cortes, C. and V, Vapnik. 1995. Support-vector networks. *Machine Learning* 20(3):273–297.

Cui, C., Z. Dong, J. Tong, Q. Shi, T. Zhang, X. Chen. 2021. Water quality evaluation and variation law of main inflow channels of Hongze Lake. *Jiangsu Water Conservancy* 9:27–33.

Duan, Q., L. Meng, Z. Fan, W. Hu, W. Xie. 2015. Research on the applicability of water information extraction method from GF-1 satellite images. *Remote Sensing for Natural Resources* 27(4):79–84.

Emhamed, A. and J. Shrivastava. 2021. Electrical load distribution forecasting utilizing support vector model (SVM). *Materials Today: Proceedings* 47(1):41–46.

Fu, D. and J. Yue. 2019. Analysis of Hongze Lake surface area and its change based on Landsat image. *Journal of Gansu Sciences* 31(2):36–38.

Gao, J. and Z. Jiang. 2012. *Conservation and Development of Five Major Freshwater Lakes in China*. Beijing: Science Press.

Ge, X. and G. Wang. 2008. Ecological and environmental problems faced by Hongze Lake and their causes. *The People of the Yangtze River* 39(1):27–29.

Guo, H., J. Huang, B. Chen, X. Guo, V. Singh. 2021. A machine learning-based strategy for estimating non-optically active water quality parameters using Sentinel-2 imagery. *International Journal of Remote Sensing* 42(5):1841–1862.

Guo, J., Y. Chen, Z. Li, F. Fang, Z. Sun, Y. Chen. 2011. Seasonal variation of chlorophyll A and its relationship with major algae in Xiaojiang Backwater of the Three Gorges. *Environmental Science* 32(4):976–981.

He, Y., S. Jin and W. Shang. 2021. Water quality variability and related factors along the Yangtze River using Landsat-8. *Remote Sensing* 13(2241):2–17.

Hou, W. J. Wang and X. He. 2021. Heavy mineral oil spectral pattern recognition based on support vector machine modeling. *Laser and Optoelectronics Progress* 58(6):630001.

Huang, Y., Z. Huang, W. Xiao, L. Zeng, H. Li. 2021. Study on ammonia nitrogen and COD_{Mn} load in inlet and outlet sections of the Yangtze River in the Three Gorges Reservoir Area. *Journal of Zhejiang A&F University* 6:1221–1230.

Jiang, X., S. Wang, L. Zhong, X. Jin, S. Sun. 2010. Seasonal variation of algae biomass in Chaohu Lake. *Environmental Science* 31(9):2056–2062.

Jindal, R., R. Thakur, U. Singh, A. Ahluwalia. 2014. Ahluwalia phytoplankton dynamics and water quality of Prashar Lake, Himachal Pradesh, India. *Sustainability of Water Quality and Ecology* 3(4):101–113.

Jun, S., K. Bai, K. Li, T. Wang. 2017. Study on the application of comprehensive water quality identification index method in Hailar River water quality evaluation. *Environmental Science and Management* 4:171–176.

Kawabe, M. and M. Kawabe. 1997. Factors determining chemical oxygen demand in Tokyo Bay. *Journal of Oceanography* 53:443–453.

Keith, D., J. Rover, J. Green, B. Zalewsky, M. Charpentier, G. Thursby, J. Bishop. 2018. Monitoring algal blooms in drinking water reservoirs using the Landsat-8 Operational Land Imager. *International Journal of Remote Sensing*. 39(9):2189–2842.

Kim, Y., J. Im, H. Ha, J. Choi, S. Ha. 2014. Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GIScience & Remote Sensing* 51(2):158–174.

Lei, S., J. Xu, Y. Li, C. Du, L. Li. 2020. An approach for retrieval of horizontal and vertical distribution of total suspended matter concentration from GOCI data over Lake Hongze. *Science of the Total Environment* 700:2–14.

Li, B., G. Yang, R. Wan, B. Liu, X. Dai, X. Chen. 2017. Temporal variability of water quality in Poyang Lake outlet and the associated water level fluctuations: A water quality sampling revelation. *Resources and Environment in the Yangtze Basin* 26(2):289–296.

Li, Y., Z. Zhang, J. Cheng, L. Zou, Q. Zhang, M. Zhang, Z. Gong, S. Xie, Y. Cai. 2021. Temporal and spatial variation of water quality in Hongze Lake from 2012 to 2018 and its causes. *Lake Science* 33(3):715–726.

Liang, K. and H. Yang. 2016. Spectral data correction method for the online detection of total nitrogen in water quality. *Chinese Journal of Environmental Engineering* 10(12):7397–7399.

Ministry of Ecology and Environment of the People's Republic of China. 2002. *GB 3838-2002. Environmental Quality Standards for Surface Water*. Beijing: Department of Science and Technology Standards, State Environmental Protection Administration.

Ministry of Environmental Protection. 1990. *GB 11892-89. Determination of Permanganate Index in Water Quality*. Beijing: Ministry of Environmental Protection.

Mohammad, H. 2021. Land–lake linkage and remote sensing application in water quality monitoring in Lake Okeechobee. Florida, USA. *Land* 10(147):1–3.

Mountrakis, G., J. Im and C. Ogole. 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* 66(3):247–259.

Naddeo, V., T. Zarra and V. Belgiorno. 2007. Optimization of sampling frequency for river water quality assessment according to Italian implementation of the EU Water Framework Directive. *Environmental Science & Policy* 10(3):243–249.

Qiao, L., J. Huang, J. Gao, Q. Huang, Y. Zhou, W. Tian. 2016. Spatial-temporal variation characteristics of chlorophyll-a concentration in Lake Hongze. *Lake Science* 28(3):583–591.

Rao, H. 2015. Determination of permanganate index in surface water by spectrophotometry. *Exchange of Scientific Papers and Cases* 7:48.

Sagan, V., K. Peterson, M. Maimaitijiang, P. Sidike, J. Sloan, B. Greeling, S. Maalouf, C. Adams. 2020. Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Reviews* 205:103187.

Shang, S., D. Feng, L. Tan, P. Liu, W. Guo. 2016. Temporal and spatial variation analysis of main driving factors of water ecosystem evolution in Jinan City. *Water Ecological Protection* 12:36–37.

Shang, S., X. Xiang, W. Guo, X. Yin, Z. Xu. 2021. Relationship between benthic functional feeding groups and environmental factors. *The People of the Yellow River* 6:77–79.

Shen, J., Z. Liu, X. Yang. 2020. *Limnology*. Beijing: Higher Education Press.

Tomar, D. and S. Agarwal. 2015. Twin support vector machine: A review from 2007 to 2014. *Egyptian Informatics Journal* 16(1):55–69.

Liu Y. 1985. UN issues new water quality standards and testing methods. *Environmental Engineering* 03:37-45.

Tan J., A. Keith and C. Indrajeet. 2015. Using hyperspectral data to quantify water quality parameters in the Wabash River and its tributaries, Indiana. *International Journal of Remote Sensing* 11(16):5467–5482.

Vakili, T. and J. Amanollahi. 2020. Determination of optically inactive water quality variables using Landsat 8 data: A case study in Geshlagh reservoir affected by agricultural land use. *Journal of Cleaner Production* 247:119134.

Vapnik, N. V. 1995. Adaptive and learning systems for signal processing, communications and control. In *The Nature of Statistical Learning Theory*, 138–167. New York: Springer-Verlag.

Wang, X., F. Shi, L. Yu, Y. Li. 2013. *MATLAB Neural Network Analysis of 43 Cases*. Beijing: Beihang University Press. 125-133.

Wu, C. 2018. *Remote Sensing Diagnosis of Ecological Health of Important Lakes and Wetlands in the Middle and Lower Reaches of Yangtze River*. Nanjing: Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences.

Xiong, J., C. Lin, R. Ma, Z. Cao. 2019. Remote sensing estimation of lake total phosphorus concentration based on MODIS: A case study of Lake Hongze. *Remote Sensing* 11(2068):2–17.

Xu Y., X. Dong and J. Wang. 2019. Comparison of chlorophyll-a concentration retrieval from four machine learning models in Taihu Lake. *Journal of Hydroecology* 7:1321–1326.

Xu, Y., C. Ma, S. Huo, B. Xi, G. Qian. 2013. Taking Cheng Hai as an example, support vector machine regression algorithm was used to predict chlorophyll A concentration. *Journal of Environmental Engineering Technology* 3:207–210.

Xun, D., D. Zhang. 2003. *Hongze Lake Volunteers*. Beijing: Local Chronicle Press.

Yang, C. 2020. *Water Quality Parameters Inversion and Eutrophication Evaluation of Hefei Huancheng River Based on Landsat-8 Satellite Image Data*. Hefei: Anhui Jianzhu University.

Yang, K. 2021. *Huai 'an City Ecological Environment Status Bulletin, Huai'an Ecological Environment Bureau, Huai'an, China*

Yang, L., K. Lei, M. Wei, F. Guo, W. Yan. 2013. Temporal and spatial changes in nutrients and chlorophyll-a in a shallow lake, Lake Chaohu, China: An 11-year investigation. *Journal of Environmental Sciences* 25(6):1117–1123.

Ye, C., C. Li, B. Wang, J. Zhang, L. zhang. 2011. Discussion on construction scheme of healthy water ecosystem in Hongze Lake. *Lake Science* 23(5):724–730.

Ye, T., L. Li, Y. Yao, L. Xia, W. Guan. 2019. Interannual variation of maximum turbidity zone in Hangzhou Bay based on Landsat images. *Journal of Wuhan University* 9:1377–1382.

Yu, L., F. Shi, H. Wang. 2015. *Intelligent Algorithm*. Beijing: Beihang University Press.

Zhang, Y., X. Qian, Y. Qian, J. Liu, F. Kong. 2009. Quantitative remote sensing of chlorophyll a in Taihu Lake based on machine learning method. *Environmental Science* 30(5):48–55.

Zhou, T., J. Zou, Y. Cui, X. Wang, C. Xie, P. Xia. 2022. Method and application of water extraction from remote sensing images based on principal component analysis and support vector machine. *Water Resources Protection* 1836(006) (in press).