Environmental Science and Pollution Research (2023) 30:79402–79422
https://doi.org/10.1007/s11356-023-27886-2

RESEARCH ARTICLE

A novel framework to predict chlorophyll-a concentrations in water bodies through multi-source big data and machine learning algorithms

Hamed Karimian[1] · Jinhuang Huang[1] · Youliang Chen[2] · Zhaoru Wang[3] · Jinsong Huang[4]

Received: 21 February 2023 / Accepted: 19 May 2023 / Published online: 7 June 2023

Abstract

Eutrophication happens when water bodies are enriched by minerals and nutrients. Dense blooms of noxious are the most obvious effect of eutrophication that harms water quality, and by increasing toxic substances damage the water ecosystem. Therefore, it is critical to monitor and investigate the development process of eutrophication. The concentration of chlorophyll-a (chl-a) in water bodies is an essential indicator of eutrophication in them. Previous studies in predicting chlorophyll-a concentrations suffered from low spatial resolution and discrepancies between predicted and observed values. In this paper, we used various remote sensing and ground observation data and proposed a novel machine learning–based framework, a random forest inversion model, to provide the spatial distribution of chl-a in 2 m spatial resolution. The results showed our model outperformed other base models, and the goodness of fit improved by over 36.6% while MSE and MAE decreased by over 15.17% and over 21.26% respectively. Moreover, we compared the feasibility of GF-1 and Sentinel-2 remote sensing data in chl-a concentration prediction. We found that better prediction results can be obtained by using GF-1 data, with the goodness of fit reaching 93.1% and MSE only 3.589. The proposed method and findings of this study can be used in future water management studies and as an aid for decision-makers in this field.

Keywords Remote sensing · Random forest · Water pollution · Eutrophication · Spatial analysis

Introduction

In recent years, sustainable development is among the hot topics that attract the concern of scholars from different fields. One of the main elements in sustainable development is environmentally friendly development (Han et al. 2023; Karimian et al. 2013; Wang et al. 2022). To reach this goal, monitoring and mitigation of pollutants in the environment is one of the main steps, and scholars have conducted researches to propose efficient methods to mitigate and monitor pollutant concentrations in the environment (Jiang et al. 2023; Xia et al. 2023). Water is an essential requirement in daily life; however, accessing drinkable water is a challenging task for many nations due to limited sources of freshwater (Ge et al. 2019). Water issues are believed to be one of the main reasons for security problems at the international level. Therefore, water-related issues attract scholars in different fields of specialty (Cai et al. 2017; Lv et al. 2022; Qiu et al. 2020). In recent years, with the increasing exploitation of environmental resources, the development of industry, and the high concentration of urban population, a large amount of industrial wastewater and domestic sewage has been discharged into lakes and rivers without proper treatment (Baladi et al. 2022; Hojjati-Najafabadi et al. 2020; Xu et al. 2023). This not only deteriorates the water quality due to the growth in the concentrations of

Responsible Editor: V.V.S.S. Sarma

✉ Youliang Chen
chenyouliang@jxust.edu.cn

[1] School of Marine Technology and Geomatics, Jiangsu Ocean University, Lianyungang 222005, China

[2] School of Civil and Surveying Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China

[3] School of Resources and Environmental Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China

[4] Zhejiang Zhipu Engineering Technology Co., Ltd, Huzhou 313000, China

oxygen-consuming substances such as nitrogen, phosphorus, and potassium in water but also damages the ecosystem (Karimian et al. 2022, Marie and Gallet 2022). Algal bloom is one of the important water pollutants that its concentration is increased by the existence of eutrophication in water. Rising temperatures indirectly affect water temperature, forming a thermal partition of the water column, changing the structure of biological communities (Cho et al. 2018; Zhou et al. 2020), and increasing the nutrients for algal bloom (Wan et al. 2020). It is believed that the outbreak of blue-green algal bloom in water bodies is mainly due to two factors: anthropogenic factor and climate conditions.

Chlorophyll-a (chl-a) is an important indicator of existing algal bloom in water (Mamun et al. 2020). As the feasibility of remote sensing has been proved in monitoring other environmental pollutants (Chen et al. 2022a; Guan et al. 2019a), scholars have conducted a series of studies based on remote sensing techniques to estimate chlorophyll-a concentrations in water bodies. The methods used can be broadly grouped into four categories. (1) The first method is the fluorescence peak/reflection peak algorithm. This method was proposed by Gower (1980), who measured the concentration of chlorophyll-a in water using a spectrometer and found that the chlorophyll-a fluorescence peak signal exhibited high reflection characteristics at a wavelength of around 685 nm. He also pointed out the possibility of inverting chlorophyll-a concentration based on fluorescence peak/reflection peak signals. However, the wavelength of the fluorescence/reflection peak signal of chlorophyll-a concentration is not unique and it may happen in different wavelengths in the fluorescence spectrum. (2) The second method is the empirical model, which was proposed by Odermatt et al. (2012). This method is based on the combination of spectral bands that can map chl-a concentrations. This combination can be included two or several bands. Scholars have conducted research on this, and the results show that the combination of more bands can yield better results (Moradi 2014). (3) The third method is the water pollution index algorithm, which was applied by Hu et al. (2012) to inverse chlorophyll-a concentration earlier, such as the normalized digital vegetation index (NDVI), enhanced vegetation index (EVI), and floating algal index (FAI). Its basic principle is that chlorophyll-a concentration in water exhibits spectral characteristics similar to terrestrial vegetation. (4) The fourth method is APProach by Elimination (APPEL) method, which was proposed by Anas et al. (2012). Its basic principle is to use the reflection spectral characteristics of strong absorption in the near-infrared band of water, while chlorophyll-a exhibits high reflectance spectral characteristics to obtain concentration information of chlorophyll-a. This method has achieved good results in inverting chlorophyll-a concentration in rivers and lakes (Ali et al. 2014; Murugan et al. 2014).

In recent years, machine learning techniques have gained remarkable attention especially due to their feasibility in discovering trend in nonlinear phenomena (Chen et al. 2022b; Fang et al. 2022; Li 2022; Wu et al. 2018). It was demonstrated that by carefully considering influencing factors, models can yield acceptable results (Guan et al. 2019b; Karimian et al. 2020). Following other environmental issues, machine learning approaches have been also practiced in water-related studies (Chen et al. 2020). With the gradual application of machine learning algorithms to the inversion of chlorophyll-a concentrations, superior inversion results have been achieved in reservoirs (Kupssinskü et al. 2020), lakes (Li et al. 2021; Zhang et al. 2022), rivers (Chen et al. 2021), and seas (Awad 2014). Random forest, a representative model based on Bagging, is one of the current popular machine learning algorithms and a typical representative of the integration algorithm, which has shown its feasibility in many fields. As an efficient machine learning algorithm for classification and regression modelling, the random forest has advantages over decision trees in solving practical problems, is less likely to cause overfitting problems, and can better handle small sample data sets (Mo et al. 2021). Based on this, several scholars have applied this technique to monitor water quality. Maciel et al. (2021) assessed water quality in Brazilian using machine learning algorithms and empirical models based on Sentinel-2 imagery and water quality data. They showed that the random forest results were more accurate and better results can be achieved through that. Zhou et al. (2021) used least squares, support vector machines, and random forests to establish an inversion model to study the soil heavy metals. They claimed that the accuracy of the random forest model was significantly higher than the other two models. Ao et al. (2019) discussed the learning ability and robustness of the random forest in-depth and confirmed the superiority of the random forest model over linear regression models.

Although there are several attempts to monitor algal bloom from space, most of these studies suffer from low spatial resolution and low inversion accuracy. Moreover, they seldom used big data from various sources. Therefore, in this study, we propose a novel algorithm to construct a chlorophyll-a concentration inversion model. Moreover, we investigate the spatial distribution of chl-a and the influencing factors. This study aims to provide data support for ecological environmental protection and algal bloom management and accumulate reference materials for ecological departments to have more efficient water management.

Materials and methods

Study area

The study area is the Huancheng River located in Zhejiang Province, and it is one of the sources of Taihu Lake, the third largest lake in China (Fig. 1). The river is located in the subtropical region with a mild and humid climate. The temperature in summer is high, and the precipitation is low that provides favorable conditions for the formation of algal bloom. According to the meteorological data, during our study period, the highest temperature in the area was 40℃, and the lowest temperature was 11℃.

Data source and pre-processing

Water quality data and pre-processing

To comprehensively collect the chl-a concentrations in the study area, a group of monitoring stations was set up in the Huancheng River at certain distances; the distance between each group of monitoring stations is approximately 500 m. Each group included three monitoring points mainly for monitoring the chlorophyll-a concentration values in the center and both banks of the river, as shown in Fig. 1. In total, 33 monitoring points were set up in the Huancheng River. We used water quality sensor AP-7000 to collect water quality data from May 1 until September 30, 2020. The collection time was set from 9:00 to 18:00 to meet the time for satellite overpass time. Sample point data recording mainly includes the following two steps: (1) use the device's built-in GPS module to collect geographic coordinates and display data in real-time using Aquameter, which can also be used to calibrate electrodes and assign each optional sensor to the appropriate AUX interface; (2) use the Aquaread standard output, SDI-12, and RS485 interfaces to connect to any telemetry or data logging devices via the probe for long-term recording. The collected data is transmitted to the server via the 4G DTU and saved in a MySQL database. Due to the impact of extreme weather, network transmission stability, etc. (Lv et al. 2022), it is necessary to pre-process the data collected by the AP-7000 sensor, including the average of the 2 data collected simultaneously, eliminating outlier data, etc.

Remote sensing data and pre-processing

The remote sensing images used in this paper are from Gaofen-1 and Sentinel-2 dataset. The GF-1 PMS camera can acquire panchromatic images in 2 m and multispectral

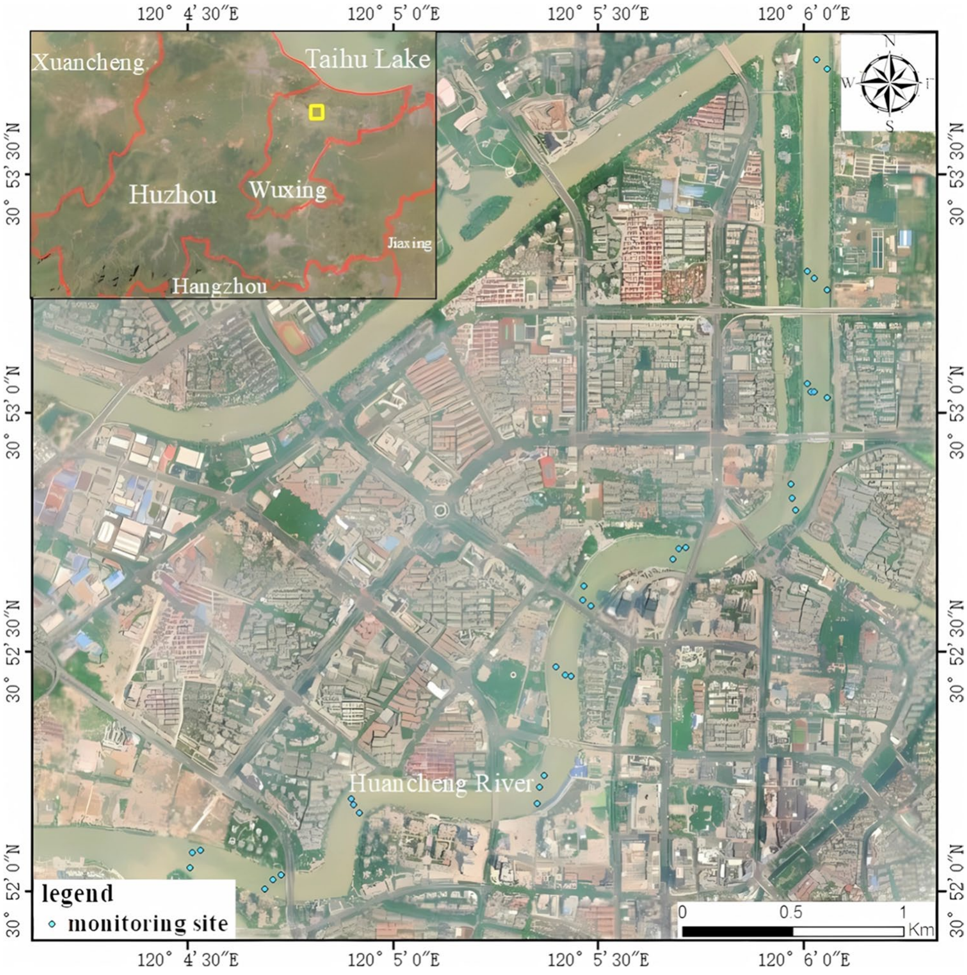


Fig. 1 The geographical location of the study area and the distribution of chlorophyll-a concentration monitoring stations

images in 8 m. During the study period, only two images of GF-1 data were available, on May 31 and July 22, 2020. Sentinel-2 carries a multispectral imager (MSI) that collects 13 spectral bands with ground resolutions of 10 m, 20 m, and 30 m, respectively, offering advantages such as high temporal resolution and multi-band combination. Sentinel 2A (L2A level) and Sentinel 2B (L2A level) data were imaged on May 13, August 1, August 11, August 16, September 5, and September 30, 2020.

The preprocessing of GF-1 image data mainly includes radiometric calibration, atmospheric correction, resampling, and land and water separation. Due to the L2A level of Sentinel-2 image data collected, there is no need for radiometric calibration and atmospheric correction. The preprocessing of Sentinel-2 image data mainly includes resampling and land water separation.

Radiation calibration converts the digital number (DN) which is assigned to each pixel to radiance (Eq. 1). This can be done through the calibration coefficient which is normally provided in image meta data (Song et al. 2022).

$$\mathrm{L} = \mathrm{DN}/\mathrm{g} + L_0 \tag{1}$$

where L is the radiance, g is the calibration coefficient and L_0 is the offset.

Because the solar radiation reflected from the surface is affected by atmospheric molecules and aerosols during atmospheric transmission (Karimian et al. 2016), atmospheric correction is used to obtain the actual surface reflectance (Li et al. 2020). In this study, atmospheric correction of multispectral data is performed based on the FLAASH (Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes) model. Resampling is done by merging GF-1 multispectral images with panchromatic band images to obtain 2 m resolution products, and Sentinel-2 images are resampled to produce 10 m resolution products.

Water and land separation is a segmentation operation on images and can be done using multi-scale segmentation. The purpose of multi-scale segmentation is to obtain vector files of river boundaries. Although GPS provides coordinates in high accuracy (Sun et al. 2021) and they were collected during the data collection stage, they are located inside the river, and we cannot obtain the vector boundaries of the river through GPS coordinates. Multi-scale segmentation is a method of classifying remote sensing images based on object-oriented idea (He et al. 2022). It takes into account the spectral characteristics and shape characteristics of an image and uses each pixel in the image as the starting point to divide the image into polygonal objects with different attributes through applying a top-down iterative merging algorithm (Yang et al. 2014; Zhang et al. 2013). This algorithm aims to ensure the homogeneity between the pixels within an object is maximized. In the analysis of remote sensing images, the spectral features directly affect the quality of segmentation results. The normalized difference water index (NDWI) is used to extract the vector boundary of the river. During the experiment, the effect of segmentation is used to find the suitable threshold interval of NDWI, and the study area vector is obtained by merging the segmented objects several times. The whole process was done in eCogntion 9.0 software to extract river waters. Finally, we used ENVI5.3 software to realize the extraction of remote sensing reflectance information for monitoring points.

$$\mathrm{NDWI} = \frac{\rho_{NIR} - \rho_{MIR}}{\rho_{NIR} + \rho_{MIR}} \tag{2}$$

where ρ_{NIR} is the near-infrared band reflectance and ρ_{MIR} is the mid-infrared band reflectance.

Inverse model of chlorophyll-a concentration

Research framework

The general idea of the random forest inversion of chlorophyll-a concentration in the Huancheng River is as follows: First, the remote sensing dataset corresponding to the collection date of the measured water quality data is screened out, among which 80% of the dataset is used to build the model, and the remaining 20% is used to evaluate the accuracy of the inverse model. Second, a suitable band combination was constructed based on the spectral characteristics, and a random forest inversion model was established. Moreover, we compare the performance of our proposed model with the empirical model and APPEL model to determine the most feasible model. Finally, the spatial and temporal distribution of chl-a concentrations in the Huancheng River is obtained. The specific process is shown in Fig. 2.

Chlorophyll-a concentrations through band combinations

Based on the remote sensing images of GF-1 and Sentinel-2, the spectral reflectance of the monitoring stations in the study area was extracted and was plotted with the wavelength as the X-axis and the reflectance (data value) as the Y-axis (Fig. 3). Figure 3 shows the spectral reflectance of each monitoring station, with each color representing one monitoring station. Our purpose in doing so is to display the overall reflectance of chlorophyll-a concentration in the station. According to the spectral curve of GF-1, it can be seen that there is an absorption valley in the spectral curve of chlorophyll-a at the wavelength of 680 nm (Fig. 3a); from that, we can infer the presence of chlorophyll-a on the surface of a water body (Juarez et al. 2008). Moreover, the reflectance at the near-infrared band has a certain elevation (Dall'Olmo et al. 2005). Compared

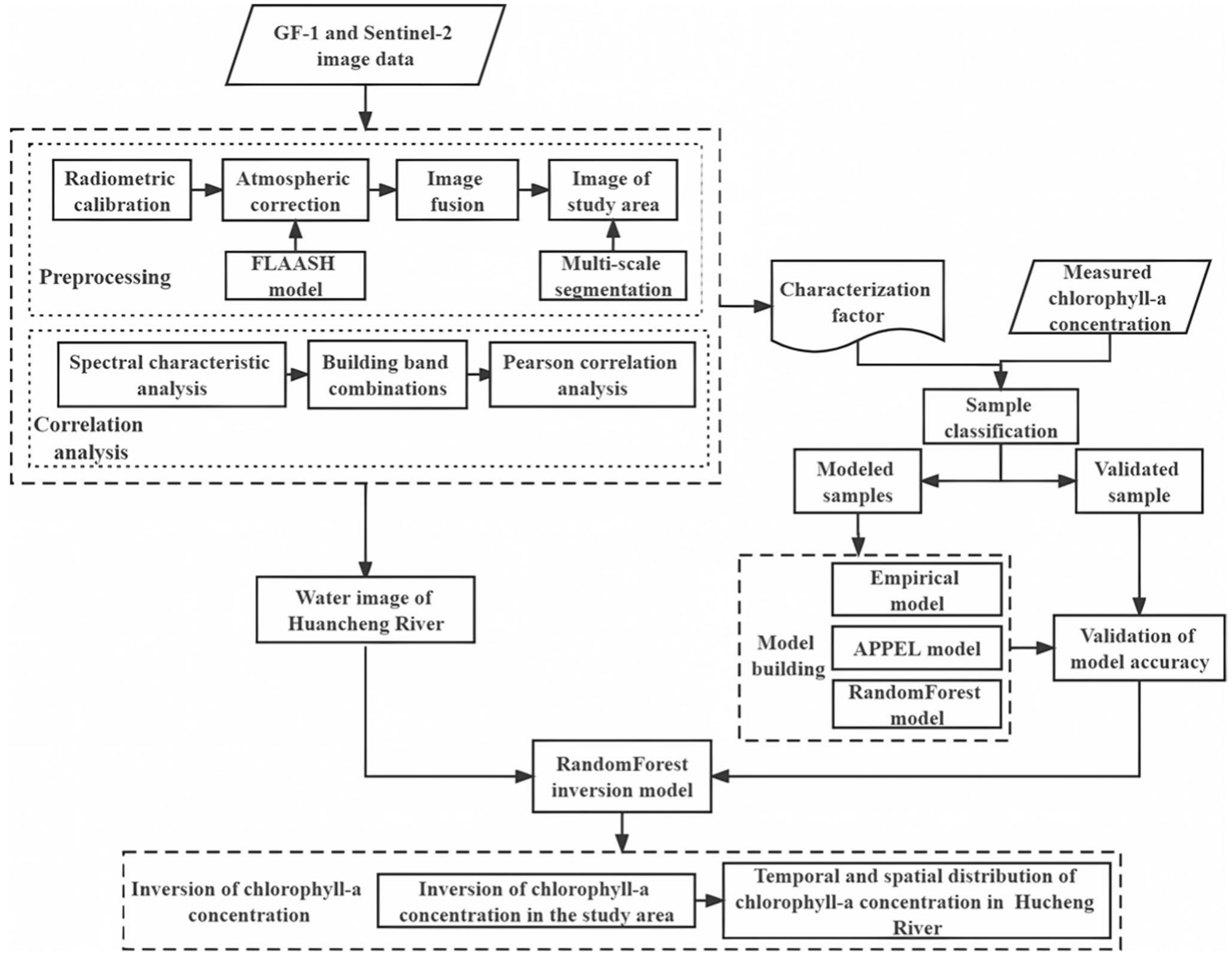
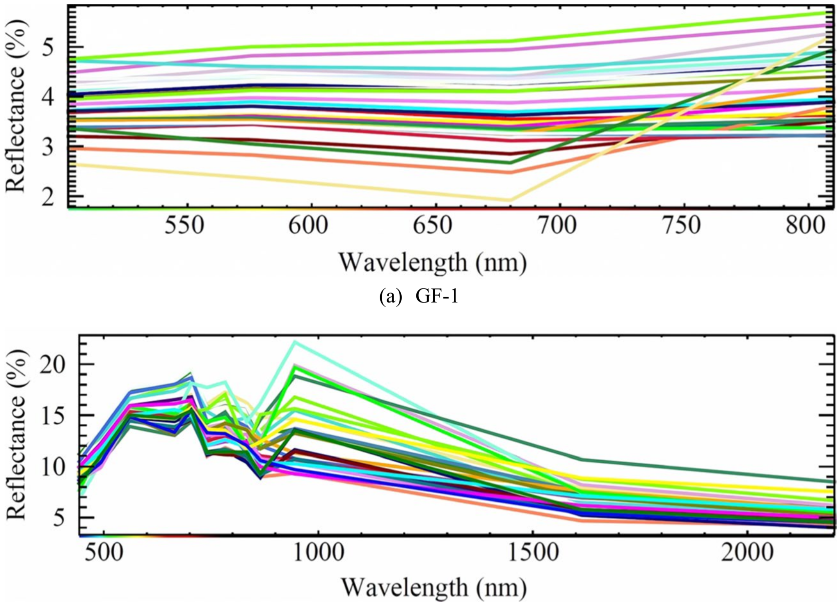

GF-1 and Sentinel-2 image data
Radiometric calibration
Atmospheric correction
Image fusion
Image of study area
FLAASH model
Multi-scale segmentation
Preprocessing
Spectral characteristic analysis
Building band combinations
Pearson correlation analysis
Correlation analysis
Characterization factor
Measured chlorophyll-a concentration
Sample classification
Modeled samples
Validated sample
Empirical model
APPEL model
RandomForest model
Model building
Validation of model accuracy
Water image of Huancheng River
RandomForest inversion model
Inversion of chlorophyll-a concentration
Inversion of chlorophyll-a concentration in the study area
Temporal and spatial distribution of chlorophyll-a concentration in Hucheng River

Fig. 2 The framework of chlorophyll-a retrieval in Huancheng River

Fig. 3 Reflection of spectral characteristics of monitoring stations based on remote sensing image data



(a) GF-1

(b) Sentinel-2

with GF-1 remote sensing image data, Sentinel-2 has a more detailed delineation of electromagnetic wavelengths. As shown in Fig. 3b, the spectral curve of chlorophyll-a concentration showed four reflection peaks and three absorption valleys. In the visible part, a reflection peak appears near 559 nm (corresponding to the green band) and an absorption valley near 664 nm (corresponding to the red band). In addition, the reflection peaks appear around 704 nm (corresponding to the B5 band), 782 nm (corresponding to the B7 band), and 945 nm (corresponding to the B9 band); the absorption valleys appear around 740 nm (corresponding to B6 band) and 864 nm (corresponding to B8A band), respectively. As can be seen from Fig. 3, an effect similar to the vegetation red edge appears between the reflection peak and the absorption valley. Therefore, the two bands corresponding between the reflection peak and the absorption valley are selected to construct a suitable inversion band model. In GF-1 data, bands b3 (680 nm) and b4 (810 nm) are selected, and in Sentinel-2 data, bands b2 (492.7 nm) and b3 (559.8 nm), b4 (664.6 nm) and b5 (704.1 nm), b6 (740.5 nm) and b7 (782.8 nm), and b8A (864.7 nm) and b9 (945.1 nm) are selected. A band inversion model is built by combining the two bands.

According to Fig. 3, it can be seen that the spectral curve of chlorophyll-a concentration in water exhibits an effect similar to the red edge of vegetation. Therefore, we construct band combinations in the form of common vegetation indices, such as the normalized difference vegetation index (NDVI) and enhanced vegetation indices (EVI).

To determine the correlation between chlorophyll-a concentration and band inversion models, correlation analysis was performed. In this study, the Pearson's correlation coefficient (Pearson) was used to determine the correlation between the two. The formula is as follows:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \tag{3}$$

where n is the number of samples, X_i and Y_i are the measured values of chlorophyll-a concentration and the reflectance of the band model at point i, respectively, and $\overline{X}$ and $\overline{Y}$ are the mean values of chlorophyll-a concentration and the mean values of the reflectance of the band model.

From the results in Table 1, the highest correlation between the reflectance of the b4/b3 combination and the measured chlorophyll-a concentration was observed in the GF-1 band combination, $p < 0.01$, indicating that the reflectance of the b4/b3 combination was considerably correlated with the measured chl-a concentrations within the 99% confidence interval, with r reaching 0.532. The results in Table 2 show that the reflectance of the (b9 − b8A)/(b9 + b8A) band combination had the highest correlation with the measured chlorophyll-a concentration, and the reflectance of the (b9 − b8A)/(b9 + b8A) combination was significantly correlated with the measured chlorophyll-a concentration at the 99% confidence interval, with r reaching 0.326. In summary, b4/b3 was chosen as the characteristic variable of the inverse model of chlorophyll-a concentration for GF-1 data, and (b9 − b8A)/(b9 + b8A) was selected as the characteristic variable of the inverse model of chlorophyll-a concentration for Sentinel-2.

Table 1 Correlation between GF-1 band combination and measured chlorophyll-a concentration

Band combination	Correlation coefficient
(b4 − b3)/(b4 + b3)	0.529**
1/(b4 − b3)	− 0.213
1/(b4 + b3)	− 0.338
b4/b3	**0.532****

Note: ** indicates a significant correlation at a confidence level (two-sided) of 0.01

Inversion method of chlorophyll-a concentration based on random forest algorithm

In the random forest technique, the decision tree is the basic unit. Its essential idea is the bagging method, which determines the outcome of the integrated evaluators by constructing multiple

Table 2 Correlation between Sentinel-2 band combination and measured chlorophyll-a concentration

Band combination	Correlation coefficient
(b3 − b2)/(b3 + b2)	0.051
1/(b3-b2)	− 0.031
1/(b3 + b2)	− 0.032
b3/b2	0.056
(b5 − b4)/(b5 + b4)	0.131
1/(b5 − b4)	− 0.041
1/(b5 + b4)	− 0.040
b5/b4	0.114
(b7 − b6)/(b7 + b6)	0.016
1/(b7 − b6)	− 0.173
1/(b7 + b6)	− 0.156
b7/b6	0.021
(b9 − b8A)/(b9 + b8A)	**0.326****
1/(b9 − b8A)	0.237
1(b9 + b8A)	− 0.280*
b9/b8	0.309*

Note: ** indicates a significant correlation at a confidence level (two-sided) of 0.01; * indicates a significant correlation at a confidence level (two-sided) of 0.05

mutually independent evaluators with the principle of average or majority voting. Random forest randomly selects individual decision trees during training. It converges to a lower generalization error as the number of evaluators increases, which has better robustness and is suitable for modelling and analyzing small sample data (Su et al. 2018). Therefore, this paper uses the random forest algorithm to establish the nonlinear relationship between the measured chlorophyll-a concentration and the spectral features.

During the experiment, first, given all the data sets:

$$\mathrm{D} = \left[\left(x_i, y_i\right)\right]\left(|\mathrm{D}| = \mathrm{n}, x_i \in R^m, y_i \in \mathrm{R}\right) \tag{4}$$

where n is the number of samples and m is the number of features per sample.

Divide all data sets into training and test sets, where the training set is:

$$S_T = \left[\left(x_1, y_1\right), \left(x_2, y_2\right), \ldots\ldots, \left(x_T, y_T\right)\right] \tag{5}$$

where T is the number of training sets.

The algorithmic framework of the random forest is shown in Fig. 4, with the following steps:

(1) Randomly select N times from the training set S_n by bootstrap method with put-back, one sample each time, resulting in N samples, and train a decision tree with the randomly selected N samples.
(2) Build multiple regression models (M_1,M_2,……,M_N) separately using the new training set obtained in step 1.
(3) Bring the test set into the trained regression tree model to get the predictive values [$M_1(X)$,$M_2(X)$,……,$M_N(X)$].
(4) The results of the predicted values of all regression trees are averaged, and the results are used as the final prediction of the random forest model.

In this study, a random forest algorithm was used to construct an inverse model of chlorophyll-a concentration using the Python language. The pre-processed image dataset is selected according to the sensor type. The input variable for the GF-1 inversion model is b4/b3, and the input variable for the Sentinel-2 inversion model is (b9 − b8A)/(b9 + b8A). The complete dataset is generated based on the measured chlorophyll-a concentration and the input variables of the model, and 80% of the dataset is randomly selected as the training data and the remaining 20% as the test data. During the random forest model training, the hyper parameters were selected using the RandomizedSearchCV provided by Scikit-learn. Table 3 provides the hyper parameter details for GF-1 and Sentinel-2.

After the model training is completed, the obtained model is applied to the corresponding remote sensing data to obtain the chl-a concentrations in those pixels without an in situ sensor. During the experiment, the pixel values of GF-1 and Sentinel-2 remote sensing images were extracted through

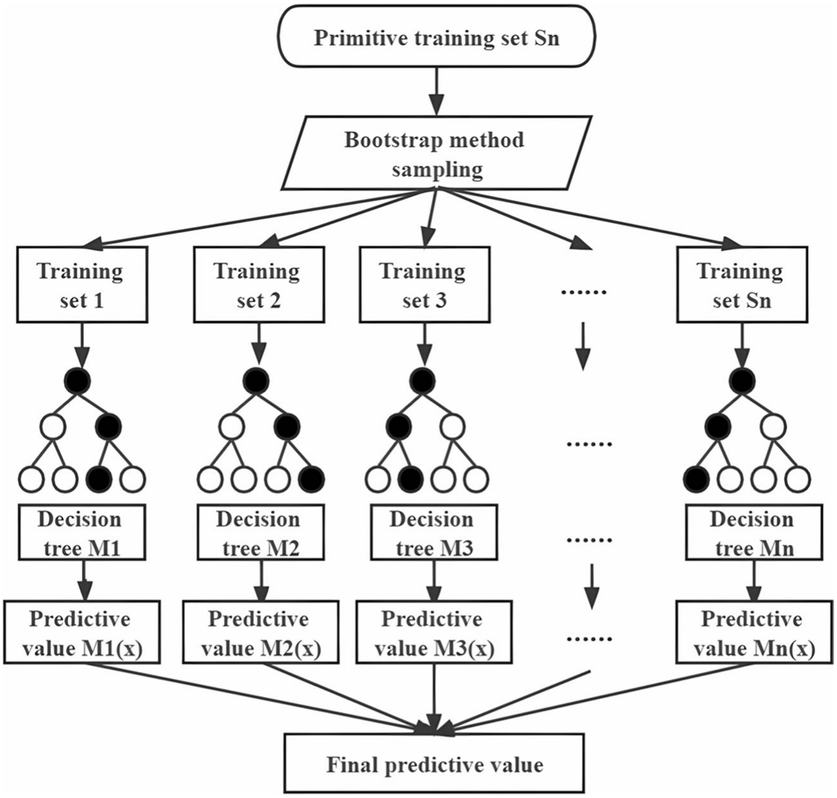


Fig. 4 Random forest algorithm framework

coding in Python. Through this step, all image element values of the b4/b3 band combination of GF-1 image and all image element values of the (b9 − b8A)/(b9 + b8A) band combination of Sentinel-2 image were extracted over the study area. It is worth mentioning that to avoid extra computation tasks, we used the segmented images in this stage which only contain the Huanghe River.

Model evaluation

Comparison based on empirical model inversions The empirical model establishes an equation algorithm for chlorophyll-a concentration mainly through the reflectance of band combinations and uses the obtained equation to invert the chlorophyll-a concentration for the study area (Dev et al. 2022; Rotta et al. 2021).

We select the b4 and b3 bands of GF-1, the near-infrared and red bands, respectively, and use the band ratio model (b4/b3) to generate the corresponding image data. The reflectance information of the monitoring stations of the two GF-1 images was extracted by ENVI5.3 software, and the extracted data were organized into a table containing reflectance data and measured chlorophyll-a concentration data, and 80% of the data were randomly selected for modelling analysis. Using reflectance as the explanatory variable and in situ measured chl-a concentrations as the dependent variable, a scatter distribution was established and a curve fit was used to construct an inverse model of chlorophyll-a concentration. Linear, exponential, logarithmic, quadratic polynomial, and multiplicative power inverse models of chlorophyll-a concentration were constructed by statistical regression analysis. The optimal model for chlorophyll-a concentration inversion was selected by the goodness of fit (R^2).

We select Sentinel-2 b8A and b9 bands, which are near-infrared (narrow) and water vapor bands, respectively, and generate the corresponding image data by the band combination model ((b9 − b8A)/(b9 + b8A)). The reflectance information of all monitoring stations of Sentinel-2 images was extracted by ENVI5.3 software to generate a data table with reflectance information and measured chlorophyll-a concentration. The same process as GF-1 treatment was used to establish the regression equation using scatter plots. The better-fitting curve equation was selected as the model for the inversion of chlorophyll-a concentration by Sentinel-2.

Table 3 Parameters of random forest model

Sensor type	Parameter	Value
GF-1	n	45
	max_depth	10
	max_feature	log2
	criterion	Poisson
	min_impurity_decrease	0
Sentinel-2	n	360
	max_depth	17
	max_feature	sqrt
	criterion	Poisson
	min_impurity_decrease	0

Validate the inversion of the random forest model based on the inversion results of the empirical model. First, we apply the empirical model to obtain chlorophyll-a concentration and compare the accuracy of the two models through evaluation indicators, such as R^2 and MSE. Second, the differences between the inversion effects of the random forest model and the empirical model are derived through comparative analysis. Finally, the inverse effect of the random forest model is evaluated.

Comparison of inversion based on the APPEL model Anas et al. (2012) proposed a model called APProach by Elimination (APPEL) for the inversion of chlorophyll-a concentration, which is a polynomial on the reflectance of the green band ($R_{rs}(\lambda_{green})$), red band ($R_{rs}(\lambda_{red})$), and near-infrared band ($R_{rs}(\lambda_{nir})$). Its basic principle is to obtain chlorophyll-a concentration information by using the property that water bodies show reflectance spectral features with strong absorption in the near-infrared band. In contrast, chlorophyll-a exhibits high reflectance spectral features. Related scholars (Ali et al. 2014; Oyama et al. 2015) have used the APPEL model to invert chlorophyll-a concentration in large lakes and obtain valid results.

The difference between the above two methods is that the empirical model estimates chlorophyll-a concentration by statistically analyzing the correlation between remote sensing data synchronized with groundwater quality analysis data, selecting the optimal band combination, and conducting statistical analysis of the correlation between remote sensing data synchronized with groundwater quality data. The APPEL model combines known spectral characteristics of water quality parameters (based on empirical knowledge that chlorophyll a concentration in water exhibits strong absorption characteristics in the near-infrared band) with statistical models and selects the optimal band as the relevant variable to estimate water quality parameter values.

The expression of the model is:

$$APPEL = R_{rs}\left(\lambda_{nir}\right) - \left[\left(R_{rs}\left(\lambda_{blue}\right) - R_{rs}\left(\lambda_{nir}\right)\right)^{*} R_{rs}\left(\lambda_{nir}\right) + \left(R_{rs}\left(\lambda_{red}\right) - R_{rs}\left(\lambda_{nir}\right)\right)\right] \tag{6}$$

$$C_{chl-a} = A + B * S_{appel} \tag{7}$$

where APPEL is the spectral index; $R_{rs}(\lambda_{blue})$, $R_{rs}(\lambda_{red})$, and $R_{rs}(\lambda_{nir})$ represent the reflectance of the blue band, red band, and near-infrared band, respectively.

The blue, red, and near-infrared bands of GF-1 data correspond to b1, b3, and b4 bands, respectively, and the APPEL model is used to generate the corresponding image data. The reflectance information of the two image data monitoring stations was extracted by ENVI5.3 software to create table data, and 80% of the data were randomly selected for analysis and modelling. The blue, red, and near-infrared bands of Sentinel-2 correspond to the b2, b4, and b8 bands, respectively, and are processed in the same way as the GF-1 data, which are used for modelling and analysis.

Inversion results are based on the APPEL model to verify the inversion effect of the random forest model. First, the goodness of fit of the two models when trained was compared by the R^2 index, and the APPEL model was applied to the inversion of chlorophyll-a concentration. Secondly, the differences between the inversion effect of the random forest model and the inversion effect of the APPEL model are derived through comparative analysis. Finally, the inverse effect of the random forest model is evaluated.

Evaluation indicators To evaluate the inversion accuracy of each inversion model, the remote sensing reflectance data corresponding to 20% of the measured chl-a data was used as the test dataset. It was evaluated using four indicators: coefficient of determination (R^2), mean square error (MSE), mean absolute error (MAE), and median absolute error (ME).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{8}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{10}$$

$$\mathrm{ME} = \mathrm{Median}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{11}$$

In the above, $\hat{y}_i$ is the inverse value of chlorophyll-a concentration, $\overline{y_i}$ is the mean of the measured chlorophyll-a concentration, y_i is the measured chlorophyll-a concentration value, subscript i indicates different stations and n is the number of samples.

Results

Inversion results of chlorophyll-a concentration based on random forest

In the experiments, the goodness of fit (R^2), root mean square error (MSE), mean absolute error (MAE), and median absolute error (ME) are used to evaluate the training results of the model. The training results of each sensor inversion model are shown in Table 4.

The inversion model was constructed using the results of the hyper parameter search and applied to the inversion of chlorophyll-a concentration in the Huancheng River based on the trained random forest model, and the inversion results are as follows:

On May 31 and July 22, the water quality in the river was good, with only minor algal bloom occurring. During the rest of the period, the river showed varying degrees of algal bloom and poor water quality conditions. On May 13, the overall chlorophyll-a concentration in the river was high, and the degree of algal bloom was more serious. Combined with the measured chlorophyll-a concentration, Fig. 5 reflects the overall chlorophyll-a concentration in the water quality of the Huancheng River during the period.

The main distribution of chlorophyll-a concentration values in the river: High values of chlorophyll-a concentrations are more often found along the river banks, and relatively low values of chlorophyll-a concentrations are found in the center of the river. Such a situation may be due to the influence of wind speed, direction, and water flow, which can easily aggregate the formed blue-green algal bloom to the riverbank, resulting in high chlorophyll-a concentration values on both sides of the river. For example, on July 22, the chlorophyll-a concentration on the bank of the river upstream was higher, 10 ~ 15 μg/L, and the chlorophyll-a concentration in the middle of the river was lower, less than 5 μg/L. The results of the inversion on August 1 showed that the high chlorophyll-a concentration area of the whole river appeared on the bank of the river, the overall concentration is 10 ~ 15 μg/L, there are also chlorophyll-a concentrations greater than 15 μg/L, and the chlorophyll-a concentration in the center of the river is less than 10 μg/L. Therefore, the algal bloom on that day gathered in the bank area of the river.

Inversion model evaluation

Comparison of inversion results based on empirical models

To select the most suitable empirical model, the goodness-of-fit (R^2) of each model was calculated, and the GF-1 empirical model had the highest goodness-of-fit for the quadratic model with R^2 of 0.507. The fitting results are

Table 4 Training results of random forest model

Sensor type	R^2	MSE (μg/L)	MAE (μg/L)	ME (μg/L)
GF-1	**0.889**	**2.196**	**0.959**	**0.686**
Sentinel-2	**0.812**	**0.974**	**0.802**	**0.649**

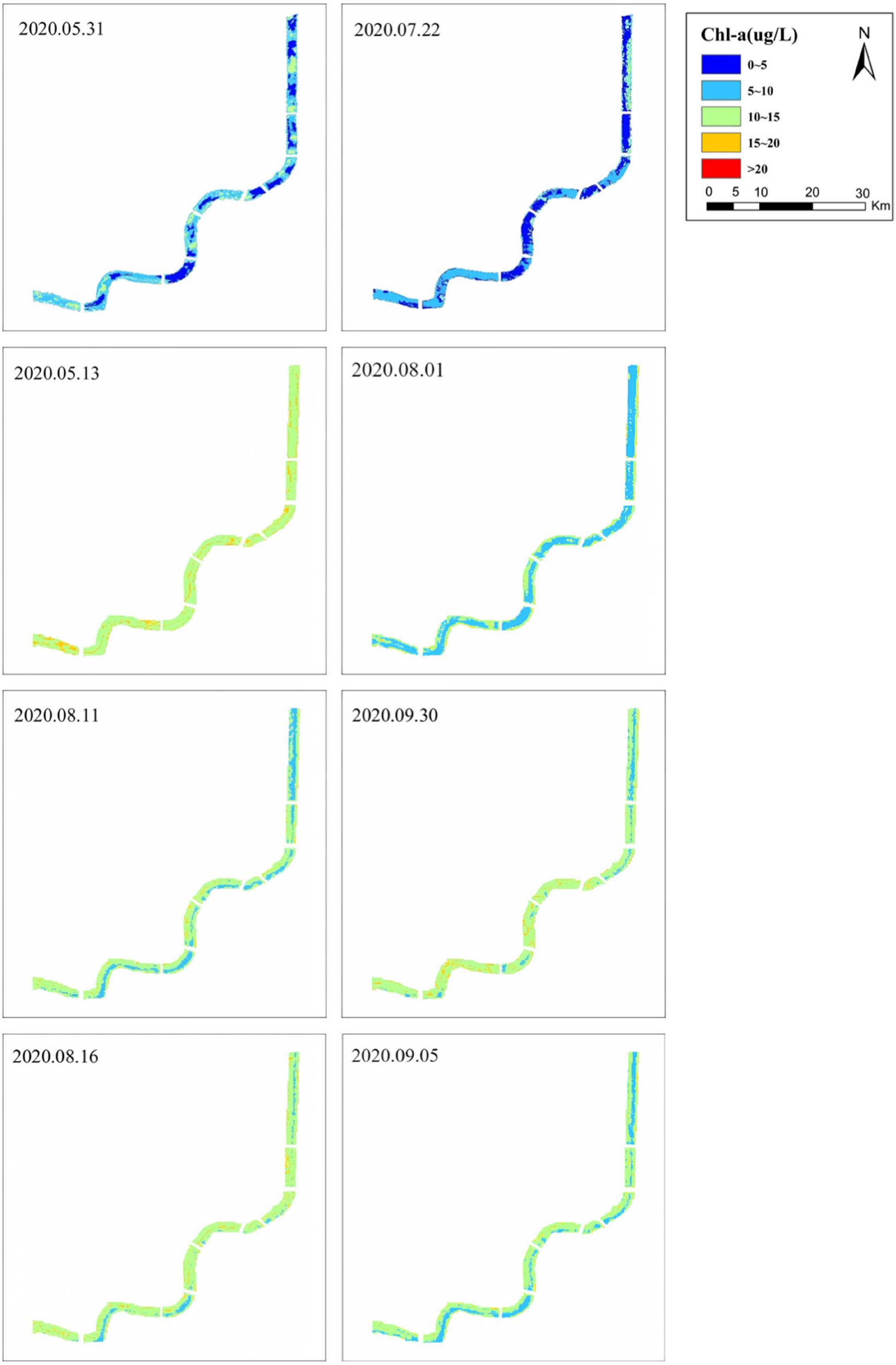


Fig. 5 Inversion results of chlorophyll-a concentration in Huzhou Huancheng River based on random forest model

shown in Fig. 6a. Compared with the other six models, it can more effectively invert the chlorophyll-a concentration in the Huancheng River. Therefore, the quadratic model was chosen as the empirical model for the inversion of chlorophyll-a concentration in GF-1. The calculation formula is as follows:

$$Y = -1.601X^2 + 13.459X - 4.657 \tag{12}$$

where *Y* is the chlorophyll-a concentration value and *X* is b4/b3.

There are some negative cases in the Sentinel-2 band combination data, so there is no logarithmic function and power function model, and only five models are fitted. The highest goodness of fit of the model is the quadratic model with R^2 of 0.246, and the fitting results are shown in Fig. 6b. Compared with the other three models, it is more effective in inverting the chlorophyll-a concentration, so the quadratic model was chosen as the empirical model for the inversion of chlorophyll-a concentration in Sentinel 2. The calculation formula is as follows:

$$Y = -3.531X^2 - 3.0931X + 13.065 \tag{13}$$

where *Y* is the chlorophyll-a concentration value and *X* is (b9 – b8A)/(b9 + b8A).

Based on the modelling results of the empirical model, it can be seen that the goodness of fit of the GF-1 and Sentinel-2 random forest models is significantly higher than the empirical models of GF-1 and Sentinel-2, as shown in Table 5. Among them, the R^2 index of the GF-1 random forest model is 38.2% higher than that of the empirical model, and the R^2 index of the Sentinel-2 random forest model is 56.6% higher than that of the empirical model.

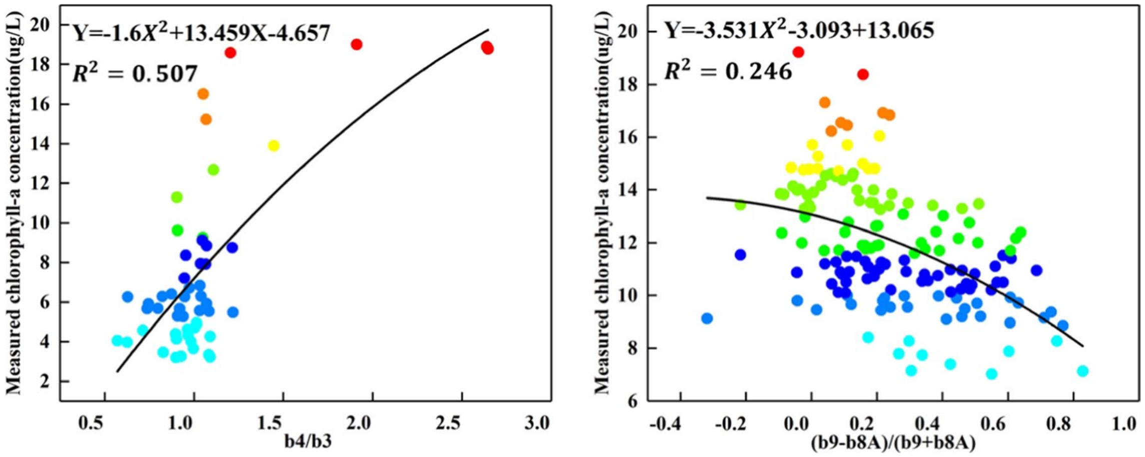
Application of empirical inversion model to GF-1 remote sensing data. From the inversion effect, it can be seen that on May 31, the inversion result of the empirical model showed that the chlorophyll-a concentration of the river was less than 10 μg/L (Fig. 7a), and the inversion result of the random forest showed that the chlorophyll-a concentration of the river was both greater than 10 μg/L and less than 10 μg/L on that day (Fig. 7b). Combined with the measured chlorophyll-a concentration on the same day, the measured concentration values were ranged from 3 to 19 μg/L. According to research investigations (Amorim et al. 2020; Qin et al. 2015), slight hydration occurs when chlorophyll-a concentration is greater than 10 μg/L. The inversion of the random forest model results in a little algal bloom in each small section of the river. In contrast, the empirical model inversion results in no algal bloom. On July 22, the inversion results of the empirical model showed a region of low chlorophyll-a concentration (less than 5 μg/L), as shown in Fig. 7c. The results of the random forest model inversion reflected both low-value areas and higher-value areas (> 10 μg/L) of chlorophyll-a concentrations, as shown in Fig. 7d, and therefore a slight water bloom phenomenon. The measured chlorophyll-a concentration on the day was less than 10 μg/L, and the measured values at a few stations were greater than 10 μg/L. The water quality condition was good on the whole. Based on the comparative analysis of the inversion results graph, the inversion results of the random forest model are closer to the results of the measured chlorophyll-a concentration cases, and the inversion effect is better.

In the inversion results on May 13, the empirical model inversions showed results between 10 and 15 μg/L (Fig. 8a). The inversion results of the random forest model showed regions greater than 15 μg/L, which are more finely represented in the resulting plot (Fig. 8b). On August 1, the inversion results of the empirical model appeared to have areas with chlorophyll-a concentrations less than 10 μg/L, but the resulting map showed slight overall hydrophobia (Fig. 8c). The inversion results of the random forest model are different from the empirical model. Although there are also chlorophyll-a concentrations below 10 μg/L, the area of the river below 10 μg/L is more extensive, the overall water quality is better, and the results are more consistent with the measured results (Fig. 8d). Therefore, the inversion results of chlorophyll-a concentration of Sentinel-2 were better in the random forest model.

Comparison of inversion results based on the APPEL model

The low goodness of fit of the APPEL model with chlorophyll-a concentration is illustrated in Fig. 9. Specifically, the goodness of fit of the APPEL model for GF-1 is only 0.001, and that of the Sentinel-2 model is only 0.004, as shown in Table 6. There is no clear trend in the fitted relationship plots of the APPEL models for GF-1 and Sentinel-2, both of which exhibit underfitting. Therefore, the obtained fits are poor. Therefore, the modelling results based on the APPEL model are not comparable to those of the random forest model due to the low R^2 index of the APPEL model.

To investigate more on the feasibility of the APPEL model in chl-a concentration prediction, the inversion results of the APPEL model on May 31 are illustrated in Fig. 10. As can be seen, the chlorophyll-a concentration in the river ranged from 5 to 10 μg/L, and no other concentration interval appeared. The results of the inversion of the random forest showed that the chlorophyll-a concentration in the river on that day was both greater than 10 μg/l and less than 10 μg/l, and the results of multiple intervals of chlorophyll-a concentration values appeared (Fig. 10 a and b). Combined with the measured data, the inversion results of the random forest model are closer to the chlorophyll-a concentration situation on that day. On July 22, the inversion results of the APEEL model were the same



(a) GF-1 (b) Sentinel-2

Fig. 6 Optimal fitting relationship between empirical model and measured chlorophyll-a concentration

as those on May 31, showing chlorophyll-a concentrations of 5 ~ 10 μg/L (Fig. 10c). The random forest model inversion results reflect that chlorophyll-a concentration has low-value areas. Areas with chlorophyll-a concentrations greater than 10 μg/L occur in parts of the river (Fig. 10d). As a result, there is a slight water bloom on the river.

In the inversion results on May 13, the inversion results of the APPEL model exhibited chlorophyll-a concentrations ranging from 10 to 15 μg/L, and the random forest model inversions showed regions greater than 15 μg/L (Fig. 11 a and b). On August 1, the inversion results of the APPEL model still showed that the chlorophyll-a concentration ranged from 10 to 15 μg/L. The inversion effect of the random forest model showed that the overall chlorophyll-a concentration was below 10 μg/L; there were relatively few areas larger than 10 μg/L (Fig. 11 c and d), and the overall water quality was better, which was more consistent with the actual measurement results.

Comparison of inversion results based on evaluation indicators

To test the feasibility of the random forest model, the remote sensing reflectance corresponding to 20% of the measured chlorophyll-a concentration data was substituted into the inversion model as the independent variable. The inversion values obtained by using the inversion model were compared with the measured values to evaluate the accuracy of the model. The empirical inversion model of GF-1 achieves an R^2 index of 0.565, and the data points of the test set are distributed overall on both sides of the function curve of $y=x$, with a small number of data points falling on the diagonal (Fig. 12a). The empirical model of Sentinel-2 has a large waviness, R^2 is only 0.194, the data points of the test set are scattered on both sides of the diagonal, and some data points deviate far from the $y=x$ function curve (Fig. 12d). The difference between the

Table 5 Fitting results of empirical inversion model

Sensor type	Model classification	Equation expressions	R^2
GF-1	Linear functions	$Y=8.283X-1.213$	0.499
	Exponential functions	$Y=2.789EXP\ (0.811X)$	0.439
	Logarithmic functions	$Y=11.181lnX+7.422$	0.480
	Quadratic functions	$\mathbf{Y=-1.601\ X^2+13.459X-4.657}$	**0.507**
	Power function	$Y=6.496\ X^{1.122}$	0.496
	Composite model	$Y=3.625\times1.933^X$	0.452
	Growth model	$Y=EXP\ (1.288+0.659X)$	0.452
Sentinel-2	Linear functions	$Y=-5.110X+13.168$	0.238
	Exponential functions	$Y=13.008EXP\ (-0.438X)$	0.231
	Quadratic functions	$\mathbf{Y=-3.531X^2-3.0931X+13.065}$	**0.246**
	Composite model	$Y=13.160\times0.655^X$	0.232
	Growth model	$Y=EXP\ (2.577-0.423X)$	0.232

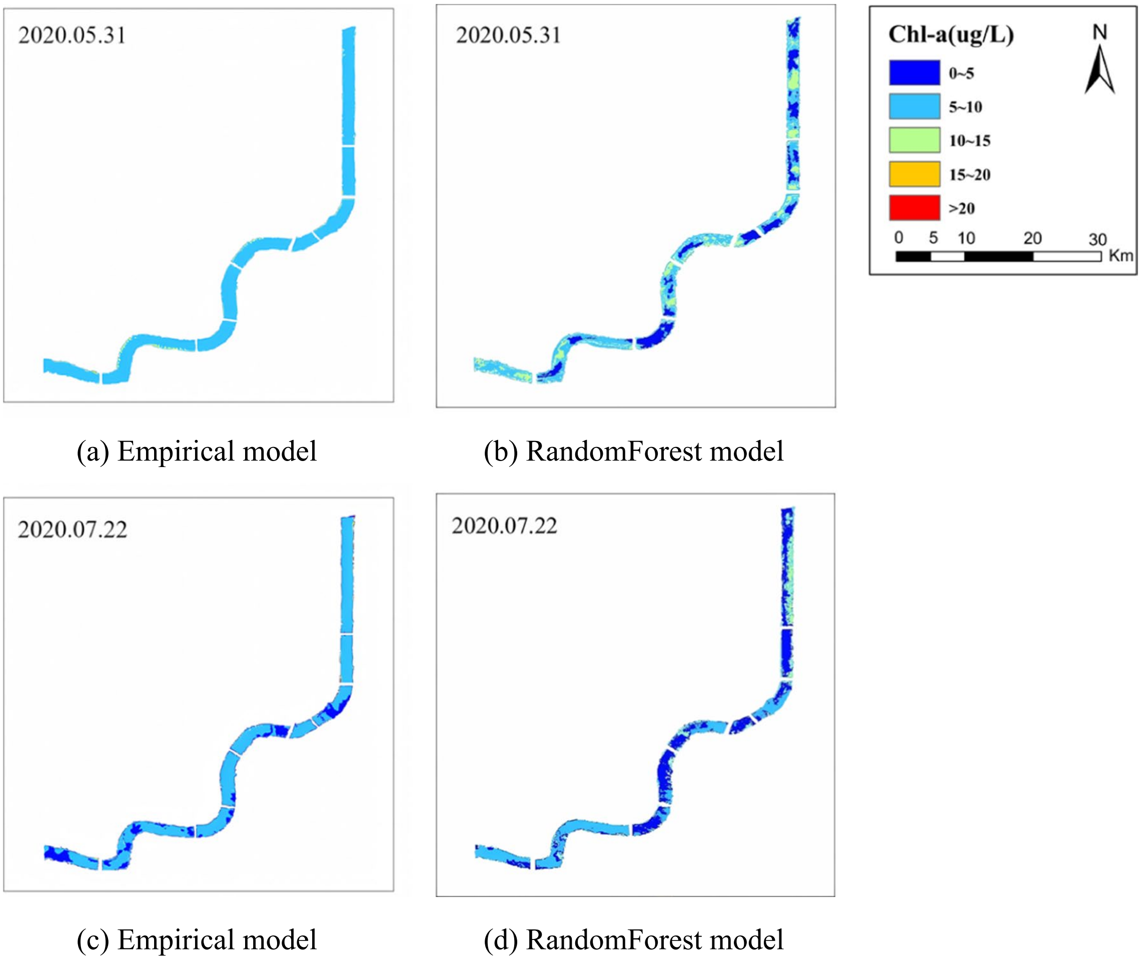


Fig. 7 Comparison of empirical model inversion results based on GF-1 data

error evaluation indexes of GF-1 and Sentinel-2 is slight. The APPEL models of GF-1 and Sentinel-2 have poor accuracy, low R^2, and significant error indicators. The red trend lines are almost parallel to the X-axis. The test datasets are not distributed along the diagonal but fluctuate above and below the red trend line (Fig. 12 b and e). Therefore, the APPEL model exhibits significant volatility and instability. The random forest inversion model of GF-1 is more stable. Compared with the other two inversion models, there is a more remarkable improvement in accuracy, with R^2 reaching 0.931 and relatively small error indicators. Compared to the empirical inversion model, more data points are falling on the diagonal, as in Fig. 12c. Sentinel-2's random forest inversion model has significantly improved R^2 compared with the previous two models. The error evaluation indexes are all less than 1, as shown by the more evenly concentrated distribution of the test data set on both sides of the diagonal (Fig. 12f). Compared to the empirical inversion model, no data points are far from the diagonal. Therefore, the random forest inversion model is more accurate, and the model is more stable.

In order to more intuitively display the accuracy of each model, we have summarized the evaluation indicators of the model in Tables 7 and 8.

Inversion model validation

To further validate the effect of the random forest model, we substitute the band combination reflectance of the test set as the dependent variable in each inversion model. The results from the random forest model are analyzed by comparing the validation values obtained from the random forest model with the measured values and the validation values from the empirical and APPEL models.

The comparison between the measured and validated chlorophyll-a concentrations based on GF-1 remote sensing data shows that the validated and measured values of the empirical model are in good agreement at some points. At other points, the validation value of the empirical model is high when the measured value is low and low when the measured value is high, as shown in Fig. 13.

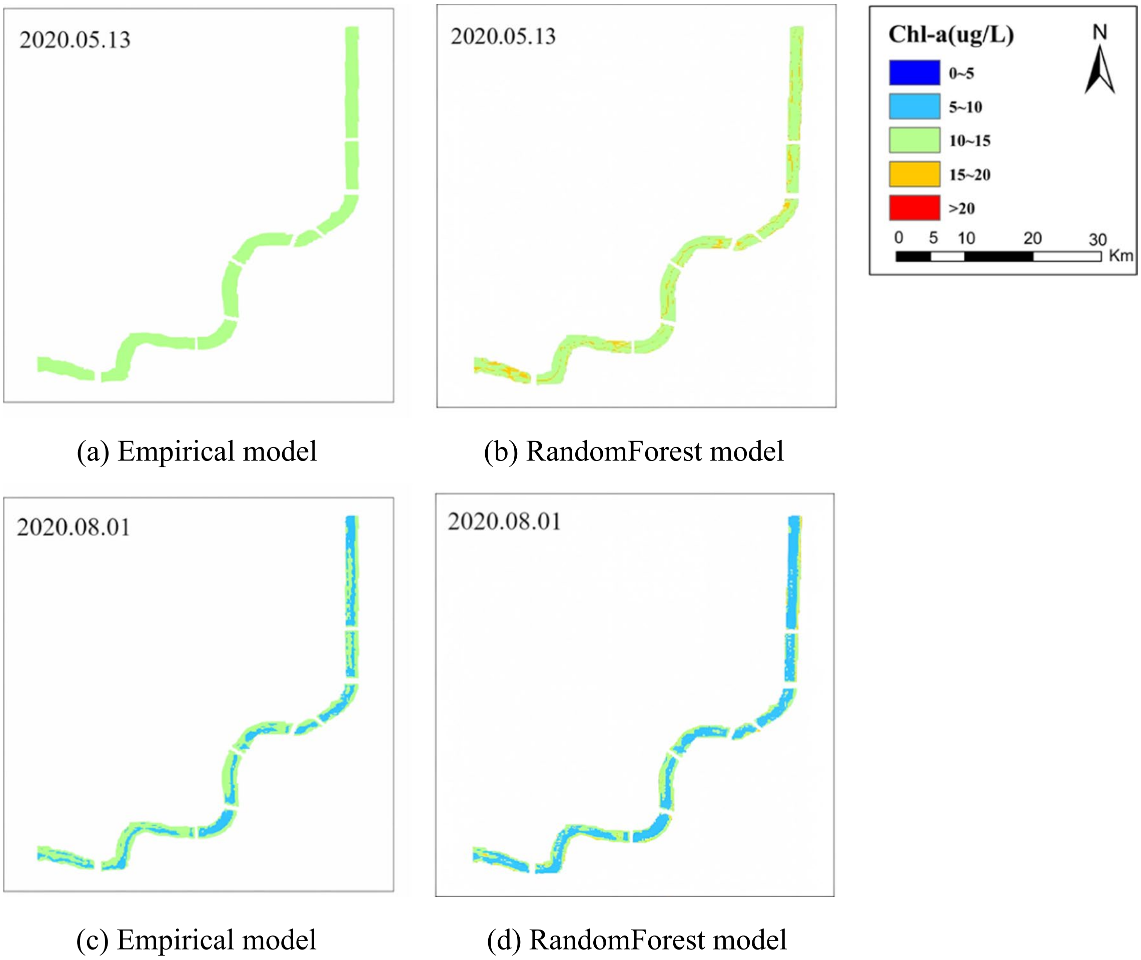


Fig. 8 Comparison of empirical model inversion results based on Sentinel-2 data

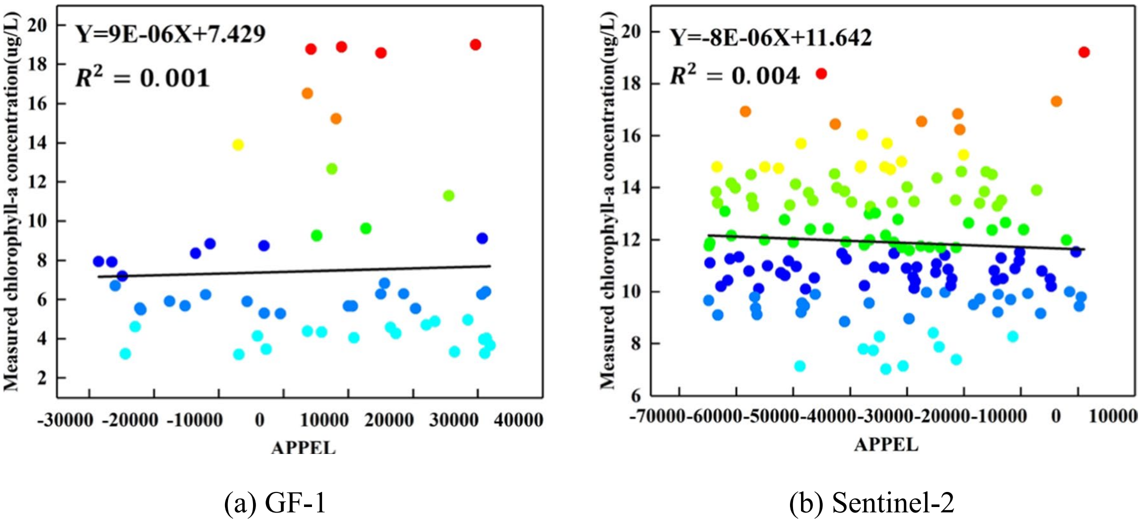


Fig. 9 Relation between the APPEL model and measured chlorophyll-a concentration. *y* is the chlorophyll-a concentration and *x* is the spectral index of APPEL

Table 6 Fitting results of the APPEL inversion model

Sensor type	Equation expressions	R^2
GF-1	$y = 9\text{E-}06x + 7.429$	**0.001**
Sentinel-2	$y = -8\text{E-}06x + 11.642$	**0.004**

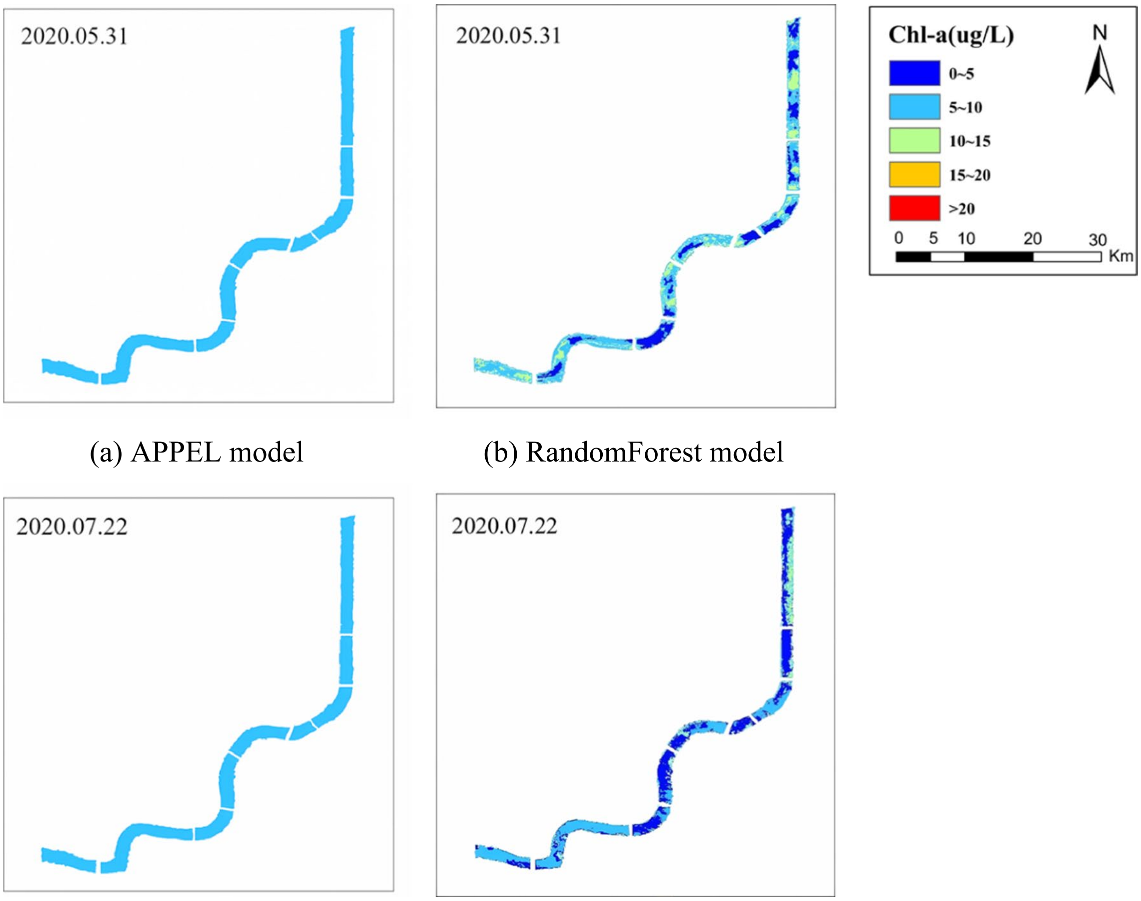
The curve trend of APPEL model validation values was relatively flat, without any fluctuation trend. The general trend was quite different from the actual measured chlorophyll-a concentration values. The trend between the random forest model validation values and the measured chlorophyll-a concentrations remained the same, reflecting the same increase and decrease.

In comparing the validated and measured values of each model based on Sentinel-2 remote sensing data, the curve trend between the validated and measured values of the empirical model was more consistent when the measured values of chlorophyll-a concentration were in the range of 10 to 13 μg/L. When the measured value is high or low, the validated value of the empirical model differs from the measured value. The validated values of the APPEL model fluctuated slightly between about 12 μg/L and the actual measured values. The difference between the validated and measured values of the random forest model is slight, the curve trend of the two is more consistent without large fluctuations, and the effect of the random forest model is better, as shown in Fig. 14.

Discussion

Inversion model feasibility analysis

In the modelling results, the accuracy of the empirical inversion model of GF-1 reached 0.565. In contrast, the empirical model of Sentinel-2 had a lower accuracy of 0.194, which might be due to the higher resolution of GF-1 remote sensing images than Sentinel-2 and the smaller data volume of GF-1. The modelling method of APPEL



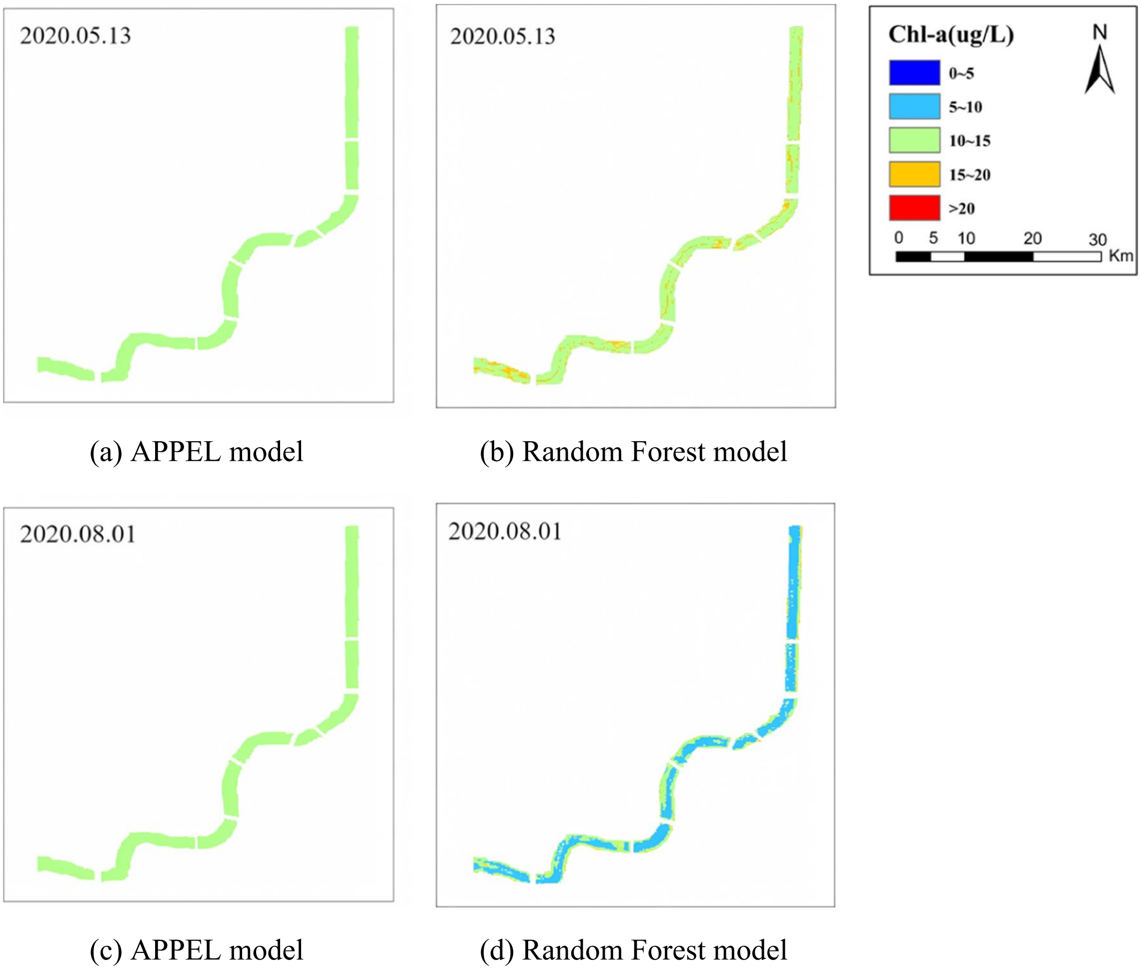
(a) APPEL model

(b) RandomForest model

(c) APPEL model

(d) RandomForest model

Fig. 10 Comparison of APPEL model inversion results based on GF-1 data



(a) APPEL model

(b) Random Forest model

(c) APPEL model

(d) Random Forest model

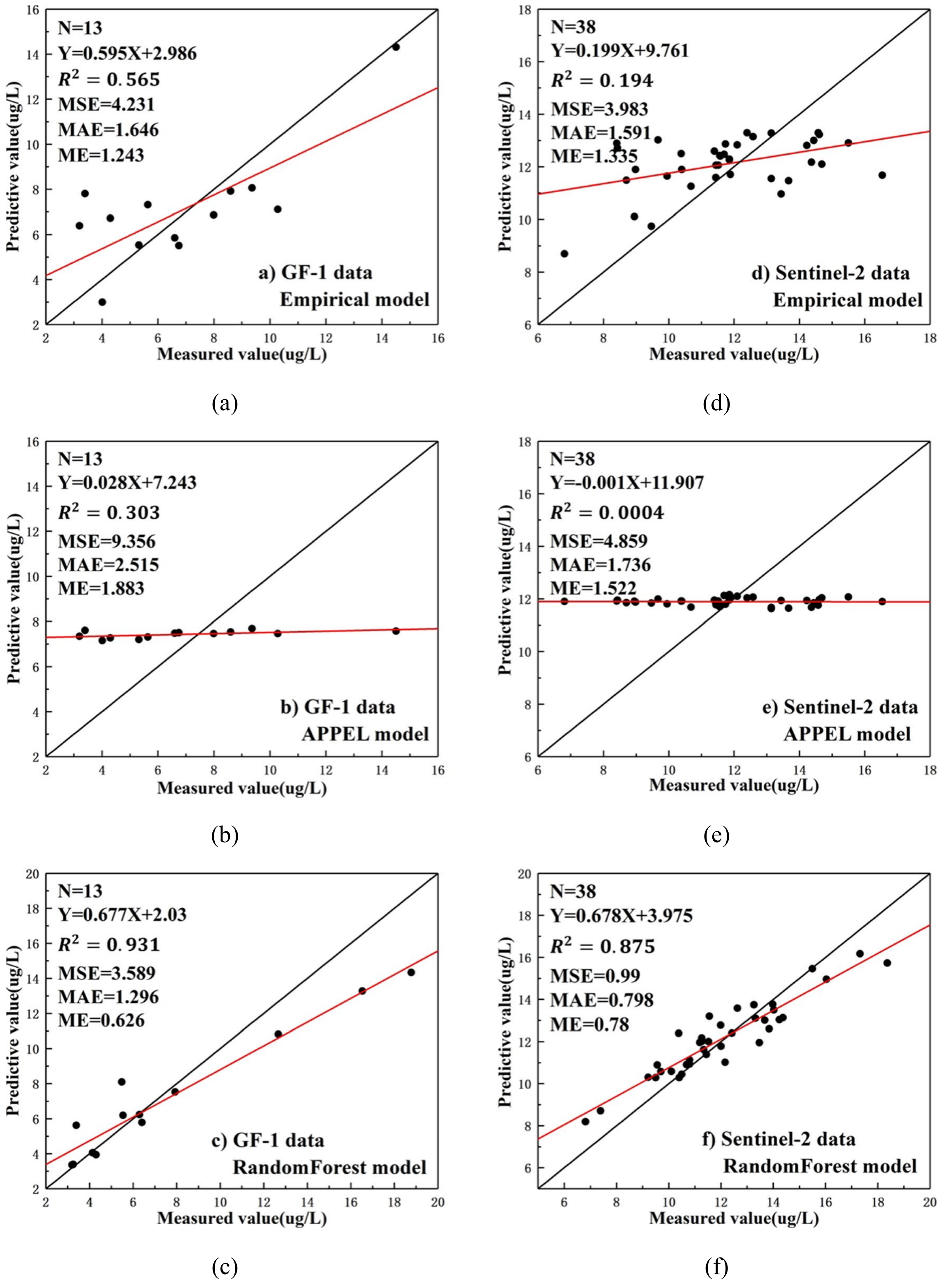
Fig. 11 Comparison of APPEL model inversion results based on Sentinel-2 data

failed to achieve the inversion of chlorophyll-a concentration in small watersheds, probably because the APPEL model was proposed for MODIS data. In contrast, the sensitive bands of chlorophyll-a concentration differed for different sensor data (Moradi 2021), indicating that the model was not generalized. Based on the modelling method of random forest, the model accuracy of GF-1 and Sentinel-2 is improved, and the random forest results are closer to the measured situation in the inversion effect.

In this study, only 5 months of measured data were available, and only 8 remote-sensing images were available in the corresponding period. Therefore, it was impossible to utilize more data for modelling, and the trained model could only be applied for short-term inversion of chlorophyll-a concentration. The inverse effect of the model for chlorophyll-a concentration in other months is unknown. Therefore, the model constructed from small sample data has some limitations. The following research is to obtain more data volume, match to more remote sensing data sources, and make up for the lack of petite sample data modelling. And the means of big data mining can be adopted, using more machine learning algorithms to establish the link between the actual measurement data and remote sensing band information, from which the relationship between the two is sought.

Causes of algal bloom and control measures

From the inversion results, we can see that the algal bloom in the Huzhou Huancheng River is mainly concentrated in summer, primarily influenced by climatic conditions (Sha et al. 2021). Therefore, local authorities have taken several measures to reduce the risk of algal bloom outbreaks. Their efforts mainly include installing aeration devices in the river, salvage by boat, and camera monitoring along the river. Installation of aeration devices, mainly by increasing dissolved oxygen concentration in the water, alleviates the degree of algal bloom in the river aggregation (Visser et al. 2016).



(a) (d)

(b) (e)

(c) (f)

Fig. 12 Accuracy evaluation results of each inversion model

Table 7 Accuracy of each model based on GF-1 data

Model	R^2	MSE (µg/L)	MAE (µg/L)	ME (µg/L)
Empirical models	0.565	4.231	1.646	1.243
APPEL model	0.303	9.356	2.515	1.883
Random forest model	**0.931**	**3.589**	**1.296**	**0.626**

Table 8 Accuracy of each model based on Sentinel-2 data

model	R^2	MSE (µg/L)	MAE (µg/L)	ME (µg/L)
Empirical models	0.194	3.983	1.591	1.335
APPEL model	0.0004	4.859	1.736	1.522
Random forest model	**0.875**	**0.99**	**0.798**	**0.78**

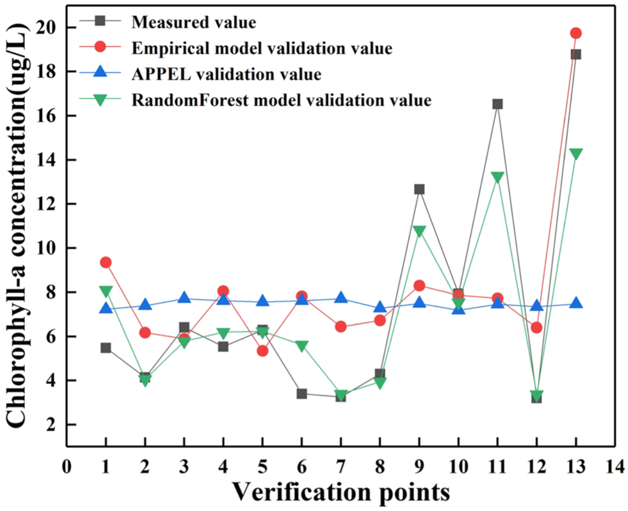


Fig. 13 Line chart of the measured value and verified value of chlorophyll-a concentration based on GF-1 remote sensing data

Local departments in Huzhou City organized fishermen and boats as emergency forces for algal bloom salvage based on professional salvage boats and tools. Regularly clean the river and salvage algal bloom during water bloom outbreaks (Fig. 15). In addition, camera devices are installed on both sides of the river to take pictures of the river's water quality. Once algal bloom appears in the river, the device delivers a risk alert to the management.

Conclusions

This paper constructed a random forest inversion model of chlorophyll-a concentration based on GF-1 and Sentinel-2 remote sensing data and actual measured

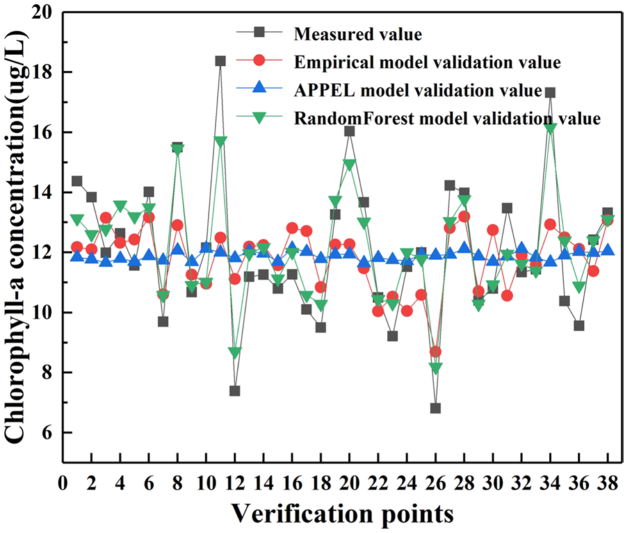


Fig. 14 Line chart of the measured value and verified value of chlorophyll-a concentration based on Sentinel-2 remote sensing data

chlorophyll-a concentration data. It was also compared and analyzed with the empirical model and APPEL model to verify the reliability and efficiency of the random forest model to invert the chlorophyll-a concentration in the study area. Therefore, in the chlorophyll-a concentration inversion study, the random forest inversion model can be used to invert the chlorophyll-a concentration in the study area more effectively and monitor the water quality condition of the area. The R^2 for the accurate evaluation of GF-1 and Sentinel-2 random forest inversion models were 0.931 and 0.875, respectively, while the R^2 for empirical models were 0.565 and 0.194, respectively, and the R^2 for APPEL models were 0.303 and 0.0004, respectively. We also found that for GF-1 and Sentinel-2, our proposed model outperforms other models, and compared with other models, the accuracy improved by over 50%. Therefore, our proposed model (the random forest inversion model) is feasible to predict algal bloom concentrations.

We need to point out that the inversion model used in this paper is limited by the time series of the measured and remote sensing data, and it was not possible to use more models to invert the study area. The inversion of chlorophyll-a concentration is limited by the time of remote sensing images, which is insufficient to construct long time series data for the inversion of chlorophyll-a concentration. We will further study future work in depth in the following areas: (1) exploring the effectiveness of inversion of chlorophyll-a concentration using environmental satellite data, medium and high-resolution data such as Zhuhai-1, and the construction of long-time image sequences and (2) considering the prediction of

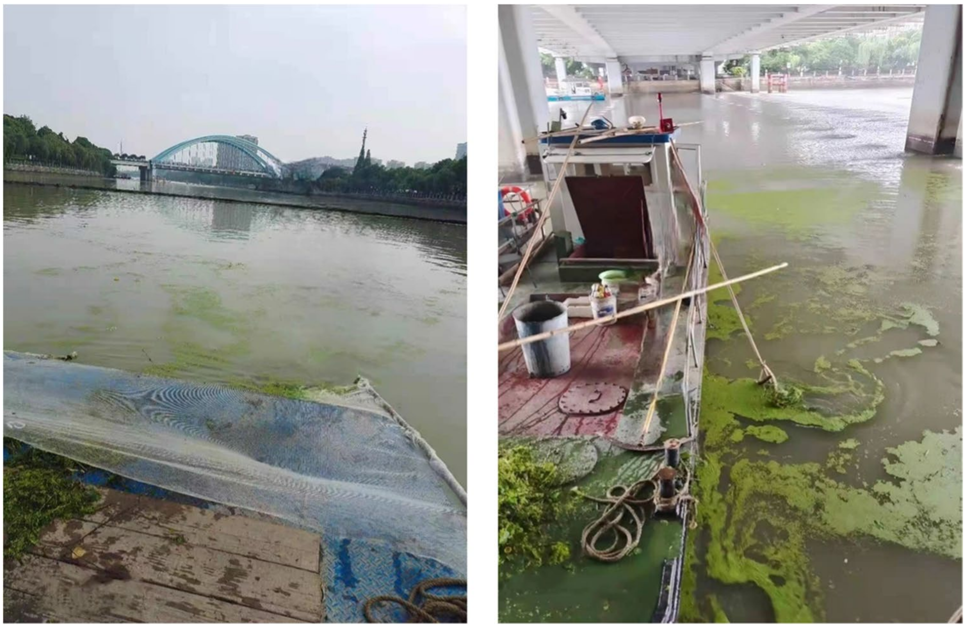


Fig. 15 Photos of salvaging algal bloom

chlorophyll-a concentration based on big data and build more complex models (such as deep learning and other models) to solve the problem of prediction and early warning of chlorophyll-a concentration through long time series of water quality monitoring data.

Author contribution Youliang Chen: conceptualization, methodology. Jinhuang Huang: verification, software, writing—original draft, formal analysis. Youliang Chen and Hamed Karimian: supervision, writing—reviewing and editing. Jinsong Huang: funding acquisition. Zhaoru Wang: provision of study materials.

Funding This work was supported by the National Natural Science Foundation of China (No. 42261072) and the Algal Bloom Remote Dynamic Monitoring Project in Wuxing District, Huzhou City.

Data availability The data will be available upon request.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Consent to participate Not applicable.

Competing interests The authors declare no competing interests.

References

Ali K, Witter D, Ortiz J (2014) Application of empirical and semi-analytical algorithms to MERIS data for estimating chlorophyll a in case 2 waters of Lake Erie. Environ Earth Sci 71:4209–4220

Amorim CA, Dantas ÊW, Moura AdN (2020) Modeling cyanobacterial blooms in tropical reservoirs: the role of physicochemical variables and trophic interactions. Sci Total Environ 744:140659

Anas EA, Karem C, Isabelle L, El-Adlouni SE (2012) Comparative analysis of four models to estimate chlorophyll-a concentration in case-2 waters using MODerate resolution imaging spectroradiometer (MODIS) imagery. Remote Sens 4:2373–2400

Ao Y, Li H, Zhu L, Ali S, Yang Z (2019) The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. J Petrol Sci Eng 174:776–789

Awad M (2014) Sea water chlorophyll-a estimation using hyperspectral images and supervised Artificial Neural Network. Eco Inform 24:60–68

Baladi E, Davar F, Hojjati-Najafabadi A (2022) Synthesis and characterization of g–C3N4–CoFe2O4–ZnO magnetic nanocomposites for enhancing photocatalytic activity with visible light for degradation of penicillin G antibiotic. Environ Res 215:114270

Cai J, Zhang Y, Li Y, Liang XS, Jiang T (2017) Analyzing the characteristics of soil moisture using GLDAS data: a case study in eastern China. Appl Sci 7:566

Chen D, Wang Q, Li Y, Li Y, Zhou H, Fan Y (2020) A general linear free energy relationship for predicting partition coefficients of neutral organic compounds. Chemosphere 247:125869

Chen B, Mu X, Chen P, Wang B, Choi J, Park H, Xu S, Wu Y, Yang H (2021) Machine learning-based inversion of water quality parameters in typical reach of the urban river by UAV multispectral data. Ecol Ind 133:108434

Chen Y, Li D, Karimian H, Wang S, Fang S (2022) The relationship between air quality and MODIS aerosol optical depth in major cities of the Yangtze River Delta. Chemosphere 308:136301

Chen Y, Li H, Karimian H, Li M, Fan Q, Xu Z (2022) Spatio-temporal variation of ozone pollution risk and its influencing factors in China based on Geodetector and Geospatial models. Chemosphere 302:134843

Cho HU, Kim YM, Park JM (2018) Changes in microbial communities during volatile fatty acid production from cyanobacterial biomass harvested from a cyanobacterial bloom in a river. Chemosphere 202:306–311

Dall'Olmo G, Gitelson AA, Rundquist DC, Leavitt B, Barrow T, Holz JC (2005) Assessing the potential of SeaWiFS and MODIS for estimating chlorophyll concentration in turbid productive waters using red and near-infrared bands. Remote Sens Environ 96:176–187

Dev PJ, Sukenik A, Mishra DR, Ostrovsky I (2022) Cyanobacterial pigment concentrations in inland waters: novel semi-analytical algorithms for multi- and hyperspectral remote sensing data. Sci Total Environ 805:150423

Fang S, Li Q, Karimian H, Liu H, Mo Y (2022) DESA: a novel hybrid decomposing-ensemble and spatiotemporal attention model for PM(2.5) forecasting. Environ Sci Pollut Res Int 29:54150–54166

Ge D, Yuan H, Xiao J, Zhu N (2019) Insight into the enhanced sludge dewaterability by tannic acid conditioning and pH regulation. Sci Total Environ 679:298–306

Gower J (1980) Observations of in situ fluorescence of chlorophyll-a in Saanich Inlet. Bound-Layer Meteorol 18:235–245

Guan M, Cheng Y, Li Q, Wang C, Fang X, Yu J (2019) An effective method for submarine buried pipeline detection via multi-sensor data fusion. IEEE Access 7:125300–125309

Guan M, Li Q, Zhu J, Wang C, Zhou L, Huang C, Ding K (2019) A method of establishing an instantaneous water level model for tide correction. Ocean Eng 171:324–331

Han Y, Wang H, Wu J, Hu Y, Wen H, Yang Z, Wu H (2023) Hydrogen peroxide treatment mitigates antibiotic resistance gene and mobile genetic element propagation in mariculture sediment. Environ Pollut 121652

He S, Lu X, Gu J, Tang H, Yu Q, Liu K, Ding H, Chang C, Wang N (2022) RSI-Net: two-stream deep neural network for remote sensing images-based semantic segmentation. IEEE Access 10:34858–34871

Hojjati-Najafabadi A, Rahmanpour MS, Karimi F, Zabihi-Feyzaba H, Malekmohammad S, Agarwal S, Gupta VK, Khalilzadeh MA (2020) Determination of tert-butylhydroquinone using a nanostructured sensor based on CdO/SWCNTs and ionic liquid. Int J Electrochem Sci 15:6969–6980

Hu C, Lee Z, Franz B (2012) Chlorophyll algorithms for oligotrophic oceans: a novel approach based on three-band reflectance difference. J Geophys Res Oceans 117

Jiang Z, Huang X, Wu Q, Li M, Xie Q, Liu Z, Zou X (2023) Adsorption of sulfonamides on polyamide microplastics in an aqueous solution: behavior, structural effects, and its mechanism. Chem Eng J 454:140452

Juarez AB, Barsanti L, Passarelli V, Evangelista V, Vesentini N, Conforti V, Gualtieri P (2008) In vivo microspectroscopy monitoring of chromium effects on the photosynthetic and photoreceptive apparatus of Eudorina unicocca and Chlorella kessleri. J Environ Monit Jem 10:1313–1318

Karimian H, Li Q, Chen HF (2012) Assessing Urban Sustainable Development in Isfahan. Appl Mech Mater 253–255:244–248

Karimian H, Li Q, Li C, Jin L, Fan J, Li Y (2016) An improved method for monitoring fine particulate matter mass concentrations via satellite remote sensing. Aerosol Air Qual Res 16:1081–1092

Karimian H, Karimian H, Chen Y, Tao T, Yaqian L (2020) Spatiotemporal analysis of air quality and its relationship with meteorological factors in the Yangtze River Delta. J Elem 25:1059–1075

Karimian H, Zou W, Chen Y, Xia J, Wang Z (2022) Landscape ecological risk assessment and driving factor analysis in Dongjiang river watershed. Chemosphere 307:135835

Kupssinskü L, Guimares TT, Souza E, Zanotta DC, Mauad FF (2020) A method for chlorophyll-a and Suspended solids prediction through remote sensing and machine learning. Sensors 20:2125

Li S (2022) Efficient algorithms for scheduling equal-length jobs with processing set restrictions on uniform parallel batch machines. Math Bios Eng 19:10731–10740

Li R, Chen N, Zhang X, Zeng L, Wang X, Tang S, Li D, Niyogi D (2020) Quantitative analysis of agricultural drought propagation process in the Yangtze River Basin by using cross wavelet analysis and spatial autocorrelation. Agric for Meteorol 280:107809

Li S, Song K, Wang S, Liu G, Wen Z, Shang Y, Lyu L, Chen F, Xu S, Tao H, Du Y, Fang C, Mu G (2021) Quantification of chlorophyll-a in typical lakes across China using Sentinel-2 MSI imagery with machine learning algorithm. Sci Total Environ 778:146271

Lv Z, Chen D, Feng H, Wei W, Lv H (2022) Artificial intelligence in underwater digital twins sensor networks. ACM Trans Sensor Netw (TOSN) 18:1–27

Maciel DA, Barbosa CCF, Novo EMLdM, Flores Júnior R, Begliomini FN (2021) Water clarity in Brazilian water assessed using Sentinel-2 and machine learning methods. ISPRS J Photogramm Remote Sens 182:134–152

Mamun M, Kwon S, Kim J-E, An K-G (2020) Evaluation of algal chlorophyll and nutrient relations and the N: P ratios along with trophic status and light regime in 60 Korea reservoirs. Sci Total Environ 741:140451

Marie B, Gallet A (2022) Fish metabolome from sub-urban lakes of the Paris area (France) and potential influence of noxious metabolites produced by cyanobacteria. Chemosphere 296:134035

Mo Y, Li Q, Karimian H, Zhang S, Kong X, Fang S, Tang B (2021) Daily spatiotemporal prediction of surface ozone at the national level in China: an improvement of CAMS ozone product. Atmos Pollut Res 12:391–402

Moradi M (2014) Comparison of the efficacy of MODIS and MERIS data for detecting cyanobacterial blooms in the southern Caspian Sea. Mar Pollut Bull 87:311–322

Moradi M (2021) Evaluation of merged multi-sensor ocean-color chlorophyll products in the Northern Persian Gulf. Cont Shelf Res 221:104415

Murugan P, Sivakumarb R, Pandiyanc R (2014) Chlorophyll-A estimation in case-II water bodies using satellite hyperspectral data. Proceedings of the ISPRS TC VIII International Symposium on Operational Remote Sensing Applications: Opportunities, Progress and Challenges, Hyderabad, India, pp 9–12

Odermatt D, Gitelson A, Brando VE, Schaepman M (2012) Review of constituent retrieval in optically deep and complex waters from satellite imagery. Remote Sens Environ 118:116–126

Oyama Y, Matsushita B, Fukushima T (2015) Distinguishing surface cyanobacterial blooms and aquatic macrophytes using Landsat/TM and ETM+ shortwave infrared bands. Remote Sens Environ 157:35–47

Qin B, Li W, Zhu G, Zhang Y, Wu T, Gao G (2015) Cyanobacterial bloom management through integrated monitoring and forecasting in large shallow eutrophic Lake Taihu (China). J Hazard Mater 287:356–363

Qiu Z, Jiao M, Jiang T, Zhou L (2020) Dam structure deformation monitoring by GB-InSAR approach. IEEE Access 8:123287–123296

Rotta L, Alcântara E, Park E, Bernardo N, Watanabe F (2021) A single semi-analytical algorithm to retrieve chlorophyll-a concentration in oligo-to-hypereutrophic waters of a tropical reservoir cascade. Ecol Ind 120:106913

Sha J, Xiong H, Li C, Lu Z, Zhang J, Zhong H, Zhang W, Yan B (2021) Harmful algal blooms and their eco-environmental indication. Chemosphere 274:129912

Song D-X, Wang Z, He T, Wang H, Liang S (2022) Estimation and validation of 30 m fractional vegetation cover over China through integrated use of Landsat 8 and Gaofen 2 data. Sci Remote Sens 6:100058

Su H, Li W, Yan XH (2018) Retrieving temperature anomaly in the global subsurface and deeper ocean from satellite observations. J Geophys Res Oceans 123:399–410

Sun P, Zhang K, Wu S, Wang R, Wan M (2021) An investigation into real-time GPS/GLONASS single-frequency precise point positioning and its atmospheric mitigation strategies. Meas Sci Technol 32:115018

Visser PM, Ibelings BW, Bormans M, Huisman J (2016) Artificial mixing to control cyanobacterial blooms: a review. Aquat Ecol 50:423–441

Wan W, Zhang Y, Cheng G, Li X, Qin Y, He D (2020) Dredging mitigates cyanobacterial bloom in eutrophic Lake Nanhu: shifts in associations between the bacterioplankton community and sediment biogeochemistry. Environ Res 188:109799

Wang X, Wang T, Xu J, Shen Z, Yang Y, Chen A, Wang S, Liang E, Piao S (2022) Enhanced habitat loss of the Himalayan endemic flora driven by warming-forced upslope tree expansion. Nat Ecol Evol 6:890–899

Wu C, Li Q, Hou J, Karimian H, Chen G (2018) PM2. 5 concentration prediction using convolutional neural networks. Sci Surv Mapp 43:68–75

Xia C, Joo S-W, Hojjati-Najafabadi A, Xie H, Wu Y, Mashifana T, Vasseghian Y (2023) Latest advances in layered covalent organic frameworks for water and wastewater treatment. Chemosphere 138580

Xu D, Zhu D, Deng Y, Sun Q, Ma J, Liu F (2023) Evaluation and empirical study of Happy River on the basis of AHP: a case study of Shaoxing City (Zhejiang, China). Mar Freshw Res. https://doi.org/10.1071/MF22196

Yang J, Li P, He Y (2014) A multi-band approach to unsupervised scale parameter selection for multi-scale image segmentation. ISPRS J Photogramm Remote Sens 94:13–24

Zhang X, Xiao P, Song X, She J (2013) Boundary-constrained multiscale segmentation method for remote sensing images. ISPRS J Photogramm Remote Sens 78:15–25

Zhang J, Fu P, Meng F, Yang X, Xu J, Cui Y (2022) Estimation algorithm for chlorophyll-a concentrations in water from hyperspectral images based on feature derivation and ensemble learning. Eco Inform 71:101783

Zhou J, Qin B, Zhu G, Zhang Y, Gao G (2020) Long-term variation of zooplankton communities in a large, heterogenous lake: implications for future environmental change scenarios. Environ Res 187:109704

Zhou W, Yang H, Xie L, Li H, Huang L, Zhao Y, Yue T (2021) Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model. CATENA 202:105222

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.