Machine learning method for quick identification of water quality index (WQI) based on Sentinel-2 MSI data: Ebinur Lake case study

Xiaohang Li, Jianli Ding and Nurmemet Ilyas

ABSTRACT

Surface water quality is an important factor affecting the ecological environment and human living environment. The monitoring of surface water quality by remote sensing monitoring technology can provide important research significance for water resources protection and water quality evaluation. Finding the optimal spectral index sensitive to water quality for remote sensing monitoring of water quality is extremely important for surface water quality analysis and treatment in the Ebinur Lake Basin in arid areas. This study used Sentinel-2MSI data at 10 m resolution to quickly monitor the water quality of the watershed. Through laboratory experiments and measurement data from the Ebinur Lake Basin, 22 water quality parameters (WQPs) were obtained. Through Z-score and redundancy analysis, 9 WQPs with significant contributions were extracted. Based on the remote sensing spectral band, four water indexes (NDWI, NWI, EWI, AWEI-nsh) and 2D modeling spectral index (DI, RI, NDI), the correlation analysis between WQPs and two kinds of spectral band indexes is carried out, and it is concluded that the overall correlation between WQP and 2D spectral modeling is more relevant. This paper calculates the evaluation and models the 2D spectrum of the Water Quality Index (WQI). The WQI is predicted and modeled through four machine learning algorithms (RF, SVM, PLSR, PLSR-SVM).The results show that the inversion effect of the two-dimensional spectral modeling index on water quality parameters (WQPs) is superior to that of the water index, and the correlation coefficient of the DI (R12-R1) SWIR-2 and BLUE band interpolation index reaches 0.787. On this basis, three kinds of two-dimensional spectral modeling indexes are used to inversely synthesize the WQI, and the correlation coefficient of the ratio index of the RI (R11/R8) SWIR-1 and near-infrared (NIR) bands is preferably 0.69. In the WQI prediction, the partial least squares regression support vector machine (PLSR-SVM) model in machine learning algorithms has good modeling and prediction effects ($R^2c = 0.873$, $R^2v = 0.87$), which can provide a good basis. The research results provide references for remote monitoring of surface water in arid areas, and provide a basis for water quality prediction and safety evaluation.

Key words | machine learning, remote sensing reflectance, Sentinel-2 MSI, water quality index (WQI), water quality parameter (WQP)

Xiaohang Li
Jianli Ding (corresponding author)
Nurmemet Ilyas
Key Laboratory of Smart City and Environment Modelling of Higher Education Institute, College of Resources and Environment Sciences,
Xinjiang University,
Urumqi 830046,
China
E-mail: *watarid@xju.edu.cn*

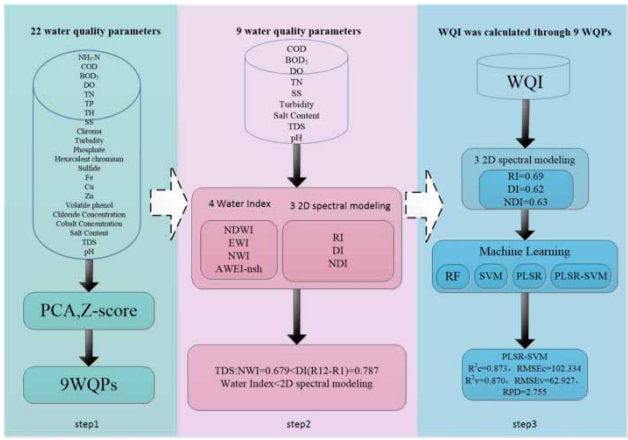
HIGHLIGHTS

- Inversion of water quality in the Ebinur Lake Basin by Sentinel-2 MSI data.
- Analysis of the contribution rate of different water quality parameters in water by Z-score and PCA.

doi: 10.2166/ws.2020.381

- Comparison of classical water quality index, correlation between spectral modeling and water quality parameters.
- Modeling and predicting water quality index (WQI) using machine learning and linear correlation methods.

GRAPHICAL ABSTRACT



INTRODUCTION

With the continuous development of water resources, water quality has become a popular issue of global concern (Pahlevan *et al.* 2019). The area of human activity continues to expand worldwide, and the impact on water quality is gradually becoming more severe (Mananze *et al.* 2018). The deterioration of water quality due to pollution is increasing due to urban cluster pollution emissions, industrial pollution emissions, and agricultural fertilizer pollution emissions (Shao *et al.* 2006; Lintern *et al.* 2018; Wang *et al.* 2019). Water pollution has had a considerable impact on human production and life and increased the health risks of ecosystems (Chen *et al.* 2020). Therefore, water quality and safety assessments are extremely important. Water quality has become the most important indicator in aquatic ecosystems. The quality of water directly affects the health of aquatic ecosystems and fully reflects the spatiotemporal heterogeneity and integrity of river water environment ecosystems (Staponites *et al.* 2019). Therefore, it is necessary to monitor the pollution of water quality in real time and record changes in water quality.

In the current situation, remote sensing can be used to accurately and quickly monitor river water quality, and there is an opportunity to regularly monitor aquatic systems using different sensor satellite products (Tyler *et al.* 2016). Remote sensing technology has become an important method in water resource surveys and monitoring, the remote monitoring of water environments, regional ecological remote sensing monitoring, urban environmental remote sensing monitoring and wetland protection. Reducing the interference of non-water factors and enhancing the expression of water body factors are key tasks in remote sensing to identify water body information (Li *et al.* 2016). Landsat image recognition technology for water bodies has been well established around the world. With the continuous deepening of machine learning methods, more and more water quality researches use this method to predict water quality (Gao *et al.* 2019; Hussein *et al.* 2019; Liang *et al.* 2020). This technology combines artificial neural network (ANN), support vector machine (SVM), and maximum likelihood (ML) classification methods to reflect

water changes (Rokni *et al.* 2015). In multivariate data research, the partial least squares regression (PLSR) model can show better results in modeling and forecasting a large number of water quality indexes (Carrascal *et al.* 2010). And PLSR is widely used in the establishment of multivariate models, which can handle the collinearity between various variables and further reduce and weaken related data sets (Sidike *et al.* 2014; Wang *et al.* 2020). Random forest (RF) has a good application basis for water quality index prediction (Meyers *et al.* 2017), support vector machine (SVM) can achieve a better effect on nonlinear relationships when used alone (Nawar *et al.* 2016; Lucà *et al.* 2017). In PLSR-SVM, not only high-quality results can be obtained, but also spectral bands can be combined to improve the inversion effect (Zhang *et al.* 2021).

Additionally, this technology applies a cluster analysis method to establish a stable water index, extract surface water information (Wang *et al.* 2018b) and establish an automatic classification set (Fisher *et al.* 2016). The remote sensing of water quality mainly includes water color remote sensing and water quality remote sensing. Water color remote sensing mainly uses remote sensing to invert the optical characteristics and spectral characteristics of water bodies, such as the suspended matter and chlorophyll a levels. Conversely, if there are no significant optical and spectral properties for a given water body, the remote sensing inversion of other WQPs that have close relationships with direct WQPs can be achieved by water quality remote sensing; such parameters include total nitrogen and total phosphorus (Ma & Dai 2005). In research study, based on the change in chromaticity and the inversion of water body radiation and WQPs through spectroscopy, chromaticity is used to predict water quality. The spectral complexity of inland lakes has also gradually been considered (Bukata *et al.* 1983).

In 1991, Anatoly Gitelson (1991) studied the remote sensing of inland water quality through aerospace remote sensing and studied the potential associated with using chlorophyll a (Lintern *et al.* 2018), dissolved organic matter (DOM), and suspended solids (SM) concentrations. BABAN and Serwan (Baban 1993) used Landsat satellites to perform a regression analysis of TM data and surface WQPs in 1993 and to simulate and predict the water quality of inland lakes. Since 1990, Chinese scholars have studied the water quality of inland rivers through remote sensing (Yu *et al.* 2008). The research on chlorophyll and suspended solids has been continuously improved, and the inversion accuracy also been improved (Ma & Dai 2005). In the inversion of the water quality of inland rivers, by selecting different spectral band combinations and combining the water quality spectral characteristic parameters, the accuracy of establishing estimation spectral and water quality estimation models has been improved. With the continuous advancement of remote sensing technology, new methods of establishing the correlations with water quality through different wave bands of remote sensing satellites have been developed.

In recent years, Landsat 8 and Sentinel-2 have been widely used for remote sensing and the monitoring of water resources. Remote sensing methods have been used to extract surface water information from images, to analyze the sensitivity of different spectral bands to the surface water quality and to distinguish surface water quality levels. Rapid extraction requires technical support (Zheng *et al.* 2015). Stéfani (Stéfani *et al.* 2017) studied the effects of total suspended solids (TSS) and chlorophyll (Lintern *et al.* 2018) levels on the physiological response of oysters, highlighting the difference between the near-infrared (NIR) bands of Landsat imagery and Sentinel-2A imagery, which plays a significant role in band quantifying TSS. Fernanda (Fernanda *et al.* 2017) found that the NIR band of Sentinel-2A is more sensitive to Chl in Brazil's super-eutrophic reservoirs. Yang (Yang *et al.* 2018) found that Sentinel-2 multispectral satellite imagery can effectively reduce noise and be used to extract water more accurately than other products in a study of the automatic extraction of urban surface water. Landsat operational land imager (OLI) data are different from data corresponding to the four Vegetation red edge (VRE) bands of Sentinel-2 multispectral imagery (MSI), with better sensitivity to Chl. Additionally, the NIR bands of alternative algorithms and concentrations of Chl pigments have been effectively obtained (Moses *et al.* 2009). As a result, MSI data can be used to effectively monitor the growth and reproduction of various algae in water (Gower *et al.* 2008). Combining NIR (0.842 μm) and two shortwave infrared (SWIR) systems based on visible light (Halgamuge & Davis 2019) can enhance water inversion and the efficiency of vegetation detection. The WQI is a water index that covers all water resource parameters (Terrado *et al.* 2010). The WQI appears to lag behind the spatial and

temporal changes in water quality samples when evaluating individual water samples, and problems related to the continuous large-scale comprehensive evaluation of water quality changes must be solved with the development of remote sensing technology (Wang *et al.* 2017). In this paper, WQPs are extracted to establish the corresponding WQI; then, the effects of the most prominent parameters of river basin water on river water quality are investigated.

The river water ecosystem in the Ebinur Lake Basin, Xinjiang, was studied. The Ebinur Lake Basin is located in an arid zone and adjacent to the largest arid zone in the temperate region of the Northern Hemisphere. The basin has typical regional characteristics, large climate changes and considerable evapotranspiration, and the artificial large-scale extraction of water for agricultural production has led to regional ecological and hydrological changes. Intensive non-point source pollution caused by the development of water and soil resources has affected the water quality of the basin to a certain extent. To this end, it is necessary to quickly identify and manage water quality through remote sensing images and provide technical support for environmental protection and safety.

In this article, the water quality in arid regions is studied through an analysis of the correlation between the water index and the WQI based on a spectral combination method, and a correlation analysis between the band combination index and the WQI is performed to identify the optimal WQI. The spectral band combination that is most suitable for water quality in arid areas is also determined. Furthermore, the effects of different parameters on the water quality in the entire basin is studied to provide important technical support for the future macrocontrol of water bodies.

MATERIALS AND METHODS

Study area

The study was performed in the arid region of the hinterland of the Eurasian continent, the Ebinur Lake Watershed of northwestern Xinjiang, China (Figure 1). The Ebinur Lake Watershed is relatively complex. This watershed is located in the northwestern part of the Junggar Basin and is the lowest lake subbasin in the Junggar Basin. The northeastern part of the area is connected to the Gurbantunggut Desert, and the northwestern and western parts are the Alatau Mountains and low hills, respectively (Abuduwaili *et al.* 2008). Additionally, the Alashankou region, which is the largest outlet in China, is part of this basin. The valley is surrounded by mountains to the north, west and south. The plain is approximately 220 km from east to west and 60–120 km from north to south. The total area of the basin is 2,080 km^2. Ebinur Lake is located in the low-lying area of the river tail basin (Yu & Jiang 2003), and the source of recharge for the lake is the surrounding surface water and groundwater. The water of Ebinur Lake is mainly composed of water from several inflowing rivers, such as the Jing River, Boertala River and Kuitun River. The complexity of the geographical environment has affected the ecological environment of the Ebinur Lake Basin to varying degrees. The water quality of river water directly affects the growth of vegetation in the basin, the degree of salinization of the soil and the living conditions for human beings (Wang *et al.* 2017).

Sentinel-2 image acquisition and preprocessing

A multispectral instrument was launched onboard the Sentinel-2A satellite on June 23, 2015, from the European Space Agency. Sentinel-2A has 13 bands and three different spatial resolutions. The MSI ranges from VIS to NIR and SWIR, with spatial resolutions of 10, 20, and 60 m, respectively (Table 1). Notably, the MSI is the only terrestrial remote sensing satellite in the world with a ground resolution of up to 20 m. Four specialized bands (B5, B6, B7, and B8a) were designed to obtain the spectral characteristics of the vegetation in the near-infrared 'red edge' region (690–800 nm), and these bands are close to B4 at the red wavelength (Peterson *et al.* 2020).

We downloaded the Sentinel 2 image from the ESA Sentinel Scientific Data Hub (https://scihub.copernicus.eu/); this data is a radiometrically calibrated L1C product. We considered the data of field survey to choose the date of the image (October 2017, cloudless). To achieve the L2A product, we conducted atmospheric correction using the sen2cor toolbox in SNAP software (Main-Knorn *et al.* 2015). Meanwhile, geometric correction is carried out to ensure the verification accuracy to ±0.5 pixels. Using ENVI5.5 software, the 20-m and 60-m resolution bands were resampled to 10 m. To verify the water spectrum, the spectrum of the

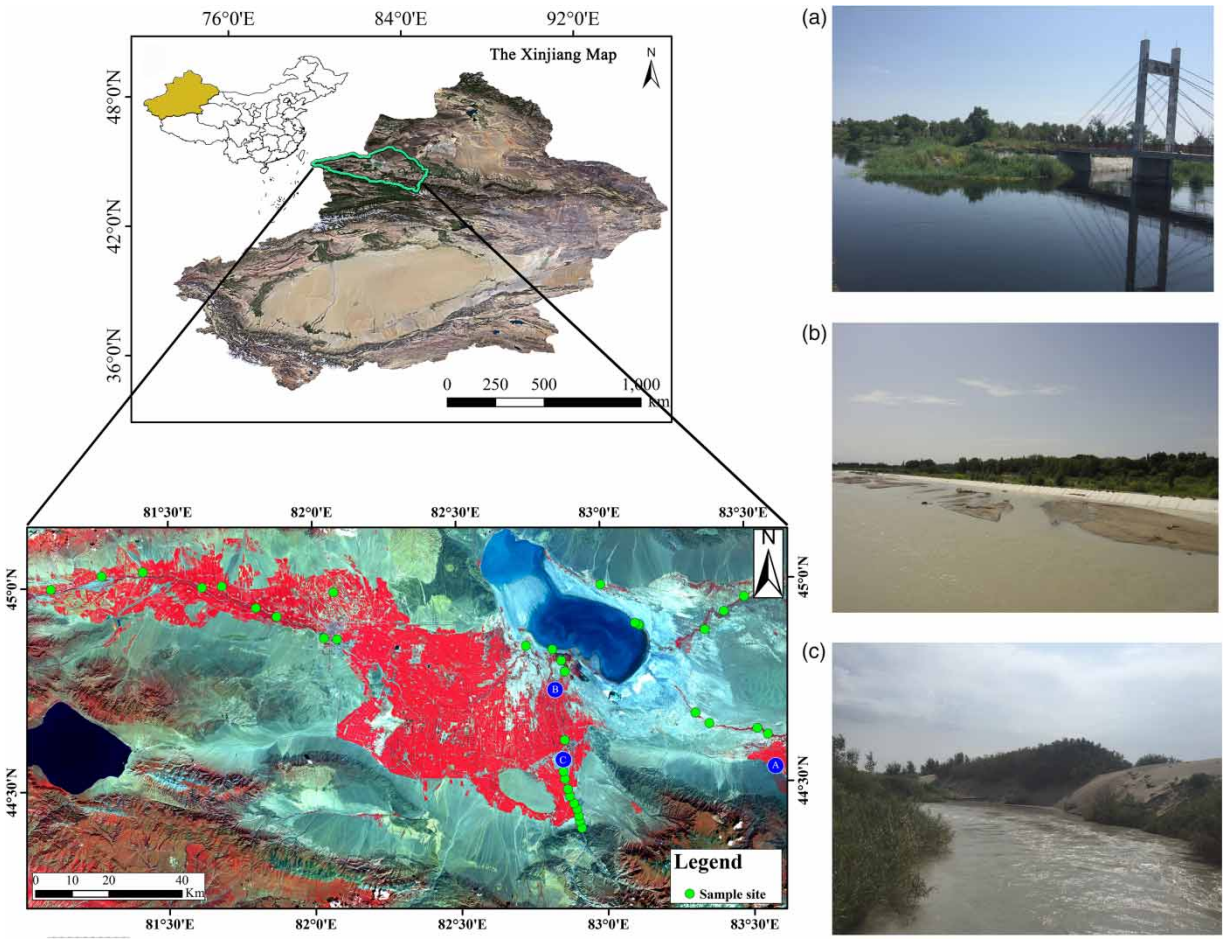


Figure 1 | The sampling points in the study area were on May 28, 2016, in the Ebinur Lake Basin. The Sentinel-2 data were selected. The image uses pseudo-RGB colors, and the bands are 8, 4, and 3. The three pictures in (a), (b), and (c) are landscape photos of the three sampling points (a), (b), and (c) in the research area, respectively.

typical sample was extracted as Figure 2. In general, the water spectrum is obvious reflection in the B3 band, and the absorption continues to increase after the B4 band. The absorption effect is best in the B8a band, which is consistent with the changes in the water in the spectral range. By comparison, these spectra were accurate in this study.

Water sampling collection and analysis

River water quality data selection is mainly based on China's surface water environmental quality standard GB3838-2002 (GB3838-2002 2002); in this process, water samples were collected from the basin, and outdoor and indoor tests were conducted.

The sampling sites were located on the Boertala River in the west, the Jing River in the south and the Kuitun River in the east. Samples were collected from 38 sample points according to the trends of the three rivers, and samples from surface water bodies with a depth of 50 cm on the surface section of the river were extracted. One-liter water samples were collected, which were stored in polytetrafluoroethylene plastic bottles, and packed into benzene board insulation boxes with ice cubes and quickly transported back to the laboratory for the determination of the water quality indices. Water quality indicators were experimentally tested at each water sample point, and the measurement range of each indicator was strictly screened according to China's surface water environmental quality standard GB3838-2002 (GB3838-2002 2002), and passed based on the 'Water and Wastewater Monitoring and Analysis Method' of the State Environmental Protection Administration (State Environmental Protection Administration of

Table 1 | Description of the information regarding the Sentinel-2 image

Band	Band name	Center wavelength (μm)	Spatial resolution (m)
1	Coastal Aerosol	0.443	60
2	Blue	0.490	10
3	Green	0.560	10
4	Red	0.665	10
5	Vegetation Red Edge	0.705	20
6	Vegetation Red Edge	0.740	20
7	Vegetation Red Edge	0.783	20
8	NIR	0.842	10
8a	Vegetation Red Edge	0.865	20
9	Water Vapor	0.945	60
10	SWIR-Cirrus	1.375	60
11	SWIR-1	1.610	20
12	SWIR-2	2.190	20

*The data come from the European Space Agency (ESA), which only released L1C-level multispectral data (MSI) of Sentinel-2 (S2) https://earthexplorer.usgs.gov/.

China 2002). Twenty-two kinds of water quality indicators were experimentally monitored (Table 2).

The following 22 parameters were analyzed: ammonia nitrogen (NH_3-N), chemical oxygen demand (COD), biological oxygen demand (BOD_5), dissolved oxygen (DO), total nitrogen (TN), total phosphorus (TP), total hardness (TH), total suspended solid(SS), Chroma, turbidity, phosphate, hexavalent chromium, sulfide, Fe, Cu, Zn, volatile phenol, the chloride concentration, the cobalt concentration, the salt content (CON), total dissolved solids (TDS), and the pH. First, the data or named z-scores were standardized. Additionally, the concentration of the data through this step was analyzed, and the parameters were further extracted.

Then, a redundancy analysis (Kaplan & Avdan 2017) was performed on the 22 WQPs. The RDA involved a sorting method that combined regression analysis with principal component analysis (PCA) and an extension of multi-response regression analysis. Conceptually, the RDA is a PCA of the fitted-value matrix of the multivariate multiple linear regression between the response variable matrix and the explanatory variables. The parameters in the RDA diagram are determined by the length and direction of the arrows, which represent the contribution of each variable to the two main components of the biplot.

Water spectral indices

There are many established water indices that can be applied to Landsat TM/ETM + data. This study selected four classic water indices for analysis. Based on the Landsat WQI, the improved WQI was developed to reflect the inversion of water quality in each band of Sentinel-2. The normalized difference water index (NDWI) was selected (McFeeters 1996; Yin *et al.* 2018; Alghamdi *et al.* 2020) by choosing the two optimal bands of water information from the remote sensing data. The optimal index was obtained by the spectral calculation of the green band and NIR of the VIS. A new water index was established; the automated water extraction index with no shadows (AWEI-nsh) can

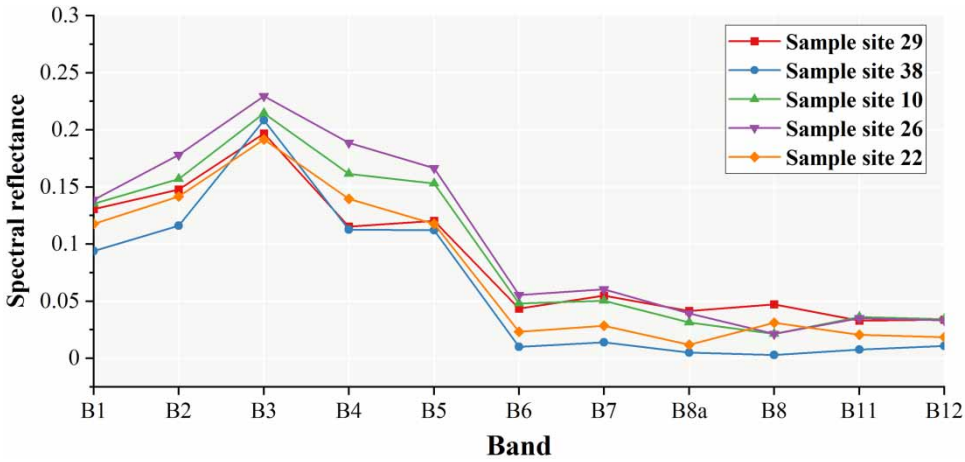


Figure 2 | The spectral reflectance of study area in Sentinel-2 MSI imagery.

Table 2 | Water index methods using Sentinel-2 MSI data

Parameters	Units	Measurement methods
NH_3-N	$mg{\cdot}L^{-1}$	Measured by spectrophotomtetry with salicylic acid
COD	$mg{\cdot}L^{-1}$	Determined according to the dichromate method (HJ 828-2017) with a standard COD digester (KY-100)
BOD_5	$mg{\cdot}L^{-1}$	Measured by the dilution and inoculation method (HJ 505-2009) with a thermostat (HWS-150)
DO	$mg{\cdot}L^{-1}$	Measured by visible light photometer (722N) according to iodometric method (GB 7489-1987)
TN	$mg{\cdot}L^{-1}$	Measured by ultraviolet spectrophotometry (HJ 535-2009) with a UV-visible photometer (UV-6100)
TP	$mg{\cdot}L^{-1}$	Measured by the ammonium molybdate spectrophotometry (HJ 636-2012) with visible light photometer (722N)
TH	$mg{\cdot}L^{-1}$	EDTA titration method
SS	$mg{\cdot}L^{-1}$	Gravimetric method
Chroma		Determined by platinum-cobalt colorimetry (GB 11903-89)
Turbidity	NTU	Measured by spectrophotometry (GB 13200-91) with a visible light photometer (722N)
Phosphate	$mg{\cdot}L^{-1}$	Measured by ion chromatography
Hexavalent chromium	$mg{\cdot}L^{-1}$	Measured by diphenylcarbazide spectrophotometry
Sulfide	$mg{\cdot}L^{-1}$	Measured by the iodine content method
Fe	$mg{\cdot}L^{-1}$	Measured by phenanthroline spectrophotometry and atomic absorption spectrophotometry
Cu	$mg{\cdot}L^{-1}$	Measured by spectrophotometry
Zn	$mg{\cdot}L^{-1}$	Measured by spectrophotometric method with dithizone
Volatile phenol	$mg{\cdot}L^{-1}$	Measured by flow inject analysis (FIA) and 4-APP spectrophotometric method
Chloride concentration	$mg{\cdot}L^{-1}$	Measured by silver nitrate titration method
Cobalt concentration	$mg{\cdot}L^{-1}$	Measured by flame atomic absorption spectrometry
Salt content	μS·cm-1	Measured with a hand-held conductivity meter
TDS	$mg{\cdot}L^{-1}$	Determined by weight method (GB 11901-89) with a ten-thousandth-scale balance (SI-234)
pH		Measured with a pH-40A portable pH acidity meter

effectively remove the non-water pixel index information and is suitable for cases in which shadows are included in the image. Luo Chongliang (Luo 2015) extracted the Ebinur Lake boundary using the enhanced water index and normalized water index (NWI); these researchers found that the two types of indices were suitable for water extraction in the Ebinur Lake Basin and that the extraction accuracy was sufficient (Table 3).

Water quality index (WQI)

As there are a variety of chemical, physical and biological WQPs, several researchers have proposed a WQI in the form of a simple expression for representing the general quality of surface waters (Zeinalzadeh & Rezaei 2017). This study constructed a WQI as a weighted multifactor environmental quality index that highlights the maximum values of major water quality indicators. The WQI can be used to extract and reflect WQPs and the water quality and composition (Wang *et al.* 2017).

First, a certain weight value is assigned to each parameter. According to the influence of different WQPs on the human body, as established by the World Health Organization (WHO 2008), the weights of different parameters are assigned values from 1–5, and the weight assignment calculation formula is as follows:

$$\mathrm{Wi} = \mathrm{w_i} / \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{W_i}$$

where Wi is the weight value, wi is the weight of each parameter, and n is the number of WQPs. The weight value of each water quality parameter is based on the surface water quality standards of the WHO (Alghamdi *et al.*

Table 3 | Water index methods using Sentinel-2 MSI data

Index	Source	Equation (Sentinel-2)
NDWI	McFeeters (1996)	$NDWI = \frac{B_3 - B_8}{B_3 + B_8}$
EWI	Luo (2015)	$EWI = \frac{B_3 - B_8 - B_{11}}{B_3 + B_8 + B_{11}}$
NWI	Luo (2015)	$NWI = \frac{B_1 - (B_8 + B_{11} + B_{12})}{B_1 + B_8 + B_{11} + B_{12}}$
AWEI-nsh	Feyisa *et al.* (2014)	$AWEI_{\_nsh} = 4 \times (\rho B_3 - \rho B_{11}) - (2.5 \times \rho B_8 + 2.75 \times \rho B_{12})$

2020). The result is multiplied by 100%. The specific formula is as follows:

$$q_i = \frac{C_i}{S_i} \times 100$$

where q_i is the ratio of the measured WQP concentration to the International WHO and China surface water quality standard concentration value, C_i is the water quality parameter concentration of the measured sample, and S_i is the International WHO and China surface water quality standard concentration value. This value mainly refers to the 2008 WHO (WHO 2008) and China surface water quality standard concentration value. To calculate the WQI value in this study, we first need to calculate the contributions of the individual WQPs to the WQI (Table 4). These contributions can be calculated by the following formula:

$$SI_i = W_i \times q_i$$

$$WQI = \sum_{i=1}^{n} SI_i$$

where SI_i is the contribution of water quality parameter i to the WQI and W_i is the weight value of the water quality parameter. The weight value is assigned by referencing a Intergovernmental Panel on Climate Change (IPCC) report, and the values of WQPs assigned by the WHO considering human health are considered. The values of parameters vary from 1–5; the greater the threat of WQPs to human health, the higher the W_i value (Ramakrishnaiah *et al.* 2009).

Table 4 | Relative weight for each indicator

Parameters	WHO standards (2011)	Weight (wi)	Relative weight (Wi)
COD	15(mg/L)	4.0	0.154
BOD_5	3(mg/L)	5.0	0.192
DO	6 (mg/L)	1.0	0.038
TN	0.5(mg/L)	3.0	0.115
SS	7(mg/L)	3.0	0.115
Turbidity	30(NTUc)	3.0	0.115
Salt content	10(mg/L)	2.0	0.077
TDS	450(mg/L)	1.0	0.038
pH	6.8–8.5	4.0	0.154
		$\sum wi = 26.0$	$\sum wi = 1.000$

The WQI is a quality index that directly and comprehensively reflects water quality (Ewaid *et al.* 2018). This index can quantify the degree of river water pollution objectively, scientifically and reasonably, and it can comprehensively and effectively reflect the effectiveness of comprehensive improvements in rivers. Compared with a single water quality assessment, WQI is a comprehensive water quality assessment method that integrates multiple water quality variables into an intuitive value (Zotou *et al.* 2020). This method can not only evaluate the overall situation of water quality (Şener *et al.* 2017), but also eliminate the differences caused by a single water body. Effective selection of the relevant water quality index can better apply WQI to water quality assessment at all levels (Pesce & Wunderlin 2000), and construct a WQI based on the actual water resources in the study area. This article contains many

variables of water quality parameters, and these parameters are not all influential in terms of overall water quality. It is a relatively better way to invert the overall water quality by selecting water quality parameters with greater weight in the PCA analysis. This is a good, comprehensive water quality evaluation method (Şener *et al.* 2017).

Constructions of 2D spectral indices

Information enhancement processing is a common data processing method for remote sensing data. Complex terrain information is difficult to analyze using single-band data. Therefore, multispectral remote sensing data are selected for two-dimensional spectrum analysis and calculations (addition, subtraction, multiplication, division, etc.). In the nonlinear combination method, the spectral information of the remote sensing data is enhanced, the values sensitive to the unique feature information-spectral index are extracted, and the feature information is qualitatively and quantitatively evaluated (Wang *et al.* 2018a). Correlation analysis is performed between the calculated multispectral index and the water quality content, such that the determination coefficient R^2 can be extracted to combine any two bands of the two-dimensional spectral index between 443 nm and 2,190 nm; consequently, the results can be combined to analyze nine different water quality and band combinations.

$$NDI(R_i, R_j) = (R_i - R_j)/(R_i + R_j) \quad (1)$$

$$DI(R_i, R_j) = R_i - R_j \quad (2)$$

$$RI(R_i, R_j) = R_i/R_j \quad (3)$$

where NDI (Equation (1)) is the normalized remote sensing index, DI (Equation (2)) is the difference remote sensing index, RI (Equation (3)) is the ratio remote sensing index, and i,j is an arbitrary band involved in any two-band data extraction process among the 1–7, 8a, 8, 9, 11, and 12 bands.

Research calibration models

In recent years, there have been more and more researches on the use of machine learning to model data, and modeling methods that can provide choices are also endless, but the three modeling methods (SVM, RF, PLSR) selected in this article are relatively mature machine learning methods, so this article chooses these three methods to model and predict the data, which can effectively improve the accuracy of data prediction.

Support vector machines (SVM)

An SVM is a supervised classifier based on small samples in machine learning that can quickly and accurately fit and predict samples; this method was proposed by Boser (Boser 2008). The SVM finds the difference between the maximal margins by points in different categories in the sample space; the results are mapped to the same sample space for classification and prediction of different categories (Zhao *et al.* 2018). This method has mainly been applied to classification and regression problems, where it is better applied. The regression model of the SVM is divided into a linear model and a nonlinear model (Kaya 2013). It is not invertible in the linear model, and the non-linear model can add the kernel function to the linear model (Nie *et al.* 2020). The method essentially finds a plane in the data such that all of the data in the set conform to the nearest plane distance.

Random forest (RF) regression

The random forest (RF) is an algorithm based on classification and regression through a large number of decision trees as weak classifiers. The generated classifiers are diverse and aggregated (Halgamuge & Davis 2019). The RF selects the bagging method and randomly returns corresponding decision trees while continuing to select different training subsets, and then it votes on these decision tree classification results to determine the final classification result (Houborg & McCabe 2018). However, the RF and the traditional decision tree algorithms employ different training processes. The traditional decision tree selects the optimal attribute each time it selects the partition attribute, whereas the RF introduces random attribute selection when selecting the partition attribute of the node. The principle is similar to the tournament selection method in genetic algorithms (Peters *et al.* 2007). To achieve two

random characteristics of the RF, classification features are randomly selected and the training subsets are randomly generated and are independent and identically distributed. With these two characteristics, various data sets can be quickly extracted without repeated accumulation, which is better for data classification.

Partial least squares regression (PLSR)

PLSR is a new multivariate statistical analysis method (Geladi & Kowalski 1986). It is a regression-based modeling method based on a multivariate dependent variable corresponding to multivariate analysis. This linear model is better than the one-by-dependent variable for regression analysis based on the internal linear height correlation. However, PLSR can cope with the problem of a small sample size. This method combines the advantages of the canonical correlation analysis, PCA and multiple linear regression analysis to improve prediction and nonmodeling prediction types as well as data cognitive analysis. The optimal number of potential variables to be predicted in the future can be determined based on the number of factors with the smallest RMSE (Douglas *et al.* 2018).

RESULTS

Water quality parameter selection

Z-score

In this study, 22 WQPs were chosen, normalized and introduced to a system with a standard value of 1 and an average of 0 for the analysis. To perform an accurate and economically feasible analysis of all the parameters, it is necessary to screen all the parameters and extract the indicators with the highest contributions.

For the 22 WQPs, the Z-score analysis is shown in Figure 3. For this box and whisker plot, the Y-axis represents the 22 WQPs, and the X-axis represents the WQPs based on normalized values. The green line in the middle of the box is the median of all data; for all the data, the median value is on the left side of 0. Additionally, the line is on the right side of 0 for DO and the pH. The left side of the blue border represents 25% of the data, and the right side represents 75% of the data. The overall quartiles of the data are small, and the middle part of the data is the most concentrated. The blue box enclosed by the black whiskers

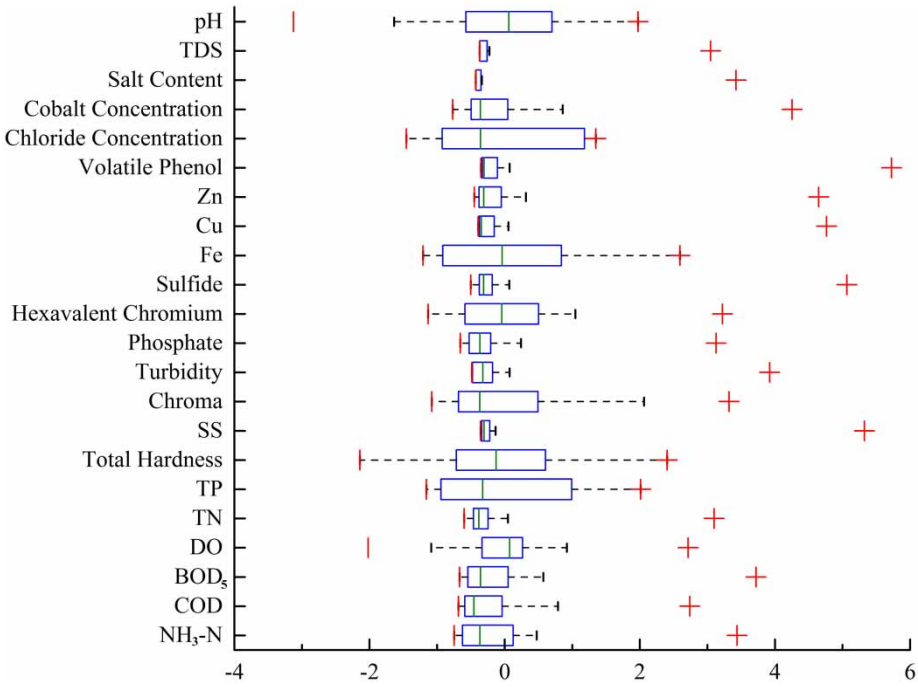


Figure 3 | Z-score variability of the data.

has the symbol '|', which denotes the maximum and minimum values of the data. The red vertical lines on the left and plus signs on the right are the extensions of the whiskers to the most extreme data points that are not considered outliers.

The figure shows that most of the small and concentrated data have large abnormal values, indicating that the water quality at the sampling points in different rivers is very different and the values at some sampling points may exceed the standard; that is, all of the data sets have outliers. The parameters with more than four outliers include the TSS, turbidity, sulfur, Cu, Zn, volatile phenol, and cobalt concentration, and the largest outliers appear for these indicators. The water quality data are standardized, and it is clear that the median value of the data appears below the average in these cases.

RDA analysis and correlation

In the first principal component of the horizontal axis (Figure 4), the positive correction coefficients of phosphate, Zn, TN, COD, BOD_5, the cobalt concentration, the salt content, hexavalent chromium, the chloride concentration, TDS, chroma, and NH_3-N are shown. Additionally, suspended solids, Fe, TP, turbidity, Cu, and sulfur volatile phenols are plotted. Moreover, on the vertical axis of the second principal component, the negative correction coefficients are DO and the pH. These points are scaled with respect to the maximum score value and the maximum coefficient length; therefore, only the relative locations can be determined from the plot.

Pearson's linear correlation is a statistic that reflects the linear correlation between two variables by the one-to-one

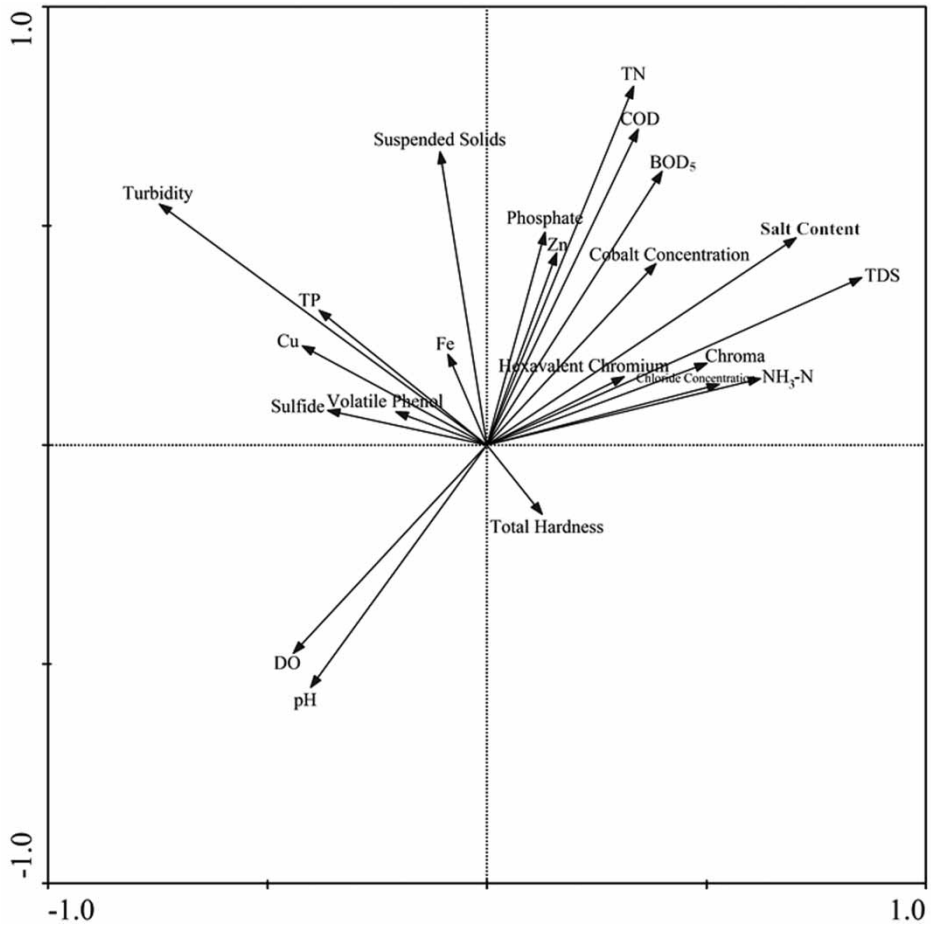


Figure 4 | Redundancy analysis of the water quality indices.

correspondence of multiple parameters. After the redundancy analysis of the data, the nine parameters with the highest contributions were selected for linear correlation analysis. It can be seen from the correlation analysis in Table 5 that COD and TDS have the highest correlation with BOD_5 and TN, and the correlation coefficient is 0.891 at the 0.01 significance level. The COD in water will reflect the size of the electrolytes in the water, and BOD_5 has a certain restriction on the TN content in the water. Additionally, the TN content also limits changes in the COD in the water. The TDS is also influenced by the COD, BOD_5, DO, TN and salt content in the water. As shown above, each parameter in the water will be affected by other factors, and the water quality will change.

Table 6 | Correlation analysis between the water quality indices and hydrology

	NDWI	NWI	EWI	AWEI-nsh
COD (mg/L)	0.530	0.592	−0.494	0.558
BOD_5 (mg/L)	0.357	0.427	0.366	−0.344
DO (mg/L)	−0.255	−0.412	−0.397	0.063
TN (mg/L)	0.520	0.636	0.574	−0.461
TSS (mg/L)	0.307	0.312	0.291	−0.118
Turbidity (NTU)	0.128	−0.005	0.007	−0.046
Salt content (Peters *et al.* 2007)	0.486	0.527	0.517	−0.134
TDS (mg/l)	0.532	0.679	0.645	−0.468
pH	−0.143	−0.393	−0.285	0.106
WQI	0.341*	0.312	0.271	−0.251

*$p < 0.05$.

Correlation between the water quality indices and water quality parameters (WQPs)

Sentinel-2 MSI imagery was used to retrieve the typical water index values in July 2016. Based on SPSS software, the correlation between the four classical water quality indices extracted from remote sensing images, and the nine WQPs were analyzed. The results are shown in Table 6.

According to the correlation analysis, the water quality parameters with high correlations were the COD, BOD_5, DO, TN, TSS, salt content, TDS, pH and NWI, suggesting that the NWI can effectively reflect most WQPs. The correlation between the NDWI and turbidity was as high as 0.128. Compared with other water indices, the band was not sensitive to water turbidity, and the turbidity correlation was generally low. The correlations between the NDWI and TDS and COD and TN were above 0.5. The EWI exhibited correlations with TN, the salt content and TDS of more than 0.5. The correlation between the AWEI-nsh and COD was 0.558.

In general, the correlation between the NWI and various water quality indicators in the classic water index was good, indicating that the WQI is sensitive to bands 1, 8, 11, and 12. Remote sensing image extraction is a sensitive reflection of the water indices.

Table 5 | Pearson's linear correlation matrix of the water quality indices

	COD	BOD_5	DO	TN	TSS	Turbidity	Salt content	TDS	pH
COD	1								
BOD_5	0.639**	1							
DO	−0.443**	−0.102	1						
TN	0.835**	0.891**	−0.243	1					
TSS	0.298	0.735**	0.253	0.672**	1				
Turbidity	0.087	0.416**	0.386*	0.297	0.664**	1			
Salt content	0.451**	0.377*	−0.401*	0.383*	0.114	−0.174	1		
TDS	0.891**	0.605**	−0.600**	0.779**	0.146	−0.137	0.520**	1	
pH	−0.624**	−0.532**	0.614**	−0.631**	−0.082	0.183	−0.336*	−0.734**	1

*$p < 0.05$.
**$p < 0.01$.

Correlation between the 2D spectrum modeling and water quality parameters (WQPs)

By using 2D spectrum modeling, the multiband pixel values of the corresponding points on each image point are extracted, and three empirical indices, normalized index, differential index and ratio index, are established. Correlation analysis was performed on nine water quality parameters and 13 bands of Sentinel-2 sampled at each point. The analysis results are shown in Figure 5.

The color bar on the right side of the figure shows the correlation between the band combination and WQPs, where dark red is the maximum value of positive correlation and dark blue is the maximum value of negative correlation. As can be seen from the figure, the correlation coefficients of the two WQPs and band modeling effects, COD and TDS, are better than 0.7.

From the correlation among the water index and the spectral band index and the measured WQPs, the correlation coefficient between the water index and the WQP is below 0.7, and the highest is 0.679. In contrast, the spectral band is established. The index correlation is better. Although the index can only reflect the combination algorithm of two bands, it can also find the correlation of the optimal band

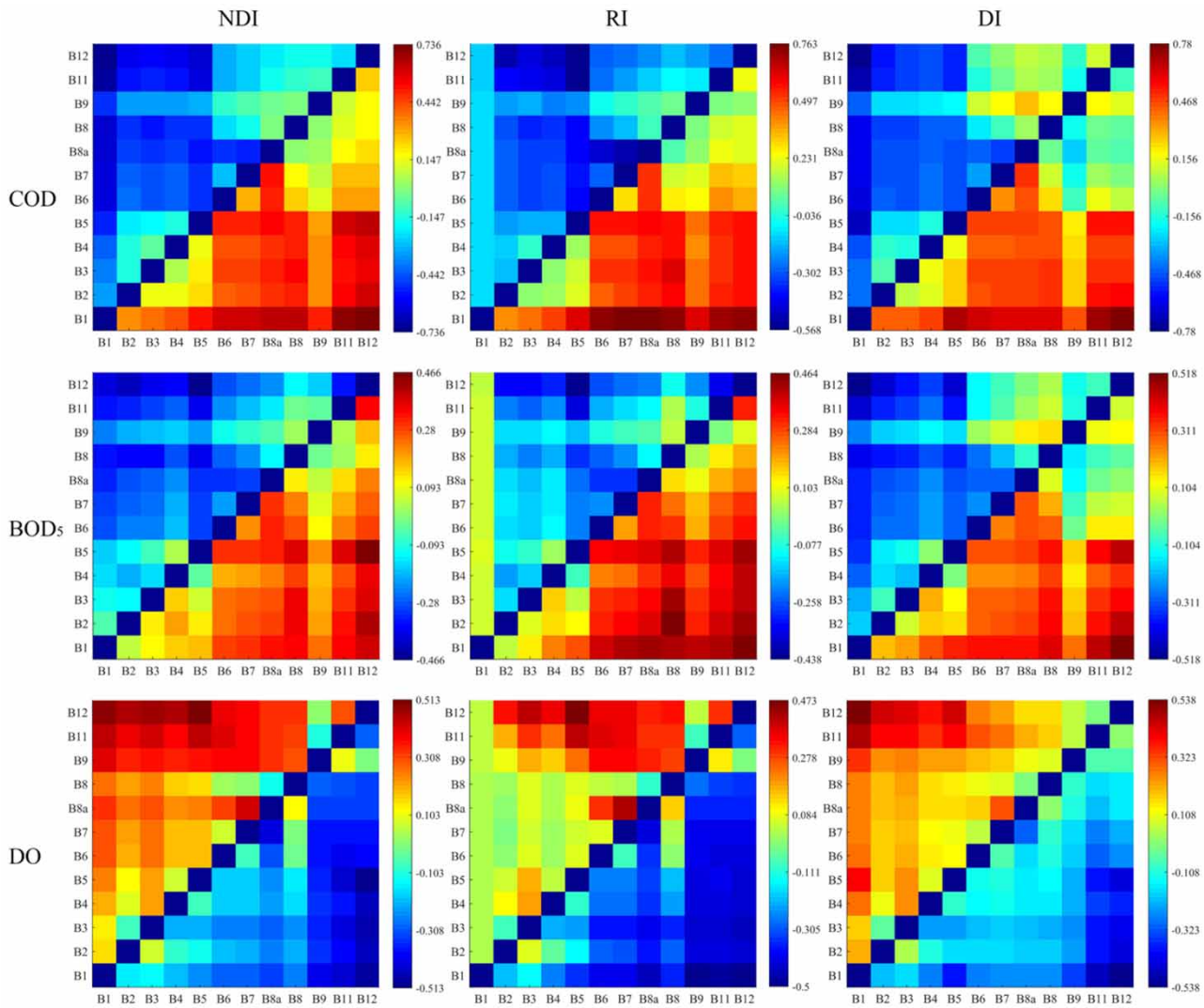


Figure 5 | 2D correlation coefficients between the optimal band of the two-band spectral modeling indices (DI, NDI, and RI) and 9 water quality parameters. The full color version of this figure is available in the online version of this paper, at http://dx.doi.org/10.2166/ws.2020.381. *(Continued.)*

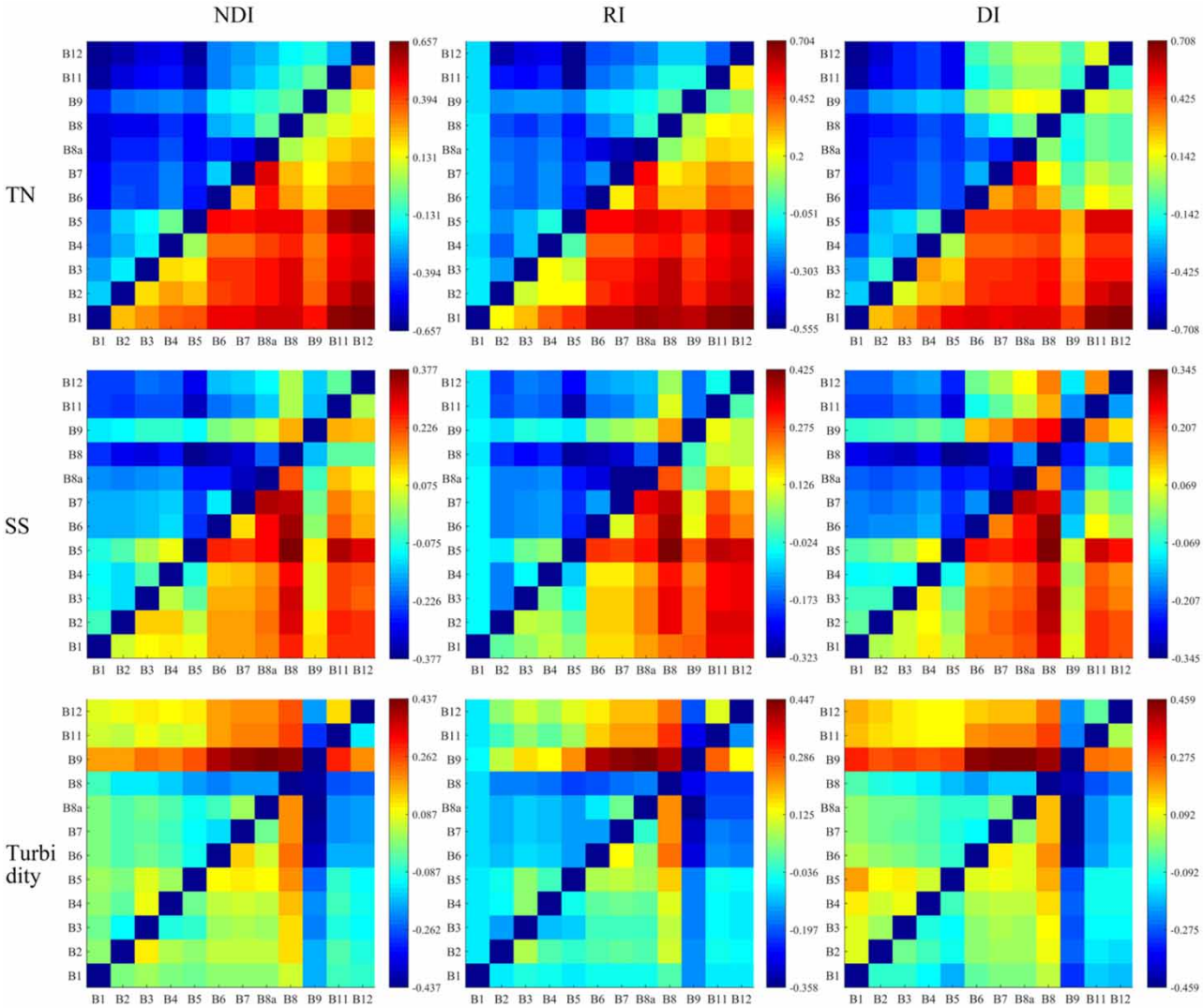


Figure 5 | Continued.

and the band combination algorithm from 13 bands. Overall, the nine WQPs selected can be reflected in different bands of the image.

Comparison of the water indices and modeled spectral indices for water quality extraction

Table 7 shows the correlation results for the nine water quality parameters and the classic water index and the 2D modeled spectral index. The classic water index with the highest correlation among the nine water quality parameters was extracted, and the three water-based indicators with the highest correlation were extracted from the three 2D modeled spectral indices. Overall, compared with the classic water index, the 2D modeled spectral index has a better effect in water quality extraction. The overall effect is best when the TDS is inverted by the DI, and the correlation coefficient reaches 0.787, while the correlation coefficient of the NWI inversion of TDS is 0.679. Based on the two correlation coefficients of the nine WQPs, the correlation coefficients of the BOD_5, TN, TSS, salt content, and TDS modeled spectral indices are not much different from those for the water indices. This result suggests that the two methods have the same effect on the inversion of these five

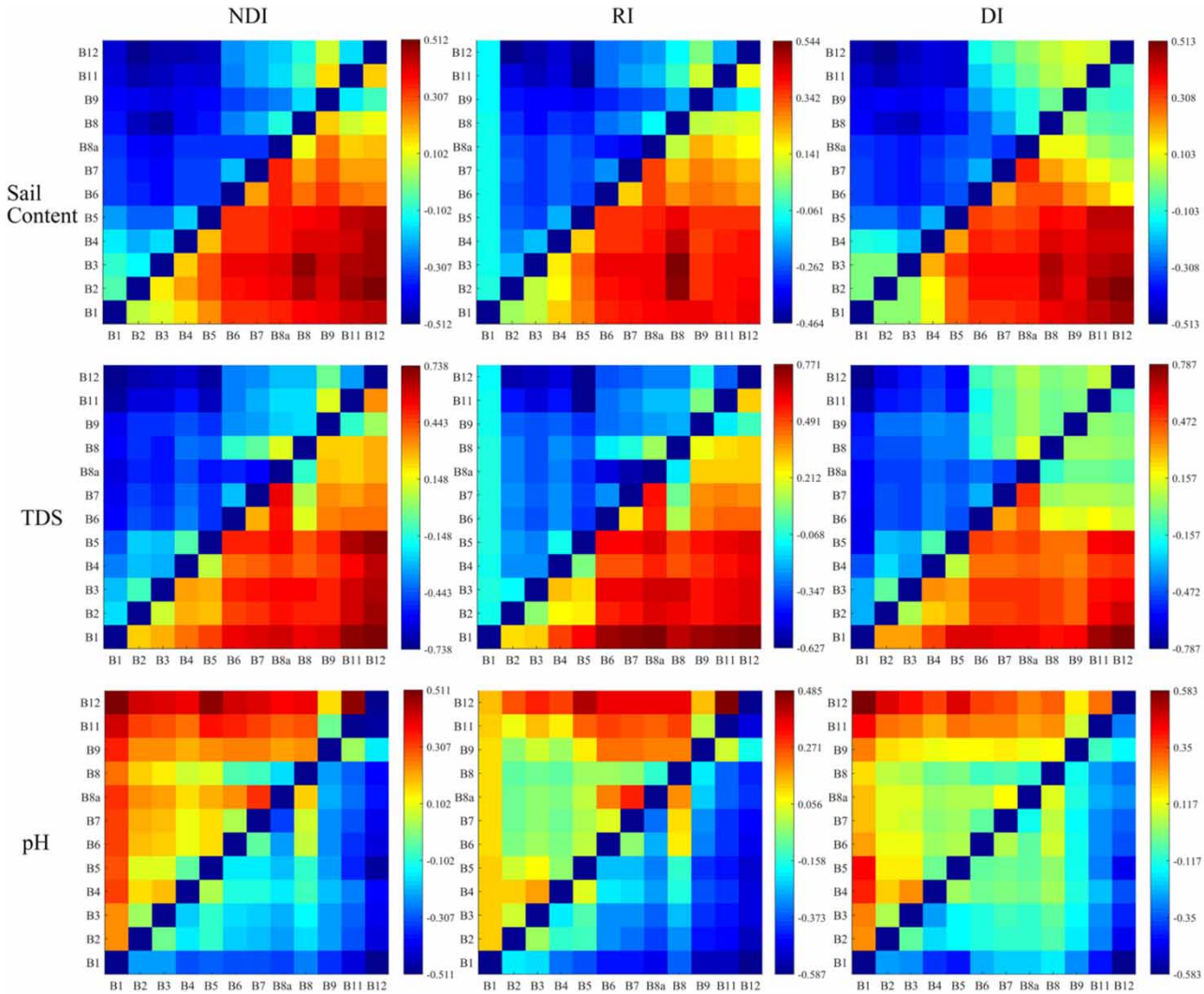


Figure 5 | Continued.

WQPs. In the inversion of DO, turbidity and the pH, the 2D modeled spectral index is much different than the water index. The inversion results of the classical water index for DO and the pH had negative correlation coefficients of −0.412 and −0.393, respectively. The inversion results of the 2D modeled spectral index showed positive correlation coefficients of 0.538 and 0.583. The two methods display a large difference in the inversion of turbidity, and the 2D modeled band index can better invert the degree of turbidity in water.

Overall, the 2D modeled band index can effectively invert the WQI, and the accuracy is higher than that of the water index. Thus, the 2D modeled band index is the optimal spectral index for water quality monitoring.

WQI and modeled spectral index

To estimate the overall water quality, a WQI index was introduced. WQI is a comprehensive evaluation of water quality indicators. It can analyze the overall effect of water quality through selected indicators, so that it can analyze the overall water quality situation more intuitively. We employed nine types of water quality selected from RDA in the WQI calculation, obtained WQI of 38 sampling points, and compared and analyzed WQI in the water index and modeled spectral index, respectively. As can be seen in Table 7, among the correlations between WQI and the four water indexes, the correlation between WQI and

Table 7 | Correlation analysis between the water quality indices and 2D spectrum modeling

WQP	Water index	Correlation coefficient	2D spectral model	Correlation coefficient
COD (mg/L)	$AWEI_{-nsh} = 4 \times (\rho B_3 - \rho B_{11}) - (2.5 \times \rho B_8 + 2.75 \times \rho B_{12})$	0.558	DI (R12-R1)	0.78
BOD_5 (mg/L)	$NWI = \frac{B_1 - (B_8 + B_{11} + B_{12})}{B_1 + B_8 + B_{11} + B_{12}}$	0.427	DI (R12-R1)	0.518
DO (mg/L)	$NWI = \frac{B_1 - (B_8 + B_{11} + B_{12})}{B_1 + B_8 + B_{11} + B_{12}}$	−0.412	DI (R12-R1)	0.538
TN (mg/L)	$NWI = \frac{B_1 - (B_8 + B_{11} + B_{12})}{B_1 + B_8 + B_{11} + B_{12}}$	0.636	DI (R12-R1)	0.708
TSS (mg/L)	$NWI = \frac{B_1 - (B_8 + B_{11} + B_{12})}{B_1 + B_8 + B_{11} + B_{12}}$	0.312	RI (R8/R5)	0.425
Turbidity (NTU)	$NDWI = \frac{B_3 - B_8}{B_3 + B_8}$	0.128	DI (R8a-R9)	0.459
Salt content (Peters *et al.* 2007)	$NWI = \frac{B_1 - (B_8 + B_{11} + B_{12})}{B_1 + B_8 + B_{11} + B_{12}}$	0.527	RI (R8/R3)	0.544
TDS (mg/l)	$NWI = \frac{B_1 - (B_8 + B_{11} + B_{12})}{B_1 + B_8 + B_{11} + B_{12}}$	0.679	DI (R12-R1)	0.787
pH	$NWI = \frac{B_1 - (B_8 + B_{11} + B_{12})}{B_1 + B_8 + B_{11} + B_{12}}$	−0.393	DI (R12-R1)	0.583

NDWI reached 0.341 at the level of 0.05, and the correlation with the other three water indexes was low.

Correlation analysis between WQI and three modeling indexes shows that RI of WQI and NDI, RI, and DI are 0.63, 0.69, and 0.62, respectively (Figure 6). The modeled water index has a better correlation with WQI than the classic water index.

Model performance in estimating the WQI

The WQI was calculated from the nine selected WQPs, and the WQI values of 38 sampling points were obtained. For prediction, 24 models were selected, and 14 values were used for prediction verification. Machine learning (SVM and RF) and PLSR methods were chosen for prediction.

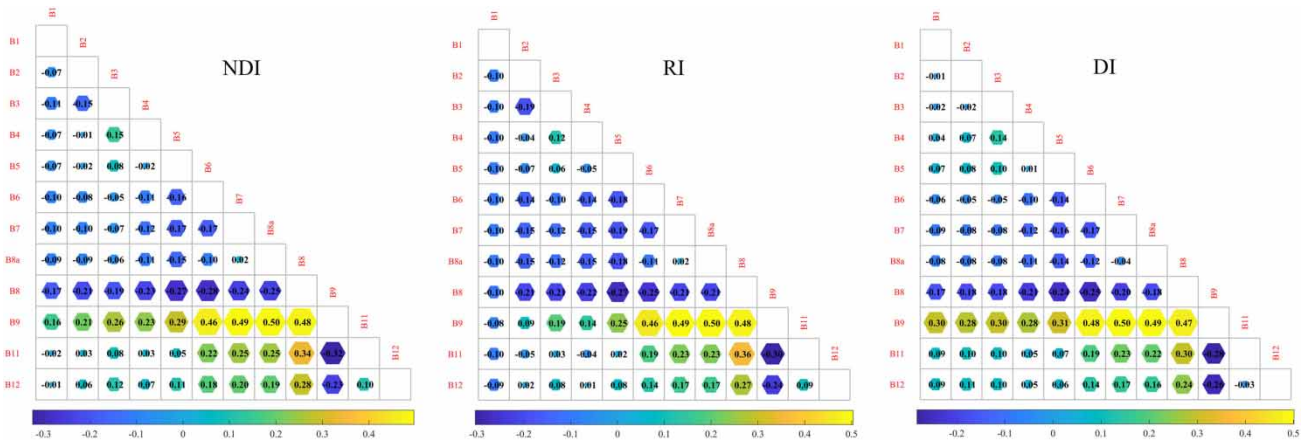


Figure 6 | Plots showing the relationship between the WQI and four models with the training dataset.

The machine learning model was used as the main model for prediction, the partial least squares model was used as the auxiliary model, and the PLSR-SVM model was established for comparison to analyze the accuracy of each model.

From Table 8, by calibration modeling with 24 values, the modeling of the machine learning methods was performed. The results indicate that the RF modeling effect yields the best correlation ($R^2c = 0.814$ and RMSEc = 109.411); however, the SVM modeling effect ($R^2c = 0.0.770$ and RMSEc = 120.629) is preferred. Machine learning is a linear process, and the linear correlation of the PLSR method was low. The correlation coefficient and root mean square error of the PLSR-SVM are better than those of the previous three methods; notably, the correlation of $R^2c = 0.873$ is better, and the degree of data dispersion is better at RMSEc = 102.334.

From the validation dataset, the modeling effect R^2v displays the following order: PLSR-SVM > RF > PLSR > SVM; additionally, the highest R^2v is 0.870. From RMSEv, SVM > RF > PLSR > PLSR-SVM, and the root mean square error (RMSEv) is small, indicating that the data are not discrete and the prediction effect is good. Additionally, RPD = Stdev/RMSE; a large RPD corresponds to a satisfactory prediction effect. The result indicates that the prediction effect of PLSR-SVM is better than that of the other methods, with RMSE = 62.927 and RPD = 2.755.

The WQI is modeled and predicted with the four modeling methods. Figure 7(a)–7(d) shows the modeling of 24 data points. In the modeling diagram, the red line around the red points is the fitted line. The surrounding red interval is the confidence interval. The smaller the interval is, the better the data modeling effect. It can be seen from the figure that the interval of the PLSR-SVM method in Figure 7(d) is smaller than that of the other three model regions. At $R^2c = 0.873$, the model effect is best. In Figure 7(e)–7(h), the predictions involve 14 data points and the original data; the blue point is the prediction point, the blue line is the prediction fitting line, and the blue interval is the confidence interval of the prediction. In the interval, the confidence interval area of Figure 7(f) and 7(h) is the smallest, although the angle between the fitted line and the 1:1 line of Figure 7(h) is the smallest. Additionally, the predicted points are all distributed around the 1:1 line, and the predicted values yield $R^2v = 0.87$ and RPD = 2.755, which reflect the good prediction effect of the PLSR-SVM method.

DISCUSSION

The reflectance of different spectral bands in remote sensing images is different from the reflectance of surface vegetation, soil and water. For such problems, differences are often analyzed (Pahlevan *et al.* 2019). Extracting the water pixels of all pixels from satellite images is the most important step in remote sensing detection, and it is also a popular issue in current research. Long-term water masks can be used to analyze time-varying water information and provide a reliable basis for surface water research (Fisher *et al.* 2016). Sentinel images can be used to extract water features under complex terrains such as mountains, snow, and cities (Kaplan & Avdan 2017). Water indices based on spectral bands can be used to quickly identify surface water and highlight certain features (Xu 2006). This article studies the extraction of the DN

Table 8 | Calibration ($n = 24$) and validation ($n = 14$) results of the WQI

	Calibration dataset			Validation dataset			
Calibration model	N	R^2c	RMSEc	N	R^2v	RMSEv	RPD
RF	24	0.814	109.411	14	0.783	97.049	1.786
SVM	24	0.770	120.629	14	0.746	118.166	1.467
PLSR	24	0.777	121.758	14	0.753	84.357	2.055
PLSR-SVM	24	0.873	102.334	14	0.870	62.927	2.755

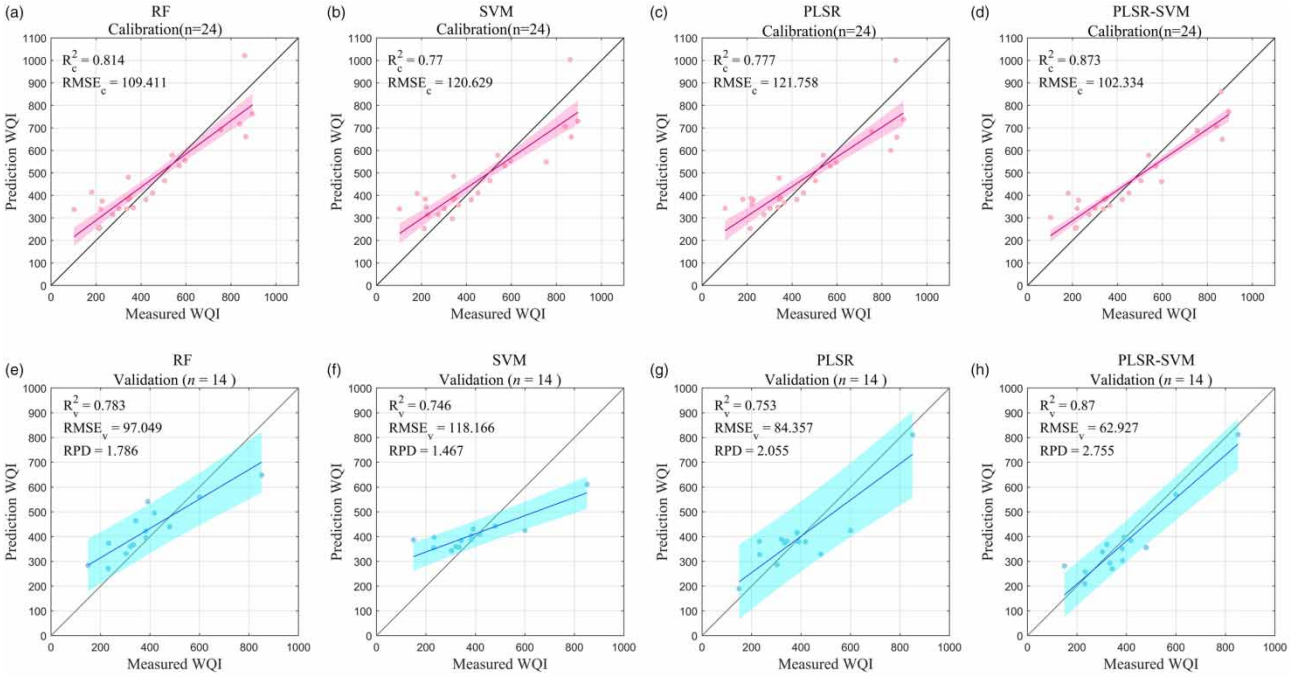


Figure 7 | 2D correlation coefficients between the optimal band of the two-band spectral modeling index (DI, NDI, and RI) and the WQI.

values corresponding to the Sentinel-2 data and sampling points and establishes a WQI and a two-dimensional spectral model based on this approach. First, the correlations between the water index and individual WQPs are studied. As shown in Table 6, the correlation between the NDWI, NWI, EWI and TDS is better than 0.5. The effect of AWEI-nsh on COD inversion reaches 0.558. Notably, the water spectral index can be used to invert the corresponding WQPs in the band calculation. Based on this method, we have established a two-dimensional spectral model of the Sentinel-2 MSI and three model indices; namely, the NDI, RI and DI. By comparing the correlation between the water index and the 2D spectrum model with that for a single WQI, the effect of the two-dimensional spectrum model is better than that of the water index. The two-dimensional spectrum model can filter the spectrum information and select the optimal band combination for analysis (Hong *et al.* 2018); such methods used in water quality monitoring will allow us to find more sensitive bands and band combinations for a single type of water quality. Through the correlation analysis of WQPs, water quality indices and spectral bands, the water quality of sampling points in the basin can be analyzed in depth.

In the Ebinur Basin, located in an arid area, many scholars have used Landsat OLI data to extract water quality information through different methods, identify water bodies in arid areas through different modeling methods, and analyze the correlation between the WQPs and water indices (Li *et al.* 2016; Wang *et al.* 2017). To extract and analyze the overall water quality of the river basin and facilitate an analysis of the overall water quality safety status, single types of WQPs are integrated into the comprehensive water index WQI. The WQI can be used to establish the relationship between multispectral bands and water quality reflectance, analyze the correlation between spectral bands and water quality, and establish a linear relationship with the sensitive bands of different indicators related to water quality (Feyisa *et al.* 2014). Based on the preliminary exploration of single types of WQPs and spectral modeling, we selected an improved 2D band model and WQI for further comprehensive water quality research. Figure 7 shows that the correlation between the three two-dimensional spectral modeling methods and the WQI is above 0.75, and the modeling effect is satisfactory. The WQI is based on a combination of different WQPs and converts a large number of WQPs into a single quantity (Sánchez *et al.*

2007; Feng *et al.* 2018). This method is currently widely used in water quality evaluation studies (Wu & Chen 2013). To conduct further prediction research on water quality, on this basis, the WQI was modeled and predicted by machine learning methods. This paper established four models: SVM, RF, PLSR and PLSR-SVM models. The PLSR-SVM provided a better model result and prediction of the basin WQI than did the other models. The PLSR-SVM and RF models have been used in other river basin water quality studies, and the prediction results have been shown to be better and more accurate than those of other models (Mananze *et al.* 2018). This finding confirms that the method used in this study is reasonable and provides a foundation for subsequent research.

In future research, it is necessary to consider the distribution of water samples and the selection of measured data in the entire basin.

CONCLUSIONS

This study explores the relationship between the water quality in the Ebinur Lake watershed and the multispectral bands of Sentinel-2 MSI data. The relevant WQPs are related to the water index and analyzed by spectral bands. The following results are obtained.

The Z-score and RDA are used to reduce 22 WQPs to nine while dividing them into different groups. COD, BOD_5, DO, TN, TSS, turbidity, the salt content, TDS and the pH are the nine selected WQPs with high contribution values. TDS, COD, and TN are the most influential WQPs.

The WQI is established through the selected nine WQPs, and modeling and prediction are performed through machine learning and linear correlation models. The PLSR-SVM model with a linear correlation and machine learning is the best model for modeling, with $R^2v = 0.87$ and $RPD = 2.755$; the predictions with this approach are very accurate, and this approach can provide an effective method for water prediction.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (41771470), the National Natural Science Foundation of China (41961059), and the National Natural Science Foundation of China (41561089). We are especially grateful to the anonymous reviewers and editors for reviewing the manuscript and offering instructive comments.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Abuduwaili, J., Gabchenko, M. V. & Junrong, X. 2008 Eolian transport of salts – A case study in the area of Lake Ebinur (Xinjiang, Northwest China). *Journal of Arid Environments* **72** (10), 1843–1852.

Alghamdi, A. G., Aly, A. A., Aldhumri, S. A. & Al-Barakaha, F. N. 2020 Hydrochemical and quality assessment of groundwater resources in Al-Madinah City, Western Saudi Arabia. *Sustainability* **12** (8), 3106.

Baban, S. M. 1993 Detecting water quality parameters in the Norfolk Broads, U.K., using Landsat imagery. *International Journal of Remote Sensing* **14** (7), 1247–1267.

Boser, B. E. 2008 A training algorithm for optimal margin classifiers. *Proceedings of Annual ACM Workshop on Computational Learning Theory* **5**, 144–152.

Bukata, R. P., Bruton, J. E. & Jerome, J. H. 1983 Use of chromaticity in remote measurements of water quality. *Remote Sensing of Environment* **13** (2), 161–177.

Carrascal, L. M., Galván, I. & Gordo, O. 2010 Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos* **118** (5), 681–690.

Chen, J., Fan, B., Li, J., Wang, X. N., Li, W. W., Cui, L. & Liu, Z. T. 2020 Development of human health ambient water quality criteria of 12 polycyclic aromatic hydrocarbons (PAH) and risk assessment in China. *Chemosphere* **252**, 126590.

Douglas, R. K., Nawar, S., Alamar, M. C., Mouazen, A. M. & Coulon, F. 2018 Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. *Science of the Total Environment* **616–617**, 147–155.

Ewaid, S. H., Abed, S. A. & Kadhum, S. A. 2018 Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis. *Environmental Technology & Innovation* **11**, 390–398.

Feng, T., Wang, C., Hou, J., Wang, P., Liu, Y., Dai, Q. S., Yang, Y. Y. & You, G. X. 2018 Effect of inter-basin water transfer on water quality in an urban lake: a combined water quality index algorithm and biophysical modelling approach. *Ecological Indicators* **92**, 61–71.

Fernanda, W., Enner, A., Thanan, R., Luiz, R., Nariane, B. & Nilton, I. 2017 Remote sensing of the chlorophyll-a based on OLI/Landsat-8 and MSI/Sentinel-2A (Barra Bonita reservoir, Brazil). *Anais Da Academia Brasileira De Ciências* **90**, 1987–2000.

Feyisa, G. L., Meilby, H., Fensholt, R. & Proud, S. R. 2014 Automated water extraction index: a new technique for surface water mapping using Landsat imagery. *Remote Sensing of Environment* **140**, 23–35.

Fisher, A., Flood, N. & Danaher, T. 2016 Comparing Landsat water index methods for automated water classification in eastern Australia. *Remote Sensing of Environment* **175**, 167–182.

Gao, G., Xiao, K. & Chen, M. 2019 An intelligent IoT-based control and traceability system to forecast and maintain water quality in freshwater fish farms. *Computers Electronics in Agriculture* **166**, 105013.

GB3838-2002. 2002 *Environmental Quality Standard for Surface Water (GB3838-2002)*.

Geladi, P. & Kowalski, B. R. 1986 Partial least-squares regression: a tutorial. *Analytica Chimica Acta* **185** (1), 1–17.

Gitelson, A. A. 1991 Aerospace remote sensing monitoring of inland water quality. *Proceedings of SPIE – The International Society for Optical Engineering* **1492**, 307–318.

Gower, J., King, S. & Goncalves, P. 2008 Global monitoring of plankton blooms using MERIS MCI. *International Journal of Remote Sensing* **29** (21), 6209–6216.

Halgamuge, M. N. & Davis, D. 2019 Lessons learned from the application of machine learning to studies on plant response to radio-frequency. *Environmental Research* **178**, 108634.

Hong, Y. S., Chen, S., Zhang, Y. H., Chen, Y., Yu, L., Liu, Y., Liu, Y., Cheng, H. & Liu, Y. 2018 Rapid identification of soil organic matter level via visible and near-infrared spectroscopy: effects of two-dimensional correlation coefficient and extreme learning machine. *Science of the Total Environment* **644**, 1232–1243.

Houborg, R. & McCabe, M. F. 2018 A hybrid training approach for leaf area index estimation via cubist and random forests machine-learning. *ISPRS Journal of Photogrammetry and Remote Sensing* **135**, 173–188.

Hussein, A. M., Abd Elaziz, M., Abdel Wahed, M. S. M. & Sillanpää, M. 2019 A new approach to predict the missing values of algae during water quality monitoring programs based on a hybrid moth search algorithm and the random vector functional link network. *Journal of Hydrology* **575**, 852–863.

Kaplan, G. & Avdan, U. 2017 Object-based water body extraction model using Sentinel-2 satellite imagery. *European Journal of Remote Sensing* **50** (1), 137–143.

Kaya, G. T. 2013 A hybrid model for classification of remote sensing images with linear SVM and support vector selection and adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* **6** (4), 1988–1997.

Li, Y., Gong, X., Guo, Z., Xu, K. & Hu, D. 2016 An index and approach for water extraction using Landsat–OLI data. *International Journal of Remote Sensing* **37** (16), 3611–3635.

Liang, Z., Zou, R., Chen, X., Ren, T., Su, H. & Liu, Y. 2020 Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach. *Journal of Hydrology: Regional Studies* **581**, 124432.

Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., Leahy, P., Wilson, P. & Western, A. W. 2018 Key factors influencing differences in stream water quality across space. *Wiley Interdiplinary Reviews: Water* **5** (1), e1260.

Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G. & Buttafuoco, G. 2017 Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. *Geoderma* **288**, 175–183.

Luo, C. L. 2015 Comparative study on water area extraction of Ebinur Lake based on water body index. *Science and Technology Innovation Herald(China)* **24**, 34–35.

Ma, R. H. & Dai, J. F. 2005 Chlorophyll-a concentration estimation with field spectra of water-body near Meiliang Bayou in Taihu Lake. *Journal of Remote Sensing(in China)* **01**.

Main-Knorn, M., Pflug, B., Debaecker, V. & Louis, J. 2015 Calibration and validation plan for the l2A processor and products of the Sentinel-2 mission. *International Archives of the Photogrammetry Remote Sensing & S* **XL-7/W3**, 1249–1255.

Mananze, S., Pôças, I. & Cunha, M. 2018 Retrieval of maize leaf area index using hyperspectral and multispectral data. *Remote Sensing* **10** (12), 1942.

McFeeters, S. K. 1996 The use of the normalized difference water index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing* **17** (7), 1425–1432.

Meyers, G., Kapelan, Z. & Keedwell, E. 2017 Short-term forecasting of turbidity in trunk main networks. *Water Research* **124**, 67–76.

Moses, W. J., Gitelson, A. A., Berdnikov, S. & Povazhnyy, V. 2009 Satellite estimation of chlorophyll-*a* concentration using the Red and NIR bands of MERIS – The Azov sea case study. *IEEE Geoscience & Remote Sensing Letters* **6** (4), 845–849.

Nawar, S., Buddenbaum, H., Hill, J., Kozak, J. & Mouazen, A. M. 2016 Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. *Soil and Tillage Research* **155**, 510–522.

Nie, F. P., Zhu, W. & Li, X. L. 2020 Decision tree SVM: an extension of linear SVM for non-linear classification. *Neurocomputing* **401**, 153–159.

Pahlevan, N., Chittimalli, S. K., Balasubramanian, S. V. & Vellucci, V. 2019 Sentinel-2/Landsat-8 product consistency and implications for monitoring aquatic systems. *Remote Sensing of Environment* **220**, 19–29.

Pesce, S. F. & Wunderlin, D. A. 2000 Use of water quality indices to verify the impact of Córdoba City (Argentina) on Suqua River. *Water Research* **34** (11), 2915–2926.

Peters, J., Baets, B. D., Verhoest, N. E. C., Samson, R., Degroeve, S., Becker, P. D. & Huybrechts, W. 2007 Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling* **207** (2), 304–318.

Peterson, K. T., Sagan, V. & Sloan, J. J. 2020 Deep learning-based water quality estimation and anomaly detection using Landsat-8/Sentinel-2 virtual constellation and cloud computing. *GIScience & Remote Sensing* **57**, 510–525.

Ramakrishnaiah, C. R., Sadashivaiah, C. & Ranganna, G. 2009 Assessment of water quality index for the groundwater in Tumkur Taluk, Karnataka State, India. *Journal of Chemistry* **6** (2), 523–530.

Rokni, K., Ahmad, A., Solaimani, K. & Hazini, S. 2015 A new approach for surface water change detection: integration of pixel level image fusion and image classification techniques. *International Journal of Applied Earth Observation and Geoinformation* **34**, 226–234.

Sánchez, E., Colmenarejo, M. F., Vicente, J., Rubio, A., García, M. G., Travieso, L. & Borja, R. 2007 Use of the water quality index and dissolved oxygen deficit as simple indicators of watersheds pollution. *Ecological Indicators* **7** (2), 315–328.

Şener, Ş., Şener, E. & Davraz, A. 2017 Evaluation of water quality using water quality index (WQI) method and GIS in Aksu River (SW-Turkey). *Science of the Total Environment* **584–585** (apr.15), 131–144.

Shao, M., Tang, X. Y., Zhang, Y. H. & Li, W. J. 2006 City clusters in China: air and surface water pollution. *Frontiers in Ecology the Environment* **4** (7), 353–361.

Sidike, A., Zhao, S. H. & Wen, Y. M. 2014 Estimating soil salinity in Pingluo County of China using QuickBird data and soil reflectance spectra. *International Journal of Applied Earth Observation and Geoinformation* **26**, 156–175.

Staponites, L. R., Barták, V., Bílý, M. & Simon, O. P. 2019 Performance of landscape composition metrics for predicting water quality in headwater catchments. *Scientific Reports* **9** (1), 14405.

State Environmental Protection Administration of China. 2002 *Water and Wastewater Monitoring and Analysis Methods*, 4th edn. China Environmental Science Press, Beijing.

Stéfani, N., Doxaran, D., Ody, A., Vanhellemont, Q., Lafon, V., Lubac, B. & Gernez, P. 2017 Atmospheric corrections and multi-conditional algorithm for multi-sensor remote sensing of suspended particulate matter in low-to-high turbidity levels coastal waters. *Remote Sensing* **9** (1), 61.

Terrado, M., Barceló, D., Tauler, R., Borrell, E., Campos, S. & Barceló, D. 2010 Surface-water-quality indices for the analysis of data generated by automated sampling networks. *TrAC Trends in Analytical Chemistry* **29** (1), 40–52.

Tyler, A. N., Hunter, P. D., Spyrakos, E., Groom, S., Constantinescu, A. M. & Kitchen, J. 2016 Developments in Earth observation for the assessment and monitoring of inland, transitional, coastal and shelf-sea waters. *Science of the Total Environment* **572**, 1307–1321.

Wang, X. P., Zhang, F. & Ding, J. L. 2017 Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake watershed, China. *Scientific Reports* **7** (1), 12858.

Wang, J. Z., Ding, J. L., Abulimiti, A. & Cai, L. H. 2018a Quantitative estimation of soil salinity by means of different modeling methods and visible-near infrared (VIS–NIR) spectroscopy, Ebinur Lake Wetland, Northwest China. *Peerj* **6**, e4703.

Wang, X., Xie, S., Zhang, X., Chen, C., Guo, H., Du, J. & Duan, Z. 2018b A robust Multi-Band Water Index (MBWI) for automated extraction of surface water from Landsat 8 OLI imagery. *International Journal of Applied Earth Observation and Geoinformation* **68**, 73–91.

Wang, G. Q., Li, J. W., Sun, W. C., Xue, B. L., Yinglan, A. L. & Liu, T. X. 2019 Non-point source pollution risks in a drinking water protection zone based on remote sensing data embedded within a nutrient budget model. *Water Research* **157** (15), 238–246.

Wang, J., Ding, J., Yu, D., Teng, D., He, B., Chen, X., Ge, X., Zhang, Z., Wang, Y., Yang, X., Shi, T. & Su, F. 2020 Machine learning-based detection of soil salinity in an arid desert region, Northwest China: A comparison between Landsat-8 OLI and Sentinel-2 MSI. *Science of the Total Environment* **707**, 136092.

WHO 2008 *Guidelines for Drinking-Water Quality*. World Health Organization, Geneva, Switzerland.

Wu, Y. P. & Chen, J. 2013 Investigating the effects of point source and nonpoint source pollution on the water quality of the East River (Dongjiang) in South China. *Ecological Indicators* **32**, 294–304.

Xu, H. Q. 2006 Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing* **27** (14), 3025–3033.

Yang, X., Qin, Q., Grussenmeyer, P. & Koehl, M. 2018 Urban surface water body detection with suppressed built-up noise based on water indices from Sentinel-2 MSI imagery. *Remote Sensing of Environment* **219**, 259–270.

Yin, J., Liu, H. & Chen, X. L. 2018 Dynamic change in the water-level fluctuation zone of the Danjiangkou reservoir and its influence on water quality. *Sustainability* **10** (4), 1025.

Yu, X. Y. & Jiang, N. 2003 Analyzing lake area change in Ebinur by integration of RS and GIS techniques. *Journal of Remote Sensing(in China)* **15** (1), 81–84.

Yu, D. H., Wang, Y. H., Deng, Z. D., Gou, Y. Z. & Zhu, B. Z. 2008 Research progress in remote sensing technology for inland water quality monitoring. *China Water & Wastewater* **24** (22), 12–16.

Zeinalzadeh, K. & Rezaei, E. 2017 Determining spatial and temporal changes of surface water quality using principal component analysis. *Journal of Hydrology: Regional Studies* **13**, 1–10.

Zhang, Z., Ding, J., Zhu, C., Wang, J., Ma, G., Ge, X., Li, Z. & Han, L. 2021 Strategies for the efficient estimation of soil organic matter in salt-affected soils through Vis-NIR spectroscopy: optimal band combination algorithm and spectral degradation. *Geoderma* **382**, 114729.

Zhao, C. H., Gao, B., Zhang, L. J. & Wan, X. Q. 2018 Classification of Hyperspectral Imagery based on spectral gradient, SVM and spatial random forest. *Infrared Physics & Technology* **95**, 61–69.

Zheng, Z., Li, Y., Guo, Y., Xu, Y., Liu, G. & Du, C. 2015 Landsat-based long-term monitoring of total suspended matter concentration pattern change in the wet season for Dongting Lake, China. *Remote Sensing* **7** (10), 13975–13999.

Zotou, I., Tsihrintzis, V. A. & Gikas, G. D. 2020 Water quality evaluation of a lacustrine water body in the Mediterranean based on different water quality index (WQI) methodologies. *Journal of Environmental Science Health, Part A* **55** (5), 537–548.

First received 4 June 2020; accepted in revised form 16 December 2020. Available online 30 December 2020