Science of the Total Environment 916 (2024) 170225

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Estimating low concentration heavy metals in water through hyperspectral analysis and genetic algorithm-partial least squares regression

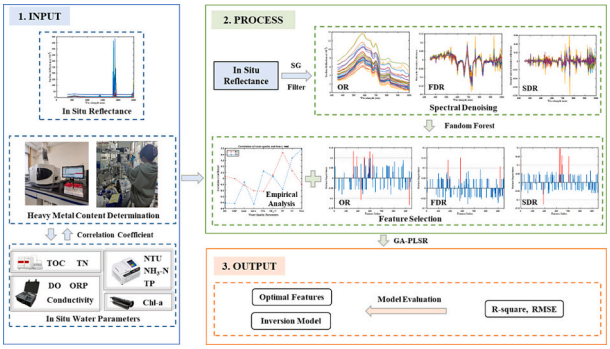
Yukun Lin [a,b], Jiaxin Gao [a,1], Yaojen Tu [a,b,*], Yuxun Zhang [a], Jun Gao [a,b]

[a] School of Environmental and Geographical Sciences, Shanghai Normal University, Shanghai 200234, China
[b] Yangtze River Delta Urban Wetland Ecosystem National Field Scientific Observation and Research Station, Shanghai 200234, China

HIGHLIGHTS

- A retrieval method based on multi-level combined features is proposed.
- The features related to TOC and Chl-a could improve Cu estimation accuracy.
- The feature bands related to TP could improve Fe estimation accuracy.
- Hyperspectral data can be applied for low concentration heavy metal inversion.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Ashantha Goonetilleke

Keywords:
Low concentration heavy metal
In situ hyperspectral
GA-PLSR
Water heavy metal retrieval

ABSTRACT

Hyperspectral spectrum enables assessment of heavy metal content, but research on low concentration in water is limited. This study employed in situ hyperspectral data from Dalian Lake, Shanghai to develop a machine learning model for accurately determining heavy metal concentrations. Initially, we employed a combination of empirical analysis and algorithm-based analysis to identify the optimal features for retrieving Cu and Fe ions. Based on the correlation coefficients between heavy metals and water quality, the feature bands for TOC, Chl-a and TP were selected as empirical features. Algorithm-based feature selection was conducted by employing the random forest (RF) approach with the original spectrum (OR), first-order derivative reflectance (FDR), and second-order derivative reflectance (SDR). For the development of a prediction model, we utilized the Genetic Algorithm-Partial Least Squares Regression (GA-PLSR) approach for Cu and Fe ions inversion. Our findings demonstrated that the integration of both empirical features and algorithm-selected features resulted in superior performance compared to using algorithm-selected features alone. Importantly, the crucial wavelength data primarily located at 497, 665, 686, 831 and 935 nm showed superior results for Cu retrieval, while wavelengths of 700, 746, 801, 948, and 993 nm demonstrated better results for Fe retrieval. These results also displayed that the GA-PLSR model outperformed both the PLSR and RF models, exhibiting an R^2 of 0.75, RMSE of 0.004, and MRE of 0.382 for Cu inversion. For Fe inversion, the GA-PLSR model outperformed other models with an R^2 of

* Corresponding author at: School of Environmental and Geographical Sciences, Shanghai Normal University, Shanghai 200234, China.
E-mail address: yjtu@shnu.edu.cn (Y. Tu).
[1] Contributed equally to the work.

https://doi.org/10.1016/j.scitotenv.2024.170225
Received 19 September 2023; Received in revised form 10 January 2024; Accepted 15 January 2024
Available online 20 January 2024

0.73, RMSE of 0.036, and MRE of 0.464. This research provides a scientific basis and data support for monitoring low concentrations of heavy metals in water bodies using hyperspectral remote sensing techniques.

1. Introduction

The relentless pursuit of global economic progress has resulted in the degradation of natural resources due to excessive exploitation and anthropogenic activities (Mondal and Bordoloi, 2023). Among these resources, water is an essential resource that plays a vital role in many aspects of our lives. However, water quality has suffered varying degrees of deterioration, and a primary concern is metal toxicity (Besser and Leib, 2007). It is estimated that around 40 % of the Earth's lakes and rivers have been affected by heavy metal contamination (Zhou et al., 2020). Metal toxicity in water is a significant issue that poses various risks to both human health and the ecosystem. Specifically, heavy metals such as lead (Pb), copper (Cu) and cadmium (Cd) can contaminate water sources through industrial activities, mining, and improper waste disposal, posing serious health hazards (Azizullah et al., 2021; Pandey and Kumari, 2023). These toxic metals can have severe effects on human health, including damage to the nervous system, kidneys, and liver (Le et al., 2019; Pinto et al., 2019). Additionally, metal toxicity has adverse effects on the environment. Metals leaching into water bodies can accumulate in aquatic organisms, disrupting their growth and reproduction, consequently leading to a decline in biodiversity and disturbing the ecological balance of aquatic ecosystems (Bashir et al., 2020). Moreover, contaminated water negatively impacts agricultural practices by impairing crop growth and reducing soil fertility (Alengebawy et al., 2021). To address these issues, effective monitoring and regular testing of water quality are crucial to identify and address metal contamination at early stages. Consequently, there is an urgent need for development of reliable, cost-effective, and efficient monitoring techniques for water heavy metals.

Remote sensing techniques are extensively utilized in assessing environmental quality by analyzing optical characteristics derived from spectrum data (Wang et al., 2018; Wang and Yang, 2019). High-dimensional hyperspectral data provides valuable insights into water quality parameters (Cai et al., 2022; Chi et al., 2016). Therefore, recent efforts have focused on developing novel approaches that utilize non-destructive hyperspectral techniques for assessing heavy metal concentration (Cheng et al., 2019; Rathod et al., 2015). However, due to the weak optical properties and low concentration, the application of hyperspectral techniques for determining heavy metals in actual environmental samples, especially in water bodies, has been rarely reported (Niu et al., 2021).

Despite being limited, there are still studies that focus on exploring the feasibility of using hyperspectral remote sensing for monitoring heavy metals in water bodies, both in situ and in the laboratory. In the laboratory, the ratio method was employed to calculate the extinction coefficient and absorption coefficient of Cu ions, Fe ions, and Cd compounds, within a heavy metal concentration range of 3000–6000 mg/L. The results indicated that the absorption coefficient spectrum ranged from 400 nm to 900 nm, while the absorption peak of Cu ions was located at 810 nm (Liang et al., 2016a). Deng et al. (2016) revealed that Fe ions showed high absorption in the purple-blue light range, followed by green light. Liang et al. (2016b) demonstrated that cadmium sulfide exhibited a maximum value at 675 nm and minimum values at 550 nm and 830 nm, while cadmium oxide exhibited a nearly linear increase in reflectance from 525 nm to 900 nm. Huang et al. (2023a) used a stepwise multivariate linear regression method with short-wavelength infrared spectroscopy to retrieve Cu ion concentrations ranging from 100 to 1000 mg/L. The results identified the most sensitive wavelengths for Cu ions were at approximately 900 nm and 1080 nm. Laboratory experiments were conducted under specific conditions to obtain precise and controlled data. These controlled experiments aim to strengthen the theoretical framework in this field and optimize the utilization of hyperspectral remote sensing imagery for analyzing heavy metal concentrations.

For in-situ measurements, Chen et al. (2010) and Chen et al. (2012) presented the potential application of heavy metal inversion in coastal regions using in situ remote sensing techniques. The results demonstrated a strong correlation between Cu and Zn concentrations (ranging from 2 to 50 mg/L) at a wavelength of approximately 711 nm, based on the symbolic regression method and a correlative relationship. Liu et al. (2013) developed a piecewise algorithm to retrieve Zn concentrations in coastal waters, based on remote sensing reflectance and suspended sediment size, within a range of 5.67–86.62 mg/L. Liang et al. (2019) obtained spectral reflectance, extinction coefficient, scattering coefficient, and absorption coefficient for a typical heavy metal-contaminated water body, and a stable reflectance peak in the range of 600–700 nm was used as a feature band to distinguish the heavy metal-contaminated water body by visual interpretation within a range of 2.5–484 mg/L. Deng et al. (2013) developed a physical inversion model using the radiative transfer theory to retrieve concentrations of Fe and Cu ions in natural water. The model provided concentration estimates within the ranges of 0–526 mg/L and 0–4 mg/L respectively, utilizing HJ-1 A high-spectral images. Rajesh et al. (2020) applied recursive linear regression to establish the relationship between heavy metal concentration and remote sensing reflectance in rivers. Rostom et al. (2017) developed linear regression models to detect Cu concentrations ranging from 0.045 to 0.07 mg/L and Fe concentrations ranging from 0.005 to 0.2 mg/L using hyperspectral data in the visible and near-infrared (VNIR) range (350–1050 nm) from in-situ measurements of lake water samples in India. In-situ measurements are deemed feasible for estimating heavy metal concentrations, however, their sensitivity to varying environmental conditions restricts their applicability to regional or larger scales (Abd-Elrahman et al., 2011).

In addition to the aforementioned restrictions, the application scenarios have mostly focused on regions with high concentrations of heavy metals, specifically Cu and Fe ions exceeding 1 mg/L. In fact, the quantity of heavy metals in freshwater is relatively low, with Cu ions typically found at less than 0.01 mg/L and Fe ions at less than 0.1 mg/L, posing challenges for accurate detection using remote sensing imagery or in-situ spectral measurements. It is worth mentioning that Guo et al. (2022) revealed the possibility for low concentration heavy metal retrieval. They found that the lowest detectable concentration of $CuSO_4$ might be less than 0.15 mg/L within the wavelength range of 460.04–496 nm, whereas the lowest detectable concentration of CdS might be less than 0.001 mg/L within the ranges of 460.04–493.59 nm and 526.89–594.79 nm. However, the impact of interfering substances such as Chl-a, TOC, TN, and TP on heavy metal retrieval in water bodies was not considered, affecting the monitoring precision in heavy metal retrieval. In order to further investigate the applicability of hyperspectral remote sensing for low concentration heavy metal retrieval, this study aimed to develop an improved method for the efficient quantification of Cu and Fe concentrations in freshwater bodies. The main objectives of this study were: 1) to investigate the optimal features for Cu and Fe ions inversion; 2) to assess the performance of GA-PLSR algorithms in the retrieval of in-situ hyperspectral data; and 3) to clarify the interrelationships between Cu and Fe concentration and water quality parameters.

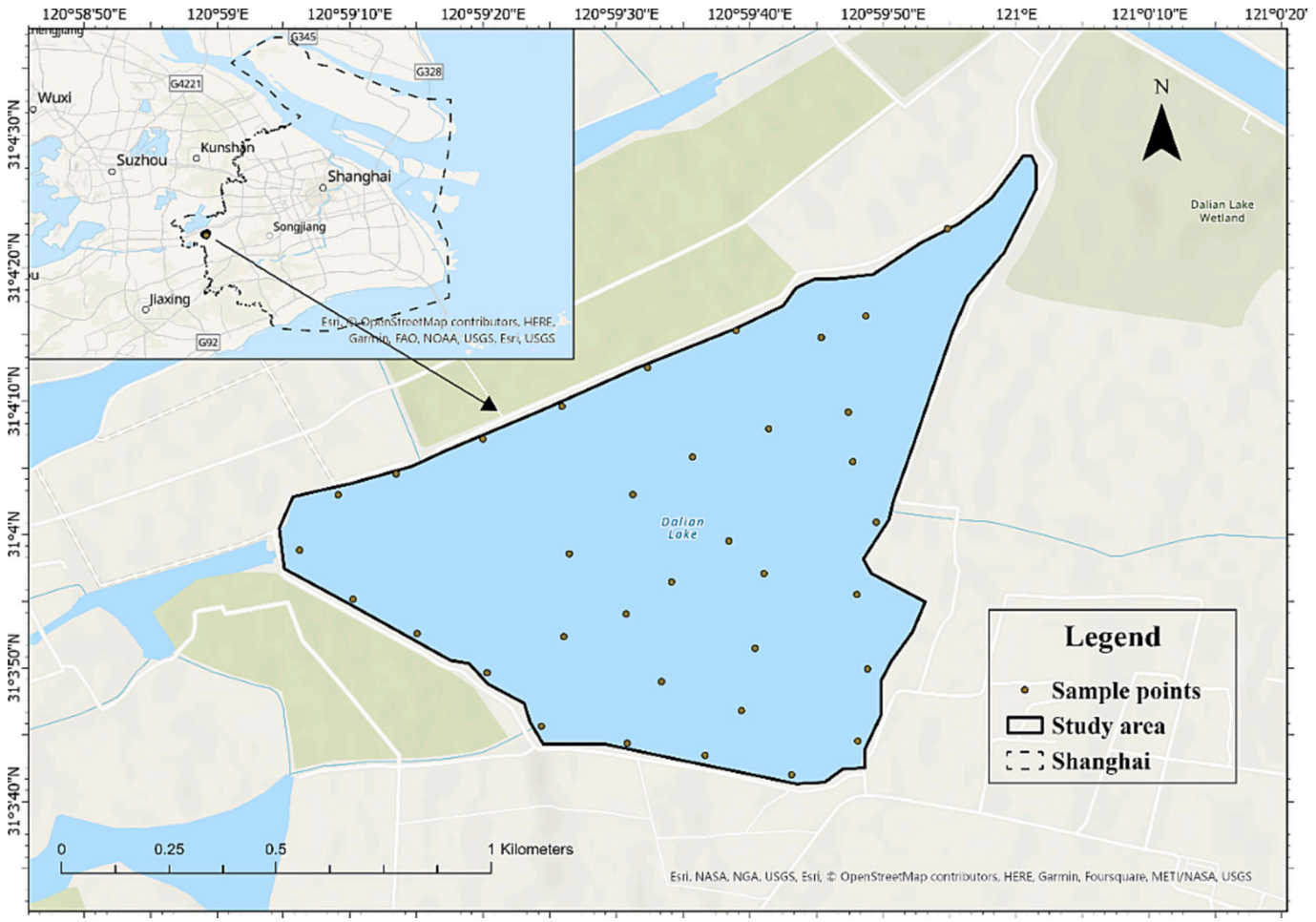


Fig. 1. Location of Dalian Lake and the distribution of sample sites.

2. Materials and methods

2.1. Study area

The Dalian Lake is located in the western part of Jinze Town, Qingpu District, Shanghai, within the Qingxi Suburban Park. It is connected to the south of Dianshan Lake through the Lanlu Port and ultimately to the Huangpu River (Liu et al., 2014). The total area of Dalian Lake is approximately 14.60 km^2, with a water area of about 1.0 km^2 (Yang et al., 2021). The average annual temperature of the lake is approximately 17.7 °C affected by subtropical monsoon climate and an annual precipitation is around 1050 mm, mostly concentrated from June to September (Shen et al., 2010). Dalian Lake is affected by tidal influences, which water level and volume are closely related to the inflow from the upstream and the tidal changes of the Huangpu River (Wang et al., 2010). The extensive farmlands surrounding the lake contribute to heavy metal pollution, due to the use of agricultural pesticides. Given its role as an important wetland ecosystem, long-term and continuous heavy metal monitoring is necessary.

2.2. Field spectral measurement data

Water sampling was conducted on April 21, 2023 in Dalian Lake. A total of thirty-five field water spectra and corresponding surface water samples (at a depth of 0–20 cm) were collected within the geographical coordinates of 31° 3' 41.31" - 31° 4' 28.40" N and 120° 59' 4.72" - 121° 0' 1.52" E. The sampling points were evenly distributed across the lake surface (shown in Fig. 1). In this study, high-resolution portable spectroradiometer SR-3500 (Spectral Evolution Inc., Lawrence, MA, USA) was used to measure the reflectance spectra of water samples on-site. The SR-3500 has a spectral range of 350–2500 nm, along with a spectral resolution of 3 nm between 350 and 1000 nm, 8 nm between 1000 and 1900 nm, and 6 nm between 1900 and 2500 nm. Additionally, all spectra were resampled to 1 nm. The water spectra were collected using an optical bare fiber probe at a measuring height of 10 cm and a field of view of 25°. To account for atmospheric variations, a white reference spectrum of a standard 99 % spectralon (99 % reflectance) was obtained prior to normalize the radiance spectrum of the samples. Each sample was measured five times and the resulting spectra were averaged to obtain a representative spectrum.

2.3. Chemical analysis for heavy metal and water quality

In the study, the water samples were obtained from different sites at a depth of −25 cm. Portable water quality analyzers and an EXO water quality analyzer were used to measure pH, temperature, dissolved oxygen (DO), oxidation-reduction potential (ORP), conductivity (Cond), and chlorophyll-a (Chl-a) directly at the sampling sites. The samples were then deposited in clean polyethylene containers and frozen promptly within a deep freezer (CNS-GB 12997-1991) for further laboratory chemical analysis. The concentrations of Cu and Fe were determined by Inductively Coupled Plasma-Optical Emission Spectroscopy (ICP-OES, Thermo Scientific, iCAP 7400). In addition to heavy metals, ammonia nitrogen (NH_3—N) was determined spectrophotometrically using salicylic acid (CNS- HJ 536-2009), total nitrogen (TN) was determined by Alkaline potassium persulfate digestion ultraviolet

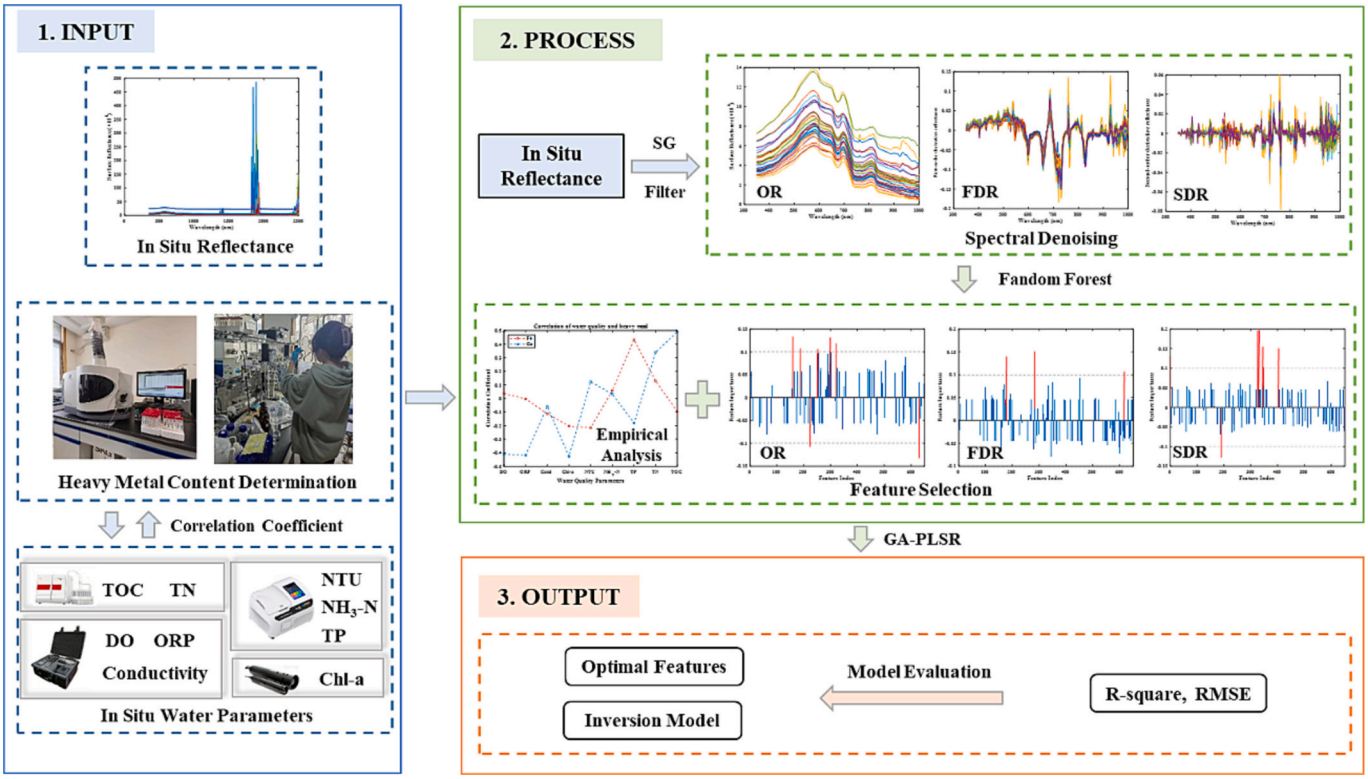


Fig. 2. Flowchart of Cu and Fe content estimation.

Table 1
Statistical measures of heavy metal concentrations and water quality in Dalian Lake in Shanghai, China.

Parameters	Minimum	Maximum	Mean	Median	Standard deviation
Cu (mg/L)	0.002	0.025	0.009	0.007	0.006
Fe (mg/L)	0.014	0.262	0.102	0.095	0.050
DO (mg/L)	7.47	9.56	8.58	8.53	0.43
ORP (mV)	116.8	210.6	146.8	151.5	20.7
Conductivity (μS/cm)	592	632	604	603	9
Chl-a (μg/L)	9.20	27.09	18.62	20.13	4.60
NTU	3	18	8	7	3
NH_3-N (mg/L)	0.11	0.52	0.30	0.27	0.10
TP (mg/L)	0.02	0.12	0.06	0.05	0.02
TN (mg/L)	0.40	1.91	0.95	0.89	0.37
TOC (mg/L)	1.56	9.19	4.61	4.61	1.25

spectrophotometric method (GB 11894-1989), total phosphorus (TP) was measured by Ammonium molybdate spectrophotometry method (GB 11893-89) and total organic carbon (TOC) was measured by non-dispersive infrared absorption method (HJ 501-2009).

2.4. Heavy metal concentration estimation

The objective of this study was to investigate the relationships between concentration of heavy metal (Cu and Fe) and hyperspectral data. This was achieved by using machine learning regression models to select the optimal features and predict the concentration of each heavy metal (Fig. 2). Firstly, the Savitzky-Golay (SG) smoothing filter was employed to reduce random noise in the in-situ field spectra. Secondly, the original spectrum (OR), first-order derivative reflectance (FDR), and second-order derivative reflectance (SDR) were utilized for feature selection by random forest (RF) respectively. Then the GA-PLSR model was employed to identify the optimal features and develop a prediction model for Cu and Fe retrieval. Finally, the relationship between the concentration of heavy metals and water quality parameters was analyzed, in order to provide an explanation for the selection of optimal features by GA-PLSR. All the statistical analyses for this study were conducted using MATLAB software (R2021a).

2.4.1. Spectral noise reduction

Since the reflectance beyond 1000 nm exhibited miscellaneous peaks, only the spectrum ranging from 350 to 1000 nm was selected due to its relatively high signal-to-noise ratio (Jupp et al., 1994). To improve the signal-to-noise ratio in the remaining spectral region, the widely used smoothing method, the Savitzky-Golay (SG) algorithm, was employed (Savitzky and Golay, 1964). According to the magnitude of spectral noise, the spectral regions were categorized into two categories. The spectral bands between 860 and 1000 nm were considerably noisy, while those between 350 and 860 nm were moderately noisy. A quadratic SG algorithm was employed with a smoothing window of 21 for the considerably noisy regions, and for the moderately noisy regions, a quadratic SG algorithm with a smoothing window of seven was utilized. Two abnormal spectra and one anomaly spectrum were excluded due to their reflectivity much higher than that of the normal water spectra. The 32 remaining field spectra were used to investigate the potential for estimating the concentration of heavy metals in water.

The SG-smoothed spectrum was taken as the original spectral reflectance (OR). The first-order and second-order derivative reflectance (FDR and SDR) spectra were then processed. Derivative spectroscopy is a fundamental technique that utilizes the differentiation and shape of spectrum for sharp peaks and absorptions (Holden and LeDrew, 1998; Rundquist et al., 1996; Zhou et al., 2021). In this study, the technique was used for eliminating background signals and resolving overlapping signals.

2.4.2. Feature selection

Feature selection is particularly crucial in hyperspectral image processing, since hundreds of bands will result in high redundancy and

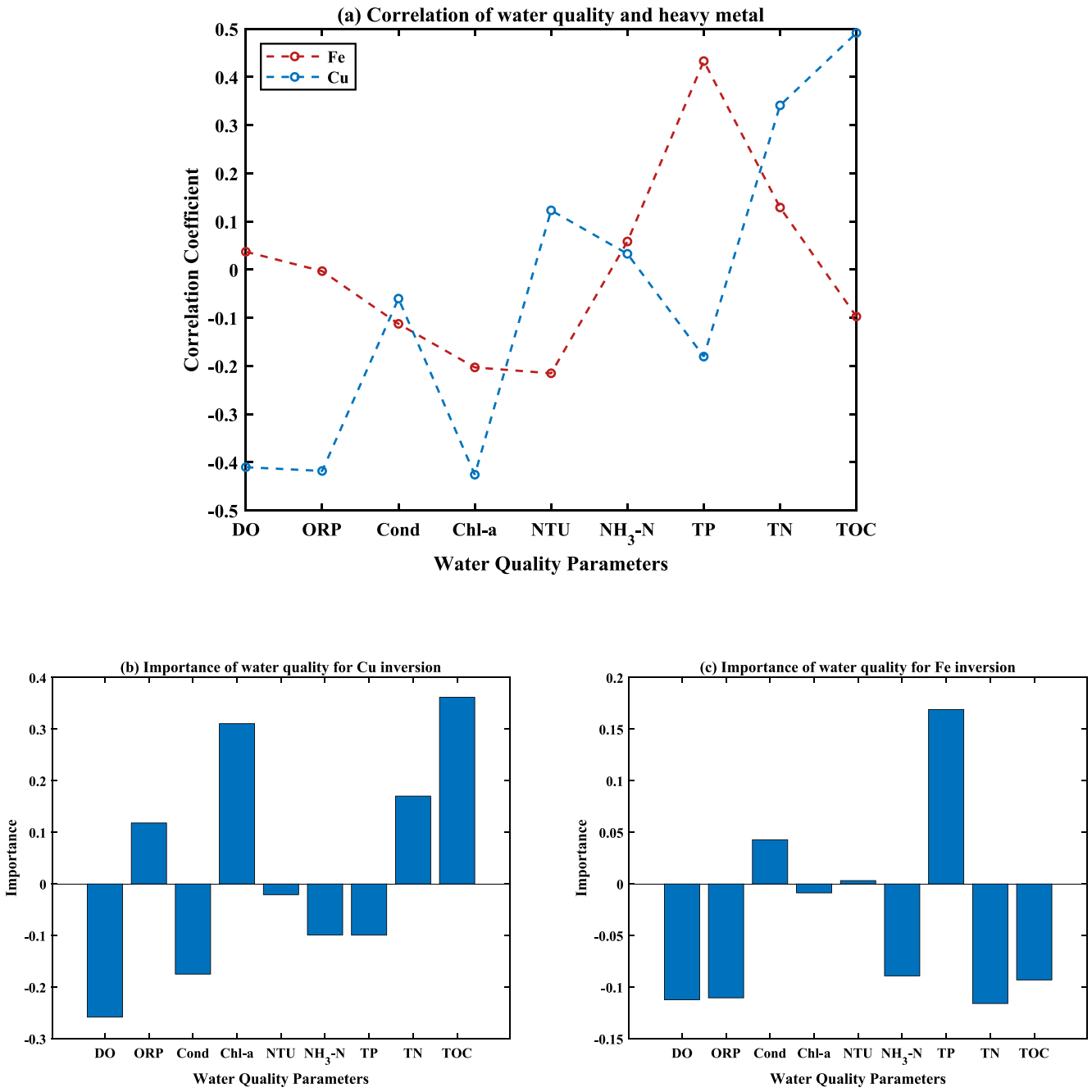


Fig. 3. Correlation between heavy metal and water quality parameters.

heavy computation. Therefore, it is necessary to identify the most sensitive spectral bands for estimating and mapping water heavy metal concentration accurately (Wang et al., 2018). In the study, the feature selection strategy integrated empirical spectral feature selection with algorithm-based feature selection.

Empirical spectral features were selected based on the theory of hyperspectral analysis for soil heavy metal prediction (Sun et al., 2023). To predict the concentration of Cu and Fe ions in water, we analyzed the water quality parameters that exhibited a strong correlation with Cu and Fe. Subsequently, we employed the feature bands of these highly correlated parameters, which had been previously identified in relevant studies, to estimate the concentration of heavy metals in water.

Algorithm-based feature selection utilized the random forest (RF) regression model to identify the most sensitive spectral bands. The RF builds multiple independent decision trees to make predictions using randomly sampling data and features (Breiman, 2001; Grömping, 2009; Liaw and Wiener, 2002). The variables of OR, FDR, SDR were utilized to establish separate RF regression models for extracting effective spectral information. Initially, 70 % of the samples were used as training data to evaluate the relative importance of variables (OR, FDR, SDR) in each trial. Subsequently, the selected features with an importance greater than 0.1 were combined as predictor variables. The predictor variables were considered optimal features if their respective R-square (R^2) values exceeded 0.7. We evaluated three datasets for 100 trials respectively and determined the optimized feature variables for subsequent models.

2.4.3. Calibration methods

A feature library was established through empirical spectral feature bands and feature variables created by RF. The integration of these different types of variables had the potential for effective monitoring of

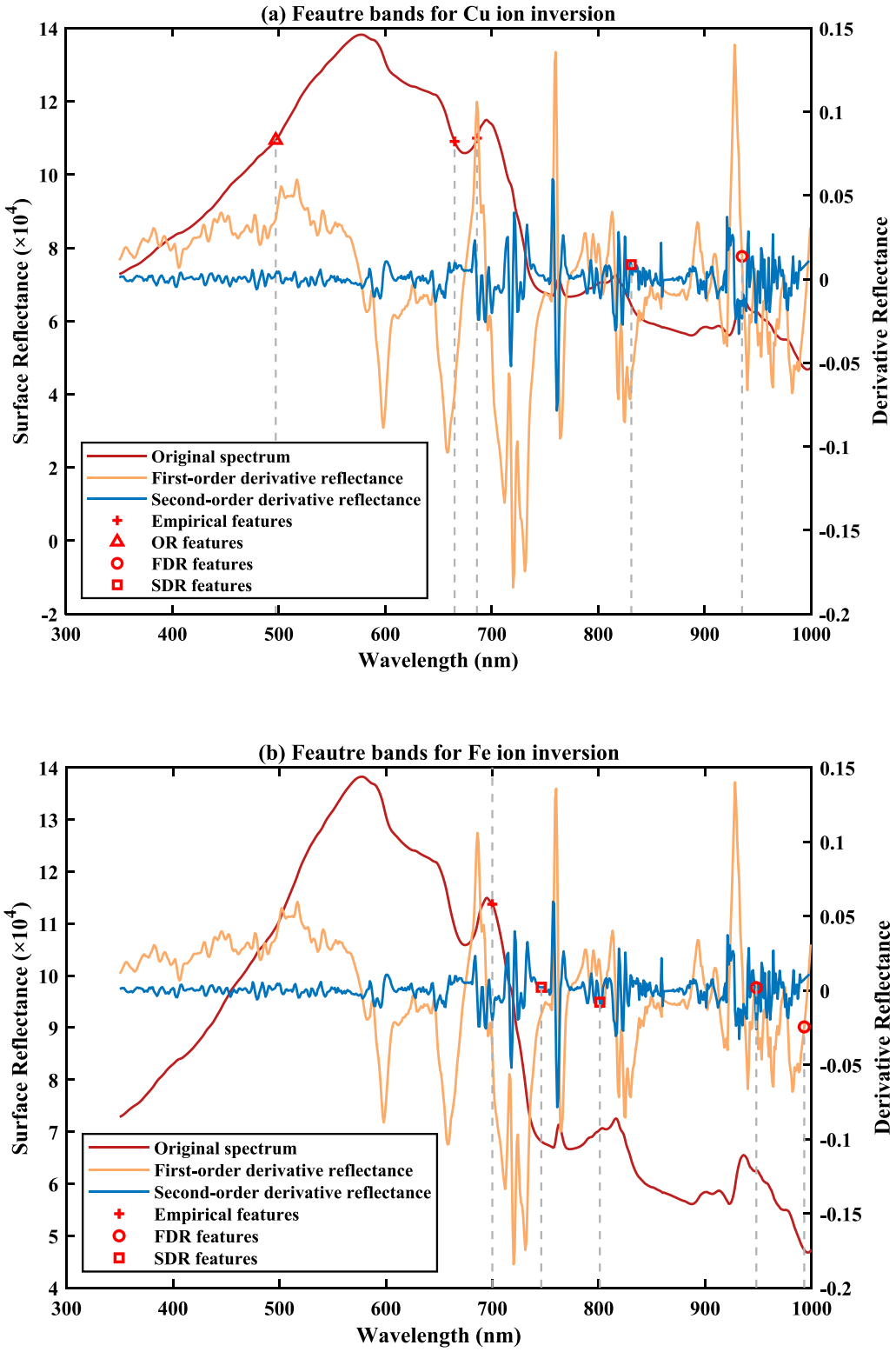


Fig. 4. The feature bands for estimation of Cu and Fe concentration from empirical features, OR features, FDR features and SDR features.

heavy metal concentration. To simplify the model and improve predictive performance, GA-PLSR was used for Cu and Fe inversion. The partial least squares regression (PLSR) is a statistical technique that is capable of addressing the issue of data collinearity, as well as situations when the number of variables significantly exceeds the sample size (Shi et al., 2014). The Genetic Algorithm (GA) is an effective feature optimization strategy for refining features and reducing model complexity (Sun et al., 2019). The integration of GA and PLSR has been proved to improve accuracy of estimating water properties (Song et al., 2013; Sudduth et al., 2015).

The initial iteration of GA consisted of randomly produced chromosomes, whereby each chromosome included binary-coded genes that determined the activation or deactivation of certain spectral bands. To minimize potential bias from the random initial generation in the esti-

Table 2
Regression results of RF, PLSR, GA-PLSR based on two feature selection methods.

Metal	Method	Algorithm-based feature selection			Empirical and algorithm-based feature selection		
		R^2	RMSE	MRE	R^2	RMSE	MRE
Cu	RF	0.45	0.004	0.437	0.53	0.004	0.424
	PLSR	0.50	0.004	0.456	0.57	0.004	0.436
	GA-PLSR	0.73	**0.003**	0.403	**0.75**	0.004	**0.382**
Fe	RF	0.39	0.041	0.565	0.47	0.045	0.563
	PLSR	0.41	0.051	0.554	0.54	0.037	0.466
	GA-PLSR	0.67	0.039	**0.464**	**0.73**	**0.036**	**0.464**

mation process, the GA-PLSR was executed 10 times. The prediction error sum of squares (PRESS) was employed to calculate the root-mean-square error of cross-validation (RMSECV), which served as a metric for choosing the most suitable number of components for PLSR. The equations were described as follows:

$$PRESS = \sum_{i=1}^{N} \left(y_i^{'} - y_i\right)^2$$

$$RMSECV = \sqrt{\frac{PRESS_k}{N}}$$

where $y_i^{'}$ represented the predicted value for the sample i, y_i represented measured concentration of the sample i, and k represented the number of components used in a PLSR model.

2.4.4. Model evaluation

The models were evaluated with the coefficient of determination for prediction (R^2), the root-mean-square error (RMSE), and the mean relative error (MRE), which were defined as

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n}}$$

$$MRE = \frac{\sum_{i=1}^{n}\frac{|y_i - \widehat{y_i}|}{y_i}}{n}$$

where n represented the number of samples, y_i represented the ith measured value, $\widehat{y}_i$ represented the ith predicted value, and $\overline{y}$ represented the average of the measured values. In general, a higher R^2 value combined with lower RMSE and MRE values indicates improved predictive performance.

3. Results

3.1. Heavy metal concentrations and water quality in Dalian Lake

According to the results presented in Table 1, the mean value of Cu content of the 35 water samples was found to be 0.009 mg/L, while the median value was determined to be 0.007 mg/L. Similarly, the mean value of Fe content of the 35 water samples was found to be 0.102 mg/L, while the median value was determined to be 0.095 mg/L. Based on the Chinese Environmental Quality Standards for Surface Water (GB3838–2002), the statistical analysis revealed that all heavy metal parameters met Class II levels. The statistical measures of mean and median also suggested a nearly symmetrical distribution of Cu and Fe concentrations. These findings indicated that the variables were independent and appropriate for regression analysis.

As for nutrient related parameters, the analysis of water quality revealed that TP ranged from Class II to Class V, with Class II-III standards accounting for 55.9 %, Class IV standards accounting for 41.2 %, and Class V standards accounting for only 2.9 %. The TN samples ranged from Class I to Class V, with Class III-IV standard accounting for 82.4 %. The NH_3-N samples ranged from Class I to Class III, with proportions of 5.9 %, 88.2 %, and 5.9 % respectively. Overall, the nutrient content in Dalian Lake was considered to be serious, resulting in a Class IV classification for the lake's water quality.

The Pearson correlation coefficients between water quality parameters and heavy metals were also investigated. Fig. 3 illustrated the correlation analysis for the seven water quality parameters. The results revealed a close relationship between TOC and Chl-a with Cu concentration in water. This relationship was demonstrated by the highest correlation coefficients of 0.49 and 0.43, respectively, as shown in Fig. 3 (a) and (b). Moreover, Fig. 3 (a) illustrated the close relationship between the concentration values of TP and Fe with correlation coefficients of 0.43. Furthermore, Fig. 3 (c) highlighted the feature importance values obtained from the RF model for assessing water quality parameters. The results from RF indicated that TOC and Chl-a were the most correlated parameters with Cu, while TP was correlated with Fe. The empirical feature bands of TOC and Chl-a in hyperspectral data were employed for Cu inversion, while the empirical feature bands of TP were utilized for Fe inversion.

3.2. Spectral features for Cu and Fe content inversion

The empirical spectral features and the RF regression model were used initially to identify the feature bands of Cu and Fe content (outlined in Section 2.4.2). This analysis confirmed that a total of 30 spectral variables including 8 empirical features, 11 OR, 9 FDR and 2 SDR were employed by GA-PLSR modeling for Cu inversion. For Fe inversion, GA-PLSR modeling utilized 25 spectral variables, including 3 empirical features, 9 OR, 10 FDR, and 3 SDR variables. The recommended number of latent variables (LVs) for both Cu and Fe inversion using GA-PLSR was determined to be five.

Fig. 4 presented the results of the feature bands. Specifically, the empirical features of 665 and 686 nm, the wavelengths of 497 nm in OR, 935 nm in FDR, 831 nm in SDR showed a significant correlation with Cu ion content (Fig. 4 (a)). This implied that the surface reflectance of 497, 665 and 686 nm, the slope of reflectance change in relation to wavelengths 935 nm, and the variation in slope concerning wavelength 831 nm could provide effective information for Cu ion inversion. Similarly, the empirical features of 700 nm, the wavelengths of 948 and 993 nm in FDR, 746 and 801 nm in SDR represented a relatively high correlation with Fe ion (Fig. 4 (b)). Hence, the surface reflectance of 700 nm, the slope of reflectance change in relation to wavelengths 948 and 993 nm and the variation in slope concerning wavelength 746 and 801 nm could provide effective information for Fe ion inversion.

3.3. Performance of GA-PLSR inversion model

The total of 32 experimental datasets were applied for calibration of modeling. These datasets were randomly divided into training and test sets, comprising 22 samples for training and 10 samples for testing, following a ratio of 2:1.

The RF, PLSR, and GA-PLSR models were employed to estimate the concentrations of the heavy metals Cu and Fe. Each model was repeated 10 times. Additionally, two feature selection methods were compared, namely the algorithm-based feature selection method and the combination of empirical and algorithm-based feature selection method. The performance of the three regression models based on two feature selection methods was evaluated using the mean values of R^2, RMSE, and MRE. The results for both heavy metals were presented in Table 2. In general, the combined approach of empirical analysis and algorithm-based analysis for feature selection achieved higher accuracy

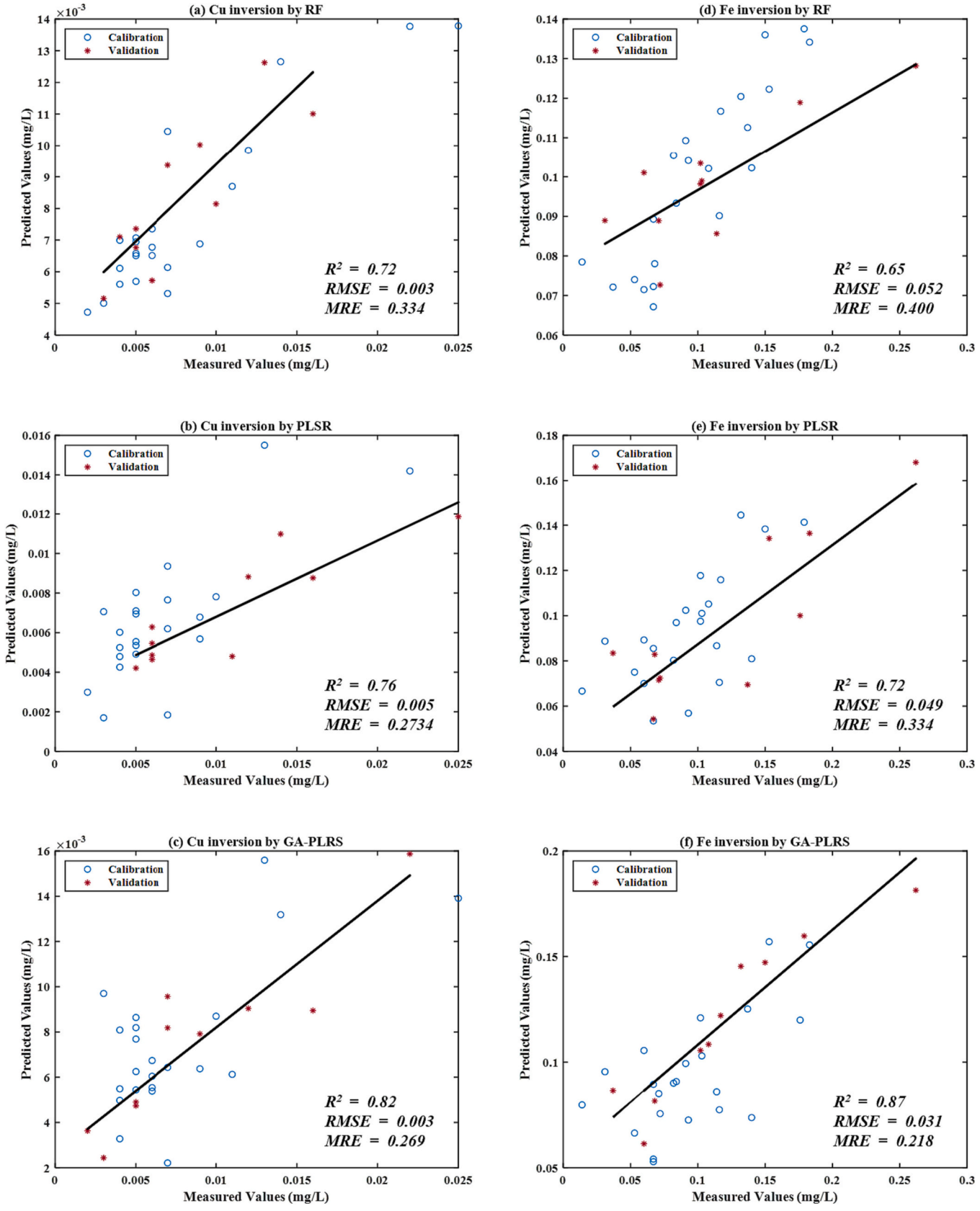


Fig. 5. Scatter plots of measured and predicted Cu and Fe ion contents of RF, PLSR and GA-PLSR models.

compared to algorithm-based feature selection method. Moreover, the RF model exhibited the lowest accuracy with the smallest R^2 values among all models. Although the PLSR model incorporating combined features demonstrated relatively good results for the parameter Cu, it failed to satisfy the accuracy requirements for both Cu and Fe inversion. On the other hand, the GA-PLSR algorithm achieved a high predictive accuracy for the inversion of both Cu and Fe. This was evident from the R^2 values greater than 0.7 and having the smallest values for both RMSE and MRE.

The scatter plots in Fig. 5 illustrated the relationship between the observed in situ heavy metal concentrations and the best-performed RF, PLSR, and GA-PLSR model predictions. All models were constructed using the optimal feature variables selected by empirical analysis and RF method. Based on the modeling results, it can be concluded that GA-PLSR had the highest R^2 value, lowest RMSE, and lowest MRE in terms of feature optimization strategy, outperforming models employing all bands. Overall, the GA-PLSR model offered several advantages. Firstly, it effectively reduced the number of input variables by selecting the most relevant wavelengths. Additionally, the GA-PLSR model demonstrated a significant improvement in regression accuracy compared to traditional machine learning approaches such as PLSR and RF models. Finally, GA-PLSR model demonstrated greater robustness and accuracy in solving complex regression problems with the value of RMSE and MRE of GA-PLSR model being smaller than PLSR and RF models.

4. Discussion

4.1. Feature selection: algorithm-based vs combination of empirical and algorithm-based approach

Compared to prediction based on algorithm-based feature selection method, the combination of empirical feature bands with algorithm-selected features improved prediction accuracy (Table 2). In terms of Cu inversion using GA-PLSR, the R^2 value showed an increase from 0.73 to 0.75. There was a slight increase in the RMSE value from 0.003 to 0.004, whereas the MRE value decreased from 0.403 to 0.382. For Fe inversion based on GA-PLSR, the R^2 value increased from 0.67 to 0.73. The RMSE value decreased from 0.039 to 0.036 and the MRE value remained unchanged. Due to the multicollinearity in hyperspectral data, there was a significant correlation between adjacent bands. Random forest had the potential to disregard certain highly correlated bands during the process of feature selection (Wei et al., 2021). Therefore, the incorporation of empirical features could supplement the feature library and subsequently improve the accuracy of heavy metal inversion. Moreover, the limited quantity of water samples (32 water samples) might influence the stability of random forest. To enhance the accuracy and reliability for heavy metal inversion, additional samples from diverse lakes were recommended in the future.

4.2. Reasons for improving the accuracy of Cu and Fe inversion

In this study, the empirical feature selection was developed considering empirical analysis of the relationship between Cu, Fe content and spectra. The results highlighted that the empirical wavelength of 665 and 686 nm combined with algorithm-selected features contained valuable information for predicting Cu content. Previous studies had demonstrated that TOC and Chl-a exhibited certain adsorption for Cu (Biswas et al., 2013; Fernandes and Henriques, 1991; Le et al., 2022; Martínez and McBride, 1999; Semeniuk et al., 2009; Tribovillard et al., 2008). Regarding TOC inversion in remote sensing, the linear combination of bands at 560, 665, 705 and 842 nm had the highest accuracy (Wang et al., 2022). For Chl-a inversion in remote sensing, reflectance peaks in green (~550 nm) and near-infrared (NIR, ~715 nm) and absorption peaks in blue (~433 nm) and red (~686 nm) spectral regions could help identify waters with varying concentrations of Chl-a (Xu et al., 2019). Therefore, the wavelengths of 433, 550, 560, 665, 686, 705, 715 and 842 nm were selected as empirical features bands to estimate Cu concentrations.

The results also demonstrated that the empirical wavelength of 700 nm combined with algorithm-selected features contained valuable information for predicting Fe content. Previous research had indicated that the formation of ferric oxide-bound phosphorus (Fe—P) may occur through the adsorption of dissolved P in the water column by Fe(III) (oxyhydr) oxides (Dan et al., 2020; Huang et al., 2023b; Yang et al., 2019). The reflectance at 680 nm, 700 nm, 769 nm, and near infrared bands was utilized to estimate TP content in waters at various depths (Abd-Elrahman et al., 2011). Therefore, the wavelength of 680, 700 and 769 nm were chosen as empirical features to estimate Fe concentrations in water.

5. Conclusion

Hyperspectral data is commonly used to estimate heavy metal content in water. However, given the low concentrations of heavy metals involved, accurately assessing such content can be challenging. To address the above issues, this study investigated the feasibility of determining heavy metal content using hyperspectral reflectance. Firstly, a combination of empirical analysis and algorithm-based analysis was utilized to select the most important spectral variables responsible for Cu and Fe ions. During the accuracy evaluation, it was observed that incorporating feature bands associated with TOC and Chl-a could improve estimation accuracy of Cu concentration. Similarly, the use of feature bands linked to TP could estimate Fe concentration effectively. Additionally, three classic machine learning models, namely RF, PLSR, and GA-PLSR, were compared. Among the three methods tested, GA-PLSR exhibited the highest accuracy, followed by PLSR and RF. In summary, this study presented a valuable approach for determining the concentration of heavy metals in low concentration regions by utilizing hyperspectral reflectance. This approach can assist government agencies in preventing and controlling inland water pollution. In future studies, the hyperspectral satellite and UAVs will be used for measuring heavy metals in water, thereby allowing for the creation of spatial distribution maps.

CRediT authorship contribution statement

Yukun Lin & Jiaxin Gao: Writing – review & editing, Writing – original draft, Software, Project administration, Methodology, Conceptualization. **Yaojen Tu:** Writing – review & editing, Conceptualization. **Yuxun Zhang:** Writing – review & editing, Resources. **Jun Gao:** Writing – review & editing, Resources, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, CAS, and Strategic Priority Research Program of the Chinese Academy of Sciences [grant number 41730642, SKLLQGZR2304, and XDB40020105].

References

Abd-Elrahman, A., Croxton, M., Pande-Chettri, R., Toor, G.S., Smith, S., Hill, J., 2011. In situ estimation of water quality parameters in freshwater aquaculture ponds using hyperspectral imaging system. ISPRS J. Photogramm. Remote Sens. 66, 463–472. https://doi.org/10.1016/j.isprsjprs.2011.02.005.

Alengebawy, A., Abdelkhalek, S.T., Qureshi, S.R., Wang, M.-Q., 2021. Heavy metals and pesticides toxicity in agricultural soil and plants: ecological risks and human health implications. Toxics 9. https://doi.org/10.3390/toxics9030042.

Azizullah, A., Taimur, N., Khan, S., Häder, D.-P., 2021. In: Häder, D.-P., Helbling, E.W., Villafañe, V.E. (Eds.), Heavy Metals Pollution in Surface Waters of Pakistan BT - Anthropogenic Pollution of Aquatic Ecosystems. Springer International Publishing, Cham, pp. 271–312. https://doi.org/10.1007/978-3-030-75602-4_13.

Bashir, I., Lone, F.A., Bhat, R.A., Mir, S.A., Dar, Z.A., Dar, S.A., 2020. Concerns and threats of contamination on aquatic ecosystems. Bioremediation Biotechnol. Sustain. Approaches to Pollut. Degrad. https://doi.org/10.1007/978-3-030-35691-0_1.

Besser, J.M., Leib, K.J., 2007. Toxicity of Metals in Water and Sediment to Aquatic Biota. Integr. Investig. Environ. Eff. Hist. Min, Animas River Watershed, San Juan County, Color, p. 14.

Biswas, H., Bandyopadhyay, D., Waite, A., 2013. Copper addition helps alleviate iron stress in a coastal diatom: response of Chaetoceros gracilis from the Bay of Bengal to experimental Cu and Fe addition. Mar. Chem. 157, 224–232. https://doi.org/10.1016/j.marchem.2013.10.006.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cai, J., Chen, J., Dou, X., Xing, Q., 2022. Using machine learning algorithms with in situ hyperspectral reflectance data to assess comprehensive water quality of urban Rivers. IEEE Trans. Geosci. Remote Sens. 60, 1–13. https://doi.org/10.1109/TGRS.2022.3147695.

Chen, C., Liu, F., He, Q., Shi, H., 2010. The Possibility on Estimation of Concentration of Heavy, pp. 4216–4219.

Chen, C., Liu, F., Tang, S., 2012. Estimation of Heavy Metal Concentration in the Pearl River Estuarine Waters from Remote Sensing Data, pp. 2575–2578.

Cheng, H., Shen, R., Chen, Y., Wan, Q., Shi, T., Wang, J., Wan, Y., Hong, Y., Li, X., 2019. Estimating heavy metal concentrations in suburban soils with reflectance spectroscopy. Geoderma 336, 59–67. https://doi.org/10.1016/j.geoderma.2018.08.010.

Chi, G., Ma, J., Shi, Y., Chen, X., 2016. Hyperspectral remote sensing of cyanobacterial pigments as indicators of the iron nutritional status of cyanobacteria-dominant algal blooms in eutrophic lakes. Ecol. Indic. 71, 609–617. https://doi.org/10.1016/j.ecolind.2016.06.014.

Dan, S.F., Lan, W., Yang, B., Han, L., Xu, C., Lu, D., Kang, Z., Huang, H., Ning, Z., 2020. Bulk sedimentary phosphorus in relation to organic carbon, sediment textural properties and hydrodynamics in the northern Beibu Gulf, South China Sea. Mar. Pollut. Bull. 155, 111176.

Deng, R., Wu, Y., Qin, Y., 2013. A Research of Remote Sensing Inversion for Heavy Metal cu and Fe in Nature Water Using HJ-1A HSI Data —— A Case Study of Beijiang Basin in Guangdong Province 227–235.

Deng, R.R., Liang, Y.H., Gao, Y.K., Qin, Y., Liu, X., 2016. Measuring absorption coefficient spectrum (400–900 nm) of hydrated and complex ferric ion in water. J. Remote Sens 20, 35–44.

Fernandes, J.C., Henriques, F.S., 1991. Biochemical, physiological, and structural effects of excess copper in plants. Bot. Rev. 57, 246–273.

Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. Am. Stat. 63, 308–319.

Guo, Y., Liang, Y., Deng, R., Li, J., Wang, J., Hua, Z., Tang, Y., 2022. Development and application of a new sensitivity analysis model for the remote sensing retrieval of heavy metals in water. Heliyon 8, e12033. https://doi.org/10.1016/j.heliyon.2022.e12033.

Holden, H., LeDrew, E., 1998. Spectral discrimination of healthy and non-healthy corals based on cluster analysis, principal components analysis, and derivative spectroscopy. Remote Sens. Environ. 65, 217–224.

Huang, C., Chen, X.-Y., Lee, M., 2023a. An improved hyperspectral sensing approach for the rapid determination of copper ion concentrations in water environment using short-wavelength infrared spectroscopy. Environ. Pollut. 333, 121984 https://doi.org/10.1016/j.envpol.2023.121984.

Huang, H., Dan, S.F., Yang, B., Ning, Z., Liang, S., Kang, Z., Lu, D., Zhou, J., Huang, Hu, 2023b. Spatiotemporal distributions of poorly-bound heavy metals in surface sediments of a typical subtropical eutrophic estuary and adjacent bay. Mar. Environ. Res. 189, 106076 https://doi.org/10.1016/j.marenvres.2023.106076.

Jupp, D.L.B., Kirk, J.T.O., Harris, G.P., 1994. Detection, identification and mapping of cyanobacteria—using remote sensing to measure the optical quality of turbid inland waters. Mar. Freshw. Res. 45, 801–828.

Le, T.T.N., Le, V.T., Dao, M.U., Nguyen, Q.V., Vu, T.T., Nguyen, M.H., Tran, D.L., Le, H. S., 2019. Preparation of magnetic graphene oxide/chitosan composite beads for effective removal of heavy metals and dyes from aqueous solutions. Chem. Eng. Commun. 206, 1337–1352. https://doi.org/10.1080/00986445.2018.1558215.

Le, T.P.Q., Le, N.D., Hoang, T.T.H., Rochelle-Newall, E., Nguyen, T.A.H., Dinh, L.M., Duong, T.T., Pham, T.M.H., Nguyen, T.D., Phung, T.X.B., Nguyen, T.Q.T., Vu, T.H., Le, P.T., Phung, V.P., 2022. Surface sediment quality of the Red River (Vietnam): impacted by anthropogenic and natural factors. Int. J. Environ. Sci. Technol. 19, 12477–12496. https://doi.org/10.1007/s13762-022-03936-z.

Liang, Y.H., Deng, R.R., Gao, Y.K., Qin, Y., Liu, X.L., 2016a. Measuring absorption coefficient spectrum (400–900 nm) of copper ions in water. J. Remote Sens 20, 27–34.

Liang, Y.H., Deng, R.R., Liu, Y.M., Lin, L., Qin, Y., He, Y.Q., 2016b. Measuring the spectrum of extinction coefficient and reflectance for cadmium compounds from 400 to 900 nm. Guang pu xue yu Guang pu fen xi= Guang pu 36, 4006–4012.

Liang, Y., Deng, R., Huang, J., Xiong, L., Qin, Y., Liu, Z., 2019. The spectral characteristic analysis of typical heavy metal polluted water——a case study of mine drainage in Dabaoshan mountain, Guangdong province, China. Spectrosc. Spectr. Anal. 39, 3237–3244.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2, 18–22.

Liu, F., Tang, S., Chen, C., 2013. Estimation of particulate zinc using MERIS data of the Pearl River estuary. Remote Sens. Lett. 4, 813–821. https://doi.org/10.1080/2150704X.2013.798711.

Liu, X., Wu, Z., Xu, H., Zhu, H., Wang, X., Liu, Z., 2014. Assessment of pollution status of Dalianhu water sources in Shanghai, China and its pollution biological characteristics. Environ. Earth Sci. 71, 4543–4552. https://doi.org/10.1007/s12665-013-2846-5.

Martínez, C.E., McBride, M.B., 1999. Dissolved and labile concentrations of cd, cu, Pb, and Zn in aged ferrihydrite− organic matter systems. Environ. Sci. Technol. 33, 745–750.

Mondal, B., Bordoloi, N., 2023. Chapter 5 - metal in water: an assessment of toxicity with its biogeochemistry. In: Shukla, S.K., Kumar, S., Madhav, S., Mishra, P.K.B.T.-M.W. (Eds.), Advances in Environmental Pollution Research. Elsevier, pp. 71–91. https://doi.org/10.1016/B978-0-323-95919-3.00017-3.

Niu, C., Tan, K., Jia, X., Wang, X., 2021. Deep learning based regression for optically inactive inland water quality parameter estimation using airborne hyperspectral imagery. Environ. Pollut. 286, 117534 https://doi.org/10.1016/j.envpol.2021.117534.

Pandey, S., Kumari, N., 2023. Chapter 8 - impact assessment of heavy metal pollution in surface water bodies. In: Shukla, S.K., Kumar, S., Madhav, S., Mishra, P.K.B.T.-M.W. (Eds.), Advances in Environmental Pollution Research. Elsevier, pp. 129–154. https://doi.org/10.1016/B978-0-323-95919-3.00004-5.

Pinto, M.M.S.C., Marinho-Reis, P., Almeida, A., Pinto, E., Neves, O., Inácio, M., Gerardo, B., Freitas, S., Simões, M.R., Dinis, P.A., Diniz, L., da Silva, E.F., Moreira, P. I., 2019. Links between cognitive status and trace element levels in hair for an environmentally exposed population: a case study in the surroundings of the Estarreja industrial area. Int. J. Environ. Res. Public Health 16. https://doi.org/10.3390/ijerph16224560.

Rajesh, A., Jiji, G.W., Raj, J.D., 2020. Estimating the pollution level based on heavy metal concentration in water bodies of Tiruppur District. J. Indian Soc. Remote Sens. 48, 47–57. https://doi.org/10.1007/s12524-019-01058-7.

Rathod, P.H., Brackhage, C., Van der Meer, F.D., Müller, I., Noomen, M.F., Rossiter, D.G., Dudel, G.E., 2015. Spectral changes in the leaves of barley plant due to phytoremediation of metals—results from a pot study. Eur. J. Remote Sens. 48, 283–302. https://doi.org/10.5721/EuJRS20154816.

Rostom, N.G., Shalaby, A.A., Issa, Y.M., Afifi, A.A., 2017. Evaluation of Mariut Lake water quality using hyperspectral remote sensing and laboratory works. Egypt. J. Remote Sens. Sp. Sci. 20, S39–S48.

Rundquist, D.C., Han, L., Schalles, J.F., Peake, J.S., 1996. Remote measurement of algal chlorophyll in surface waters: the case for the first derivative of reflectance near 690 nm. Photogramm. Eng. Remote. Sens. 62, 195–200.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639.

Semeniuk, D.M., Cullen, J.T., Johnson, W.K., Gagnon, K., Ruth, T.J., Maldonado, M.T., 2009. Plankton copper requirements and uptake in the subarctic Northeast Pacific Ocean. Deep Sea Res Part I Oceanogr. Res. Pap. 56, 1130–1142.

Shen, G., Wang, Z., Qian, X., Zhu, Y., Zhang, H., 2010. The features of agricultural nonpoint source pollution in the Dalian Lake area of Shanghai. Acta Agric. Shanghai 26, 55–59.

Shi, T., Chen, Y., Liu, Y., Wu, G., 2014. Visible and near-infrared reflectance spectroscopy-an alternative for monitoring soil contamination by heavy metals. J. Hazard. Mater. 265, 166–176. https://doi.org/10.1016/j.jhazmat.2013.11.059.

Song, K., Li, L., Tedesco, L.P., Li, S., Duan, H., Liu, D., Hall, B.E., Du, J., Li, Z., Shi, K., 2013. Remote estimation of chlorophyll-a in turbid inland waters: three-band model versus GA-PLS model. Remote Sens. Environ. 136, 342–357.

Sudduth, K.A., Jang, G., Lerch, R.N., Sadler, E.J., 2015. Long-term agroecosystem research in the Central Mississippi River basin: hyperspectral remote sensing of reservoir water quality. J. Environ. Qual. 44, 71–83.

Sun, W., Skidmore, A.K., Wang, T., Zhang, X., 2019. Heavy metal pollution at mine sites estimated from reflectance spectroscopy following correction for skewed data. Environ. Pollut. 252, 1117–1124. https://doi.org/10.1016/j.envpol.2019.06.021.

Sun, W., Liu, S., Wang, M., Zhang, X., Shang, K., Liu, Q., 2023. Soil copper concentration map in mining area generated from AHSI remote sensing imagery. Sci. Total Environ. 860, 160511 https://doi.org/10.1016/j.scitotenv.2022.160511.

Tribovillard, N., Bout-Roumazeilles, V., Algeo, T., Lyons, T.W., Sionneau, T., Montero-Serrano, J.C., Riboulleau, A., Baudin, F., 2008. Paleodepositional conditions in the Orca Basin as inferred from organic matter and trace metal contents. Mar. Geol. 254, 62–72.

Wang, X., Yang, W., 2019. Water quality monitoring and evaluation using remote sensing techniques in China: a systematic review. Ecosyst. Heal. Sustain. 5, 47–56. https://doi.org/10.1080/20964129.2019.1571443.

Wang, Z.Q., Shen, G.X., Qian, X.Y., Zhu, J., Zhu, Y., 2010. Seasonal effect of agricultural non-point source pollution on water environment of Dianshan Lake Basin in Shanghai city. J. Anhui Agric. Sci. 38, 20227–20229.

Wang, F., Gao, J., Zha, Y., 2018. Hyperspectral sensing of heavy metals in soil and vegetation: feasibility and challenges. ISPRS J. Photogramm. Remote Sens. 136, 73–84. https://doi.org/10.1016/j.isprsjprs.2017.12.003.

Wang, S., Shen, M., Liu, W., Ma, Y., Shi, H., Zhang, J., Liu, D., 2022. Developing remote sensing methods for monitoring water quality of alpine rivers on the Tibetan plateau. GIScience Remote Sens. 59, 1384–1405. https://doi.org/10.1080/15481603.2022.2116078.

Wei, H.-E., Grafton, M., Bretherton, M., Irwin, M., Sandoval, E., 2021. Evaluation of point hyperspectral reflectance and multivariate regression models for grapevine water status estimation. Remote Sens. https://doi.org/10.3390/rs13163198.

Xu, M., Liu, H., Beck, R., Lekki, J., Yang, B., Shu, S., Liu, Y., Benko, T., Anderson, R., Tokars, R., Johansen, R., Emery, E., Reif, M., 2019. Regionally and locally adaptive models for retrieving chlorophyll-a concentration in inland waters from remotely sensed multispectral and hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. 57, 4758–4774. https://doi.org/10.1109/TGRS.2019.2892899.

Yang, B., Lan, R.-Z., Lu, D.-L., Dan, S.F., Kang, Z.-J., Jiang, Q.-C., Lan, W.-L., Zhong, Q.-P., 2019. Phosphorus biogeochemical cycling in intertidal surface sediments from the Maowei Sea in the northern Beibu gulf. Reg. Stud. Mar. Sci. 28, 100624.

Yang, Z.H.A., Difang, W., Chengjin, C.A.O., Minsheng HUANG, X.W., Bowen, Y.U., Chang, L.I.U., Haochen, D.U., Mengzhuo, L.I., 2021. Investigation of the environmental status of water at the Dalian Lake demonstration area in the Jinze water source area of Taipu River. J. East China Norm. Univ. (Natural Sci.) 2021, 64.

Zhou, Q., Yang, N., Li, Y., Ren, B., Ding, X., Bian, H., Yao, X., 2020. Total concentrations and sources of heavy metal pollution in global river and lake water bodies from 1972 to 2017. Glob. Ecol. Conserv. 22, e00925 https://doi.org/10.1016/j.gecco.2020.e00925.

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. https://doi.org/10.1016/j.catena.2021.105222.