Environmental Science and Ecotechnology 25 (2025) 100570

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Machine learning reveals distinct aquatic organic matter patterns driven by soil erosion types

Yingxin Shang [a], Kaishan Song [a, *], Zhidan Wen [a], Fengfa Lai [a], Ge Liu [a], Hui Tao [a], Xiangfei Yu [b]

[a] Northeast Institute of Geography and Agroecology, CAS, Changchun, 130102, China
[b] Jilin Jianzhu University, Changchun, 130118, China

ARTICLE INFO

Article history:
Received 17 June 2024
Received in revised form
29 April 2025
Accepted 2 May 2025

Keywords:
CDOM
Remote sensing
FT ICRMS
Soil erosion
Lake

ABSTRACT

Chromophoric dissolved organic matter (CDOM), characterized by unique optical properties, is an essential indicator for understanding aquatic organic matter dynamics within global carbon cycles. Soil erosion, a major source of CDOM received by lakes, transports terrestrial organic matter to water bodies, altering sources, bioavailability and molecular complexity of CDOM significantly. Yet, the spatial patterns of CDOM in lakes from different soil erosion regions are still unknown. Here, we developed a robust machine learning framework ($RMSE_{calibration}$ = 0.87 m^{-1}) to estimate CDOM concentrations in lakes by integrating over 1300 *in situ* water samples with Landsat 8 OLI surface reflectance data. We then applied this model to map the spatial distribution of CDOM across lakes larger than 0.1 km^2 in 2020. Our analysis revealed distinct spatial patterns, with mean CDOM absorption coefficients at 355 nm of 3.73 m^{-1} in freeze-thaw erosion regions, 6.31 m^{-1} in wind erosion regions, and 3.72 m^{-1} in hydraulic erosion regions, reflecting significant variations driven by erosion intensity. Two axes of PCA analysis explained over 48% variations of CDOM for different soil erosion types. Chemical characterization indicated that polycyclic aromatic predominated in wind and hydraulic erosion regions, whereas freeze-thaw erosion regions exhibited higher proportions of peptides and unsaturated aliphatic compounds. This study highlights the crucial connection between terrestrial soil erosion processes and aquatic DOM composition, providing vital insights for evaluating global carbon cycling and carbon storage within inland ecosystems.

1. Introduction

Climatic changes and social development related to the economy and environmental protection affect the carbon balance of aquatic systems [1]. Dissolved organic matter (DOM) in aquatic environments is mainly derived from terrestrial and aquatic microorganisms, which provide energy for food webs and absorb ultraviolet radiation [2,3]. Chromophoric dissolved organic matter (CDOM) is an important watercolor parameter for evaluating the content of aquatic DOM. Allochthonous organic matter from terrestrial sources is a significant part of aquatic DOM [4,5] and is generally influenced by regional watershed attributes, such as land use, soil erosion, and climatic conditions. The transport of organic carbon from soils to lakes and, ultimately, to oceans plays an important role in the terrestrial carbon cycle.

With the rapid development of China, soil erosion has become a significant problem. Natural and anthropogenic factors, including precipitation and seasonal water flows, affect soil and water loss. Different watershed slopes contribute to the terrestrial carbon pool transformation into aquatic DOM in waterways through varying flow rates and soil leaching times. As soil erosion is influenced by precipitation, catchment runoff, land use, and land cover types [6], the transport of mobile sediment in overland flow often carries high rates of DOM and nutrients from terrestrial areas to bodies of water [7,8]. However, few studies have mainly focused on the individual effects of land use on soil erosion or terrestrial sources of DOM in watersheds. Special absorption parameter at 254 nm ($SUVA_{254}$) is calculated with the CDOM absorbance at 254 nm and dissolved organic carbon concentration (DOC), which indicates the aromaticity of CDOM [9,10]. Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) is a powerful technique for

* Corresponding author.
E-mail address: songkaishan@iga.ac.cn (K. Song).

https://doi.org/10.1016/j.ese.2025.100570

detecting the molecular structure of DOM that could help explain the sources of DOM in the biogeochemical processes of aquatic systems and terrestrial cycles. The soil type, erosion rate, and precipitation conditions can somewhat affect the CDOM sources [11]. The variations in sources and the composition of DOM in aquatic systems can be evaluated using the aromaticity index (AI), molecular lability index (MLB), and other molecular parameters to detect the degradation theory of aquatic DOM. The large-scale area of China has uneven climate conditions, causing variations in the types and degrees of soil erosion that affect DOM import into receiving waterways. Understanding the spatial variations in CDOM in different erosion regions is essential to illustrate a more detailed terrestrial path for DOM transport to various areas. This understanding will increase the accuracy of assessing carbon cycle processes for carbon neutrality in aquatic and terrestrial ecosystems.

CDOM, one of the three watercolor parameters, offers significant advantages when remote sensing techniques are used to explore spatial distribution. This approach saves substantial manpower, material, and financial resources compared with traditional *in situ* field trips. Although empirical or semi-analytical algorithms for CDOM have been widely developed for regional lakes [12,13], general CDOM models for large-scale areas of China remain sparse. Due to the complex optical properties in different regions, developing empirical or semi-analytical algorithms for large-scale areas is challenging. Machine learning algorithms, including eXtreme gradient boosting (XGBoost) and support vector machine (SVM), can help establish CDOM algorithms without requiring an optical type or regional classification [14,15]. XGBoost represents an efficient and scalable instantiation of the gradient boosting framework [16]. This algorithm constitutes an optimized, distributed gradient boosting library, and it incorporates the feature sampling approach used in random forests to compute variable importance and increase model efficiency. We believe that using remote sensing retrieval of CDOM combined with the Google Earth Engine (GEE) platform to explore the CDOM distribution across different erosion regions can elucidate the quantity, quality, and related alteration processes of DOM imports. This broad perspective will enrich our understanding of global carbon cycles.

Given the lack of a general CDOM algorithm for large-scale regions in China and the limited studies on the linkage between CDOM and soil erosion, this study aims to address these gaps despite many works emphasizing the important role of terrestrial sources for CDOM. We seek to establish a large-scale machine learning model for the CDOM absorption coefficient with high accuracy for Chinese lakes in different soil erosion regions and to evaluate the spatial trends and potential factors influencing CDOM content concerning soil erosion types and intensity. The objectives of the study are to (1) develop and validate a uniform and simple model for CDOM retrieval in lakes across China; (2) map CDOM parameters in lakes (>0.1 km^2) to identify spatial patterns using Landsat 8 Operational Land Imager (OLI) data across different soil erosion types and degrees in China; (3) identify the potential factors contributing to CDOM content in different soil erosion regions; and (4) discuss the relationships between CDOM content and terrestrial soil organic matter (SOM) to propose the potential linkage for terrestrial organic matter transport to receiving waterways.

2. Methods and materials

2.1. Field trips and laboratory analysis in the study regions

The study area encompasses the entire catchment in China. Variations in soil and water losses across these catchments are influenced by differing soil properties resulting from various lithology and land use types driven by human activities. The first-level division of soil erosion in China is categorized into three types: hydraulic erosion (northeast monsoon area), wind erosion (northwest arid area), and freeze−thaw erosion (Qingzang Plateau). This classification aligns with the basic division of natural geography in China. According to the Soil Erosion Classification and Grading Standard (SL 190-96), soil erosion is classified into three degrees (weak, moderate, and severe) across different erosion regions. The degree classification and the corresponding standards for soil erosion levels are shown in Supplementary Materials Table S1.

In total, 1337 water samples from 255 lakes distributed across different soil erosion regions in China were collected from 2015 to 2021. The samples were stored in polyethylene bottles at 4 °C during field trips. Each sample was filtered on the night of collection to ensure the dataset's accuracy. The filtered samples were transported to the laboratory for CDOM parameter analysis within two days. A global positioning system (GPS) receiver recorded the sampling locations (Supplementary Material Table S2). The CDOM absorption coefficient from 200 to 800 nm was measured with an ultraviolet−visible spectrophotometer (Shimadzu UV-2600PC) after filtering through 0.22-μm polycarbonate membrane filters. The CDOM absorption coefficient at 355 nm ($a_{CDOM}(355)$) was used to represent the CDOM concentration in this study. $SUVA_{254}$ (CDOM coefficient at 254 nm divided by DOC) was used to evaluate the aromaticity of CDOM [9].

2.2. Image Acquisition and model calibration

The *in situ* datasets in the ice-free period for the CDOM absorption coefficient ($a_{CDOM}(355)$) were used to match up the surface reflectance product of the cloud-free Landsat 8 OLI images with a time window of ±7 days (the effect of cloud contamination <8%) from the GEE platform. The Landsat 8 OLI images downloaded on the GEE platform have been widely used for monitoring CDOM and other water quality parameters for inland waters and have been proven to perform well [17]. The surface reflectance product conducted the process of atmospheric correction with the Land Surface Reflectance Code (LaSRC) software. The band collected by each image on the GEE platform was divided by 10,000, and the value was divided by π to obtain the remote sensing reflectance ($Rrs(\lambda)$) for constructing the CDOM model. The original lake boundary datasets were obtained from the HydroLAKES dataset, and the disturbance of the nearshore region was attenuated by reducing the area by 3 pixels [18]. The normalized difference water index (NDWI) was used to extract pixels for the lakes. The mean reflectance of a 3 × 3 pixel area around the GPS coordinates of the sampling sites was calculated to establish the models [19]. In total, 1337 samples from 255 lakes were selected to calibrate ($n = 928$) and validate ($n = 409$) the model performance. Four machine learning algorithms, including SVM, gradient boosting decision tree (GBDT), XGBoost, and random forest (RF), were used to calibrate the CDOM absorption models. Comprehensive descriptions of these models can be found in the Supplementary Materials (S1). The XGBoost algorithm is an optimized distributed gradient boosting library based on second-order Taylor series expansion. It incorporates the feature sampling method of random forests to calculate variable importance, thereby increasing model diversity, reducing overfitting, and improving model efficiency. Therefore, we chose the XGBoost model as the final calibrated model. The models were developed using packages in R software. Finally, we used the optimal model to retrieve the CDOM coefficient in lakes (>0.1 km^2) across China using the Landsat 8 OLI surface reflectance product embedded on the GEE platform to demonstrate spatial CDOM variation in different soil erosion regions across China.

2.3. Datasets for natural and anthropogenic factors

Land use data (http://www.globallandcover.com/Chinese/GLC30Download/index.aspx), climate conditions (precipitation, temperature, and wind speed) (http://data.cma.cn/), and social data (http://data.cnki.net/Yearbook) were used to evaluate the factors contributing to CDOM sources and quantity. These datasets were assembled with mapped lakes using the third catchment boundary. The soil carbon density datasets were taken from a previous study (https://doi.org/10.11922/sciencedb.603). The spatial distribution data of soil erosion in China are based on the industry standard SL 190-96, "Soil Erosion Classification and Grading Standard" of the People's Republic of China. According to the soil erosion types and intensity classification, three degrees (weak, moderate, and severe) in different soil erosion regions were classified. The degree classification and the corresponding standards for soil erosion levels are shown in Supplementary Materials Table S1.

2.4. FT-ICR MS analysis for molecular composition

The lake samples were collected and filtered through a 0.45 μm Whatman glass fiber filter membrane during the field trips. The samples were further filtered in the laboratory through a 0.2-μm pre-cleaned polycarbonate filter membrane and extracted using solid-phase extraction (Agilent Bond Elut PPL) [20]. The molecular compounds of DOM were analyzed using a negative ion Apollo II electrospray ionization with an Ultra FT-ICR mass spectrometer. Detailed extraction and data processing methods can be found in a previous study [10]. C, H, O, N, S, and P represent the numbers of carbon, hydrogen, oxygen, nitrogen, sulfur, and phosphorus atoms in all the DOM formulas, respectively. A modified AI (calculated by (1+C−0.5O−S−0.5H)/(C−0.5O−S−N−P)) and MLB (calculated by dividing the number of formulas with H/C ≥ 1.5 by the total number of formulas) were calculated [21]. The percentages of the formulas (CHO, CHNO, CHOS, and CHNOS) were determined. The compounds were classified into polycyclic aromatics, highly aromatic (HA) compounds, highly unsaturated (HU) compounds, unsaturated aliphatic (UA) compounds, saturated compounds, and peptides [22–24]. Detailed information on the FT-ICR MS analysis is provided in the Supplementary Materials S3.

2.5. Statistical analysis

The determination coefficient (R^2) and root mean square error (*RMSE*) were used to assess model performance between measured and predicted $a_{CDOM}(355)$ [25]. The difference analysis was carried out using the SPSS package (version 16.0, Chicago, Illinois, USA), with a significance level of $p < 0.05$. The spatial distribution of the sampling sites and the mapping results were produced using ArcGIS 10.1 (Environmental Systems Research Institute, CA). Principal component analysis (PCA) was conducted using Origin 9.0 (Microcal Software, Inc., MA, USA) based on the CDOM coefficient and its potential factors in different soil erosion regions.

3. Results

3.1. Model calibration and validation across China

We used 1349 samples collected from field trips and found a high variation in $a_{CDOM}(355)$ (0.038–15.991 m^{-1}) across China, matching the Rrs of the Landsat 8 images. Subsequently, we calibrated and validated universal models for CDOM absorption retrieval. Preliminary correlation analyses of $a_{CDOM}(355)$ with all possible single-band and band combinations of Landsat 8 images were conducted to select the most sensitive spectral variables ($R^2 > 0.50$) based on 1337 pairs of matched data. Considering the correlation coefficients, we identified the most sensitive spectral variables as inputs for estimating $a_{CDOM}(355)$ (green, red, blue, near infrared (NIR), shortwave infrared 1 (SWIR1), and shortwave infrared 2 (SWIR2). To develop a remote sensing model of $a_{CDOM}(355)$ with good performance and universality, the statistical metrics of four machine learning algorithms, namely SVM, GBDT, XGBoost, and RF, were computed to calibrate the models for the $a_{CDOM}(355)$ estimates.

Among these models, RF and SVM exhibited poor performance, with low determination coefficients (R^2 = 0.49–0.62, *RMSE* = 1.39–1.96 m^{-1}). The XGBoost model, displaying a good fitting performance (slope close to 1), outperformed the GBDT model (R^2 = 0.87, *slope* = 0.73) (Fig. 1). Similar results were observed for validating the models, with XGBoost demonstrating good performance (Fig. 2). Moreover, the XGBoost model achieved the highest accuracy with the lowest *RMSE* ($RMSE_{XGB,cal}$ = 0.87 m^{-1}, $RMSE_{XGB,val}$ = 0.94 m^{-1}). Thus, a machine learning model based on several bands (blue, green, red, NIR, SWIR1, and SWIR2) of Landsat reflectance provides a reliable estimate of CDOM content, and the universal remote sensing CDOM model offers a reasonable method for estimating CDOM in lakes across China.

3.2. CDOM distribution across China

We mapped the spatial characteristics of CDOM for over 20,000 lakes during the 2020 non-freezing period using the calibrated XGBoost model mentioned in Section 3.1 with cloud-free Landsat 8 images. We generated an averaged CDOM distribution map for lakes across China with a water area >0.1 km^2 on the GEE platform. Regarding CDOM content, the mean $a_{CDOM}(355)$ of lakes across China was 3.68 m^{-1}. Among these lakes, 6832 lakes had a CDOM concentration range of 3–4 m^{-1}, and the other 4530 lakes had a CDOM range of 2–3 m^{-1} (Fig. 3). The mean CDOM of the lakes

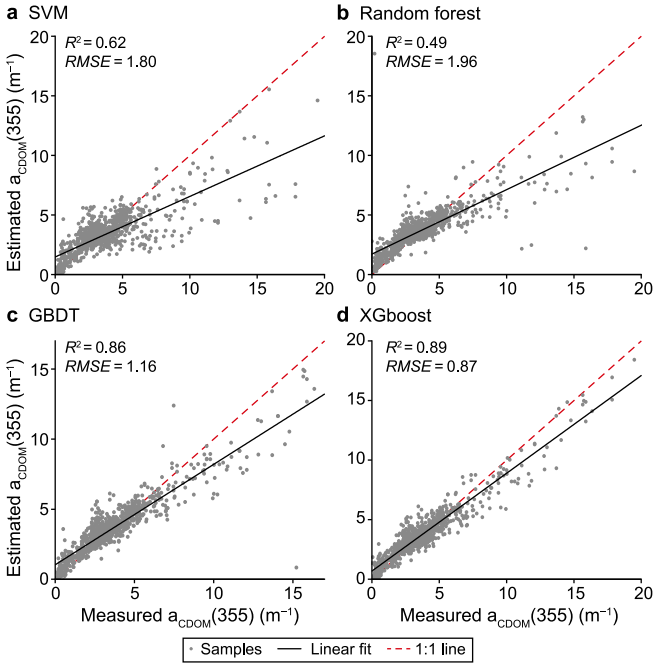

Fig. 1. The calibration results for $a_{CDOM}(355)$ (n = 928). **a**, Support vector machine (SVM); **b**, Random forest; **c**, Gradient boosting decision tree (GBDT); **d**, eXtreme Gradient Boosting (XGBoost). The grey circles represent the samples for calibration, the black line represents the linear fit lines, and the red line represents the 1:1 goodness-fit-lines.

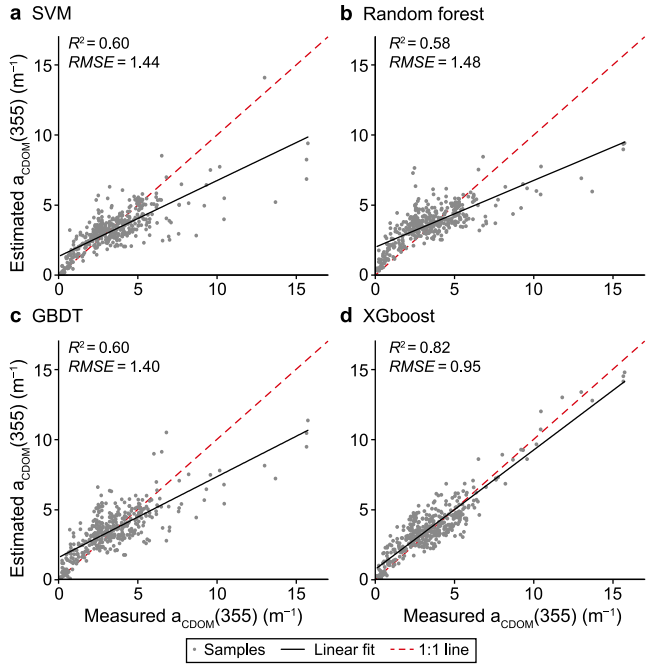

Fig. 2. The validation results for $a_{CDOM}(355)$ (n = 409). **a**, Support vector machine (SVM); **b**, Random forest; **c**, Gradient boosting decision tree (GBDT); **d**, eXtreme Gradient Boosting (XGBoost). The grey circles represent the samples for calibration, the black line represents the linear fit lines, and the red line represents the 1:1 goodness-fit-lines.

varied significantly, especially across different soil erosion regions. The mean $a_{CDOM}(355)$ values for freeze–thaw erosion regions, wind erosion regions, and hydraulic erosion regions were 3.73, 6.31, and 3.72 m^{-1}, respectively. Significant variations were observed among different erosion regions based on the erosion intensity ($p < 0.05$). For the freeze–thaw erosion regions, as the erosion intensity increased, the CDOM absorption significantly increased, while an inverse relationship was observed for wind erosion and hydraulic erosion ($p < 0.05$) (Fig. 4).

3.3. CDOM variation factors based on erosion types

Natural and anthropogenic factors contribute to variations in CDOM concentration. In this study, the water area of lakes

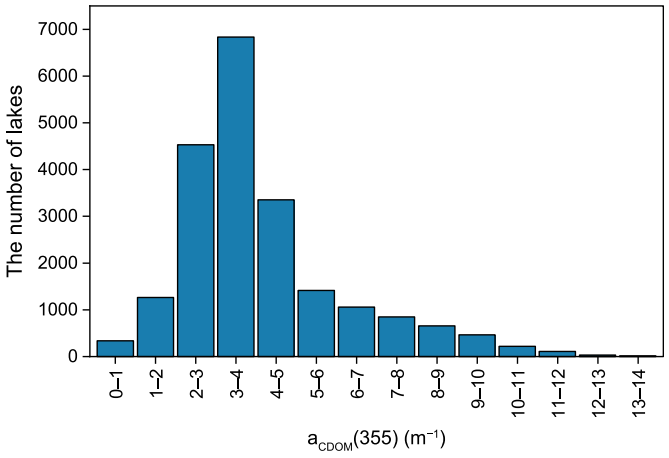

Fig. 3. The number of lakes based on the range of $a_{CDOM}(355)$ across China.

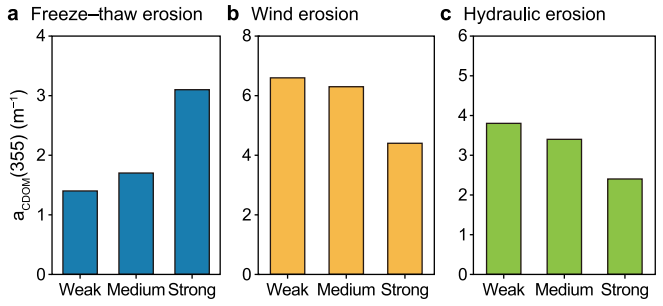

Fig. 4. The comparison of mean $a_{CDOM}(355)$ based on erosion types and intensities. **a**, Freeze–thaw erosion; **b**, Wind erosion; **c**, Hydraulic erosion.

increased, while the $a_{CDOM}(355)$ of lakes significantly decreased. Similar trends were also observed for different types of erosion (Fig. 5). The mean $a_{CDOM}(355)$ of lakes in the wind erosion region was consistently the highest among the different water area levels. Regarding human activities, this study considered that $a_{CDOM}(355)$ was not only related to the natural attributes of lakes but also influenced by the intensity of human activity and level of social development. This is supported by the finding that the $a_{CDOM}(355)$ of lakes located in the western part of Hu's line was significantly higher than that of lakes located east of Hu's line for three erosion regions across China (Fig. 6). The PCA (Fig. 7) showed that $a_{CDOM}(355)$ for wind erosion was positively correlated with wind speed. By contrast, $a_{CDOM}(355)$ for the freeze–thaw erosion region was mostly affected by land use. Moreover, the $a_{CDOM}(355)$ in the hydraulic erosion region was comprehensively influenced by multiple factors, including natural precipitation/temperature, human activities (population), and land use types.

3.4. Molecular analysis of CDOM variations by FT-ICR MS

Regarding the molecular formulas of CDOM in lakes across different erosion regions based on FT-ICR mass spectra, most DOM samples were dominated by CHO-containing and CHNO-containing formulas. High relative contents of CHOS- and CHNOS-containing compounds were observed in CDOM from hydraulic erosion regions and wind erosion regions. HU compounds and UA compounds dominated the DOM composition in the lakes. Polycyclic aromatic and aromatic compounds were more abundant in DOM from the wind erosion and hydraulic erosion regions than in the freeze–thaw erosion regions. Conversely, lakes in the freeze–thaw

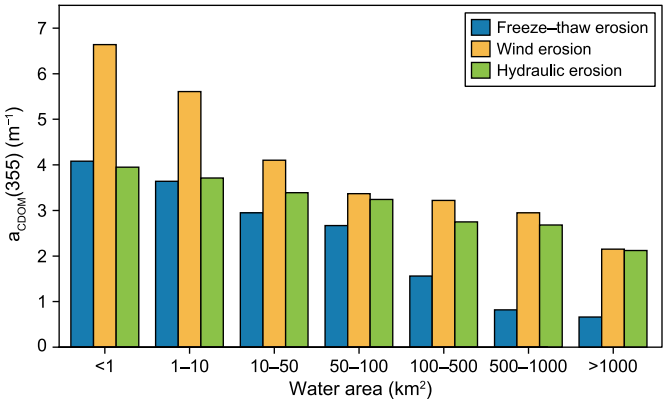

Fig. 5. The trends of chromophoric dissolved organic matter (CDOM) absorption with water area for soil erosion regions.

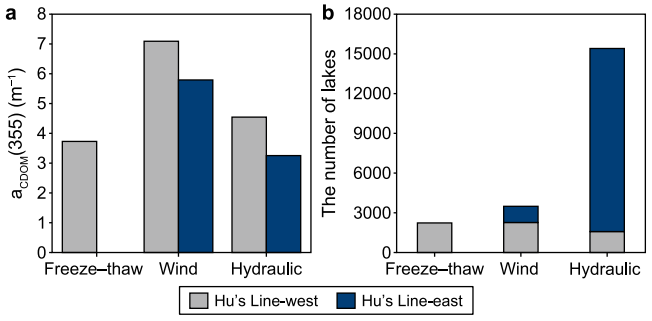

Fig. 6. The trends of chromophoric dissolved organic matter (CDOM) absorption with population for soil erosion regions. **a**, The comparison of $a_{CDOM}(355)$; **b**, The comparison of lake numbers.

erosion regions exhibited relatively high levels of peptide and UA compounds. The richness of MLB in DOM from lakes in the freeze−thaw erosion regions was significantly higher than in other areas, while a contrasting trend was observed in AI (Fig. 8).

4. Discussion

4.1. Model uncertainties

Sentinel-2A MSI and Landsat-8 OLI are among the latest satellite sensors that can be applied for inland watercolor remote sensing. More empirical algorithms have been applied to Landsat-8 images to observe CDOM absorption based on band ratio or machine learning methodologies compared with the CDOM application of Sentinel-2 for inland waters [26−28]. The Landsat series has many image products (e.g., TM, ETM+, and OLI) for a longer period. Using the Landsat series would be more suitable for longer-term CDOM inversion. The first step in developing a general inversion algorithm was to determine the reflectance bands used in the algorithms. The reflectance bands for blue, green, red, NIR, SWIR1, and SWIR2 were sensitive to the CDOM coefficients, as mentioned in Section 3. The strong absorption in the blue region of the CDOM spectrum contributed to setting the blue region reflectance as the denominator for machine learning algorithms [29]. Other bands, such as green and red, have also been used to propose inversion models for different complex waters [30,31]. In this study, empirical single-band models or band ratio models of CDOM absorption coefficients were applied with poor performance. Thus, four machine-learning algorithms were explored to retrieve CDOM concentrations. Compared with analytical and semi-analytical algorithms, empirical inversion methods have the advantages of simplicity and high efficiency, as they aim to determine straightforwardly a function that best reflects the relationships between CDOM parameters and reflectance information based on *in situ* samples [32]. The XGBoost model (R^2 = 0.89) outperformed the empirical models and other machine learning algorithms. According to Prairie [33], with a validation number of 409, $t_{0.05}$ is 1.96. The *RMSE* of XGBoost was 0.94, indicating that the precision of the predicted CDOM coefficient for a 95% confidence level satisfied the CDOM estimation to some extent. The comparison between measured $a_{CDOM}(355)$ and predicted $a_{CDOM}(355)$ showed that the predicted $a_{CDOM}(355)$ was higher than the measured $a_{CDOM}(355)$ at the corresponding levels of soil erosion (Fig. 9). Specifically, high levels of CDOM content could be underestimated to some extent, even though the training data for calibration had a large-scale distribution across China to improve the portability of the CDOM absorption coefficient algorithm. The XGBoost model demonstrated effectiveness across various CDOM properties, presenting strengths and weaknesses. Notably, its performance remains largely unaffected by disturbances arising from aerosol conditions [15]. Considering the co-variances and a gradient descent on the residual errors in the XGBoost model improves its ability to generalize through a regularization design [34].

Previous studies have reported using bands 1−7 of Landsat 8 as input elements of the algorithms and $a_{CDOM}(254)$ as the output element to retrieve CDOM [35]. In addition, although reflectance at longer wavelengths is generally insensitive to CDOM, practical validation has shown that algorithm performance can be improved using additional longer wavelengths [17,36]. We tested and compared the model's accuracy of the blue, green, and red bands to that of the model using the blue, green, red, NIR, SWIR1, and SWIR2 bands. We found that the model with the SWIR1 and SWIR2 bands exhibited a more stable performance and that the validation results also showed good performance with the SWIR1 and SWIR2 bands. However, SWIR bands with strongly absorbing aerosols could lead to uncertainties in remote sensing reflectance [37,38], which should be considered in further studies.

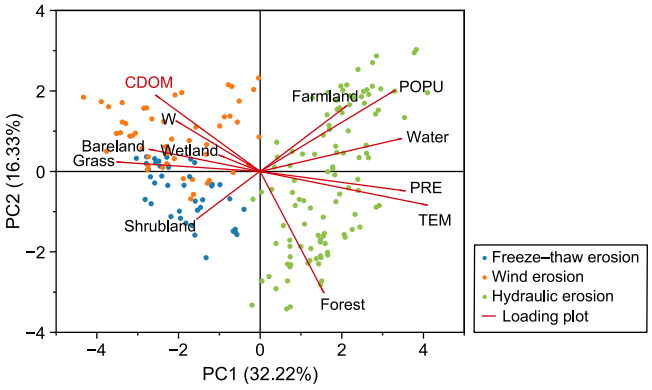

Fig. 7. The PCA analysis for chromophoric dissolved organic matter (CDOM) in different regions. W: wind; POPU: population; PRE: precipitation; TEM: temperature. The shrubland, grass, bareland, water, and farmland represent the proportion of land use types individually. The circles with different colors represent samples in different erosion regions. The red line represents the loading of each parameter on the principal component (PC) analysis axis.

4.2. Natural and anthropogenic contributions from a macro perspective

Human land use and climatic conditions have significantly changed CDOM sources in aquatic ecosystems [39]. In the hydraulic erosion regions, PCA showed that temperature and precipitation positively influenced soil-derived CDOM concentration and microbial humic−CDOM in waters. Previous studies have shown that these effects of agricultural land use can be explained by the preferential mobilization of topsoil DOM from agricultural watersheds [39]. Southern China is the main hydraulic erosion region with a large population, and human activities significantly affect DOM import. Other factors, such as agricultural activities and forest areas, often cause soil erosion, leading to soil particles with DOM ending up in neighboring streams through high runoff [40]. It has been widely observed that agricultural soils may experience more severe erosion than forest soils because of increased exposure to open-air oxidation and DOM export [41].

In regions affected by wind erosion, DOM concentration significantly correlated with wind speed. However, we observed that the mean wind speeds ranged from 1.3 m s^{-1} in areas with weak wind erosion to 0.77 m s^{-1} in regions experiencing strong wind erosion. Thus, we noted a decrease in CDOM concentration from areas with

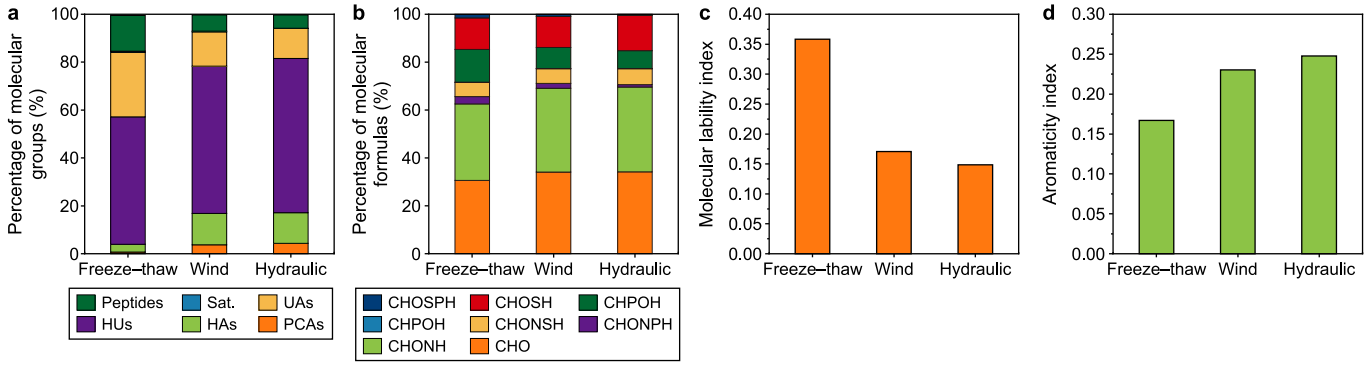

Fig. 8. The molecular distribution of aquatic dissolved organic matter (DOM) in different soil erosion regions. **a**, The percentage of molecular groups; **b**, The percentage of molecular formulas; **c**, Molecular lability index; **d**, Aromaticity index. PCAs: polycyclic aromatics; HAs: highly aromatic compounds; HUs: highly unsaturated compounds; UAs: unsaturated aliphatic compounds; Sat.: saturated compounds.

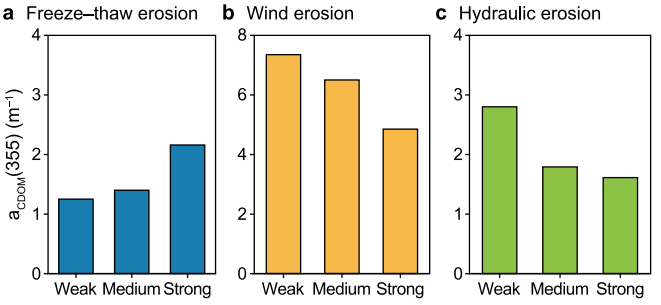

Fig. 9. The comparison of measured $a_{CDOM}(355)$ and predicted $a_{CDOM}(355)$ in different regions. **a**, Freeze–thaw erosion; **b**, Wind erosion; **c**, Hydraulic erosion.

weak wind erosion to those with strong wind erosion, suggesting the efficacy of soil erosion protection measures. In regions affected by freeze–thaw erosion, the presence of grass and shrubland could influence terrestrial CDOM concentration to some degree. In addition, the reduction in solar irradiation levels exceeding 7 kWh m^{-2} day^{-1} for freeze–thaw erosion, less than 5 kWh m^{-2} day^{-1} for wind erosion, and below 3 kWh m^{-2} day^{-1} for hydraulic erosion could impede photochemical degradation and lead to a decrease in aromaticity and the humic component ratio, as evidenced by SUVA (Supplementary Material Fig. S1). The $SUVA_{254}$ of DOM demonstrated varying trends depending on soil erosion intensity for the same types, highlighting disparities in human activities and climate conditions across different erosion regions. In summary, CDOM appears to be significantly influenced by the interplay between land use and climatic conditions in various erosion regions [42].

4.3. Effect of soil erosion on CDOM distribution from a molecular perspective

The effects of human activities, land use, and soil erosion, along with the runoff effect, accelerate the loss of loose particles on the surface and cause numerous water quality issues [1]. Soil erosion, in the form of transported suspended sediment, is often associated with organic matter and nutrients from terrestrial areas being carried to bodies of water [7]. In the freeze–thaw erosion regions, the lowest contents of polycyclic aromatics and AI showed that strong solar radiation in this region triggers strong photochemical activities, which decrease aromatic compounds through sufficient photodegradation [43]. Prolonged water residence times in lakes may enhance the oxidation process of HU compounds. Conversely, the enhanced frequency and intensity of freeze–thaw cycles influence dissolved soil carbon through the burst of cells in soil microorganisms in terrestrial ecosystems, increasing the transformation of organic matter and nutrients into water bodies via runoff flows [44]. In the case of wind erosion, the high contents of polycyclic aromatics and AI can be attributed to the terrigenous input of DOM in lakes. The high relative content of saturated compounds in DOM in this region may be due to the heavy eutrophic state of the lakes [45]. Moreover, wind erosion disturbs water areas, causing organic matter on both sides of the lake banks to flush into the water bodies. The fluctuation of water bodies with wind disturbance brings more terrestrial DOM into lakes, thus affecting the high aromatic levels for terrestrial organic matter from a macro perspective and the content of black carbon and polyphenols at the molecular level [46].

Compared with wind erosion, the highest AI and PACs and the high levels of HAs in hydraulic erosion in South China indicate the presence of many aromatic compounds. The highest content of CHNO-containing compounds suggests that the source of DOM is the degradation product of vascular plants [26]. In addition, sulfur-containing compounds indicate that there are sources of organic sulfur from human activities or pollution [47]. This result coincides with the highest content of terrestrial-derived organic matter in the hydraulic erosion region. Terrestrial sources of DOM in soil or other ground components are transported to waters through precipitation, surface runoff, and underground runoff. For weak hydraulic erosion, increasing precipitation and runoff contribute to terrestrial sources, while significant soil sediments entering waters cause sediment resuspension [48,49]. Therefore, modifications in the DOM compositions in lakes seem to be due to surface soil erosion and aquatic production [50].

4.4. Implications of CDOM in terrestrial soil organic carbon

The transfer of organic carbon from soils to lakes and rivers plays a vital role in the overall aquatic and terrestrial ecosystem carbon cycle [50]. DOM constitutes only a small fraction of SOM (<0.25%), but its mobile and active characteristics within the soil contribute to its high bio-availability. Significant relationships between aquatic CDOM and terrestrial soil organic carbon (SOC) density were established through meta-analysis and *in situ* measurements, with an *R*-value of 0.65 (Fig. 10), indicating that the enrichment of SOM affects the high concentration of DOM in aquatic systems. Soil carbon cycling and nutrient cycling are influenced by the diversity of soil microorganisms and spatially variable distribution and rainfall intensity [1]. These significant relationships are crucial for adjusting factors, such as rainfall

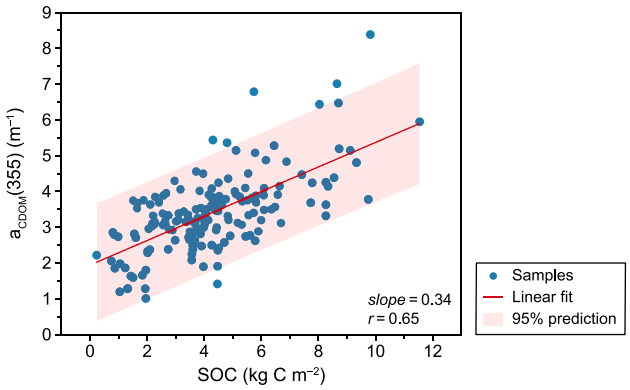

Fig. 10. The relationship between terrestrial soil organic carbon (SOC) and aquatic chromophoric dissolved organic matter (CDOM) The blue circle represents the average of $a_{CDOM}(355)$ and SOC in each basin. The red line is the linear fit line, and the red shadow band represents the 95% prediction interval.

erosion and soil erodibility, to generate soil loss maps and assess water management levels. The detailed transfer mechanisms from land to water bodies remain unknown despite the establishment of general relationships between soil DOM and aquatic DOM. Moreover, the role of soil microorganisms in biogeochemical cycling, transforming non-available nutrient elements into available forms and subsequently increasing aquatic nutrient content, warrants further investigation. Because DOM is integral to the global carbon cycle, understanding the linkage between aquatic DOM and soil DOM, including their sources, factors, and dynamics, is essential for advancing our understanding of global carbon cycles and evaluating carbon storage in inland ecosystems.

5. Conclusion

The remote sensing estimation of CDOM in lakes using Landsat 8 products on the GEE platform offers a new and effective way to evaluate DOM distribution and sources based on erosion regions on a national scale. The large number of *in situ* samples across China and the strong performance of the XGBoost model ($RMSE_{XGB,cal} = 0.87\ m^{-1}$) ensure the model's reliability. The mean $a_{CDOM}(355)$ values for freeze–thaw erosion regions, wind erosion regions, and hydraulic erosion regions vary significantly. From a molecular perspective, detailed variations in DOM compounds in diverse erosion regions were demonstrated, showing links to natural and anthropogenic factors. This study evaluated variations in CDOM distribution in lakes across China, considering terrestrial SOM systems as a whole. We found that modifications in the DOM compositions of lakes were also related to surface soil erosion, climatic conditions, and human activities. The differences in CDOM across the three erosion regions demonstrate that considering soil erosion classification is essential for assessing regional carbon cycles more precisely.

CRediT authorship contribution statement

Yingxin Shang: Writing - Original Draft, Methodology, Formal analysis. **Kaishan Song:** Supervision, Conceptualization, Writing - Review & Editing. **Zhidan Wen:** Investigation, Resources. **Fengfa Lai:** Investigation, Validation. **Ge Liu:** Software, Validation. **Hui Tao:** Visualization. **Xiangfei Yu:** Software.

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.

Acknowledgement

The research was jointly supported by the National Natural Science Foundation of China (42371390, 42471358), the Science & Technology Fundamental Resources Investigation Program (2021FY100406), the Jilin Provincial Department of Ecology and Environment(2024-01), Youth Innovation Promotion Association of Chinese Academy of Sciences of China granted for Dr. Yingxin Shang, the Staying Postdoctoral Researcher Support Program of Jilin Province granted for Dr. Yingxin Shang (2024), the Natural Science Foundation of Jilin Province, China (20220508017RC), the National funded postdoctoral researcher program (GZC20232638) and Young Scientist Group Project of Northeast Institute of Geography and Agroecology, China (2023QNXZ01). The authors thank all staff and students for their persistent assistance with field sampling and laboratory analysis. We thank the three anonymous reviewers for their constructive comments and suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2025.100570.

References

[1] V. Markogianni, A. Mentzafou, E. Dimitriou, Assessing the impacts of human activities and soil erosion on the water quality of plastira mountainous Mediterranean Lake, Greece, Environ. Earth Sci. 75 (10) (2016) 1–17.
[2] J. Aitkenhead-Peterson, R. Smart, M. Aitkenhead, M. Cresser, W. McDowell, Spatial and temporal variation of dissolved organic carbon export from gauged and ungauged watersheds of Dee Valley, Scotland: effect of land cover and C:N, Water Resour. Res. 43 (5) (2007) 1–11.
[3] H. Zhan, R. Zhou, P. Wang, Q. Zhou, Selective hydroxyl generation for efficient pollutant degradation by electronic structure modulation at Fe sites, P. Natl. Acad. Sci. USA 120 (26) (2023) e2305378120.
[4] S. Ahonen, K. Vuorio, R. Jones, H. Hämäläinen, K. Rantamo, M. Tiirola, A. Vähätalo, Assessing and predicting the influence of chromophoric dissolved organic matter on light absorption by phytoplankton in boreal lakes, Limnol. Oceanogr. 69 (2024) 422–433.
[5] Q. Zhou, C. Song, P. Wang, Z. Zhao, Y. Li, S. Zhan, Generating dual active species by triple-atom-sites through peroxymonosulfate activation for treating micropollutants in complex water, P. Natl. Acad. Sci. USA 120 (13) (2003) e2300085120.
[6] A. Sharma, K. Tiwari, P. Bhadoria, Effect of land use land cover change on soil erosion potential in an agricultural watershed, Environ. Monit. Assess. 173 (2011) 789–801.
[7] A. Sharpley, R. McDowelland, P. Kleinmann, Phosphorus loss from land to water: integrating agricultural and environmental management, Plant Soil 237 (2001) 287–307.
[8] Q. Zhou, Y. Liu, T. Li, H. Zhao, S. Daniel, W. Liu, O. Kurt, Cadmium adsorption to clay-microbe aggregates: implications for marine heavy metals cycling, Geochem. Cosmochim. Acta 290 (2020) 124–136.
[9] J. Weishaar, G. Aiken, B. Bergamaschi, M. Fram, R. Fujii, K. Mopper, Evaluation of specific ultraviolet absorbance as an indicator of the chemical composition and reactivity of dissolved organic carbon, Environ. Sci. Technol. 37 (2003) 4702–4708.
[10] D. He, K. Wang, Y. Pang, C. He, P. Li, Y. Li, S. Xiao, Q. Shi, Y. Sun, Hydrological management constraints on the chemistry of dissolved organic matter in the Three Gorges Reservoir, Water Res. 187 (2020) 116413.
[11] R. Marinho, J. Martinez, T. de Oliveira, W. Moreira, L. de Carvalho, P. Moreira-Turcq, T. Harmel, Estimating the colored dissolved organic matter in the negro river, amazon basin, with in situ remote sensing data, Remote Sens. 16 (4) (2024) 613.
[12] R. Zhang, R. Deng, Y. Liu, Y. Liang, W. Zhang, Developing new colored dissolved organic matter retrieval algorithms based on sparse learning, IEEE J-STARS 13 (2020) (2020) 3478–3492.
[13] H. Liu, H. Xu, Y. Wu, Z. Ai, J. Zhang, G. Liu, S. Xue, Effects of natural vegetation restoration on dissolved organic matter (DOM) biodegradability and its temperature sensitivity, Water Res. 191 (2021) 116792.

[14] A. Ruescas, M. Hieronymi, G. Mateo-Garcia, S. Koponen, G. Camps-Valls, Machine learning regression approaches for colored dissolved organic matter (CDOM) retrieval with S2-MSI and S3-OLCI simulated data, Remote Sens. 10 (5) (2018) 786.
[15] Z. Cao, R. Ma, T. Duan, N. Pahlevan, J. Melack, M. Shen, K. Xue, A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes, Remote Sens. Environ. 248 (2020) 111974.
[16] J. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2001) 1189–1232.
[17] Z. Zhao, K. Shi, Y. Zhang, Y. Zhou, B. Qin, Increased dominance of terrestrial component in dissolved organic matter in Chinese lakes, Water Res. 249 (2024) 121091.
[18] X. Hou, L. Feng, Y. Dai, C. Hu, L. Gibson, J. Tang, Z. Lee, Y. Wang, X. Cai, J. Liu, Y. Zheng, C. Zheng, Global mapping reveals increase in lacustrine algal blooms over the past decade, Nat. Geosci. 15 (2022) 130.
[19] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, R. Moore, Google earth engine: planetary-scale geospatial analysis for everyone, Remote Sens. Environ. 202 (2017) 18–27.
[20] T. Dittmar, S. Lennartz, H. Buck-Wiese, D. Hansell, C. Santinelli, C. Vanni, B. Blasius, J. Hehemann, Enigmatic persistence of dissolved organic matter in the ocean, Nat. Rev. Earth Environ. 2 (8) (2001) 570–583.
[21] B. Koch, T. Dittmar, From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter, Rapid Commun. Mass Spectrom. 20 (5) (2006) 926–932.
[22] M. Seidel, M. Beck, T. Riedel, H. Waska, I. Suryaputra, B. Schnetger, J. Niggemann, M. Simon, T. Dittmar, Biogeochemistry of dissolved organic matter in an anoxic intertidal creek bank, Geochem. Cosmochim. Acta 140 (1) (2014) 418–434.
[23] Y. Shang, Z. Wen, K. Song, G. Liu, F. Lai L. Lyu, S. Li, H. Tao, J. Hou, C. Fang, C. He, Q. Shi, D. He, Natural versus anthropogenic controls on the dissolved organic matter chemistry in lakes across China: insights from optical and molecular level analyses, Water Res. 221 (2022) 118779.
[24] C. He, Y. Yi, D. He, R. Cai, C. Chen, Q. Shi, Molecular composition of dissolved organic matter across diverse ecosystems: preliminary implications for biogeochemical cycling, J. Environ. Manag. 344 (2023) 118559.
[25] Y. Shang, K. Song, F. Lai, L. Lyu, G. Liu, C. Fang, J. Hou, S. Qiang, X. Yu, Z. Wen, Remote sensing of fluorescent humification levels and its potential environmental linkages in lakes across China, Water Res. 230 (2023) 119540.
[26] H. Chen, W. Kong, Q. Shi, F. Wang, C. He, J. Wu, Q. Lin, X. Zhang, Y. Zhu, C. Liang, Y. Luo, Patterns and drivers of the degradability of dissolved organic matter in dryland soils on the Tibetan Plateau, J. Appl. Ecol. 59 (3) (2022) 884–894.
[27] C. Griffin, J. McClelland, K. Frey, G. Fiske, R. Holmes, Quantifying CDOM and DOC in major Arctic rivers during ice-free conditions using Landsat TM and ETM+ data, Remote Sens. Environ. 209 (2018) 395–409.
[28] I. Joshi, E. D'Sa, C. Osburn, T. Bianchi, D. Ko, D. Oviedo-Vargas, A. Arellano, N. Ward, Assessing chromophoric dissolved organic matter (CDOM) distribution, stocks, and fluxes in Apalachicola Bay using combined field, VIIRS ocean color, and model observations, Remote Sens. Environ. 191 (2017) 359–372.
[29] D. Bowers, G. Harker, P. Smith, P. Tett, Optical properties of a region of freshwater influence (the Clyde Sea), Estuar. Coast Shelf Sci. 50 (5) (2000) 717–726.
[30] A. Mannino, M. Russ, S. Hooker, Algorithm development and validation for satellite-derived distributions of DOC and CDOM in the US Middle Atlantic Bight, J. Geophys. Res. 113 (2008) C7.
[31] O. Sagi, L. Rokach, Approximating XGBoost with an interpretable decision tree, Inf. Sci. 572 (2021) 522–542.
[32] Y. Zhang, L. Zhou, Chromophoric dissolved organic matter in inland waters: present knowledge and future challenges, Sci. Total Environ. 759 (2020) 143550.
[33] Y. Prairie, Evaluating the predictive power of regression model, Can. J. Fish. Aquat. Sci. 53 (3) (1996) 490–492.
[34] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, California, USA, 2016, pp. 785–794.
[35] X. Sun, Y. Zhang, Y. Zhang, K. Shi, Y. Zhou, N. Li, Machine learning algorithms for chromophoric dissolved organic matter (CDOM) estimation based on Landsat 8 images, Rem. Sens-Basel 13 (18) (2021) 3560.
[36] W. Zhu, Q. Yu, Y. Tian, B. Becker, T. Zheng, H. Carrick, An assessment of remote sensing algorithms for colored dissolved organic matter in complex freshwater environments, Remote Sens. Environ. 140 (2014) 766–778.
[37] M. Wang, L. Jiang, Atmospheric correction using the information from the short blue band, IEEE Trans. Geosci. Rem. Sens. 56 (10) (2018) 1–14.
[38] D. Wang, R. Ma, K. Xue, S. Loiselle, The assessment of Landsat-8 OLI atmospheric correction algorithms for inland waters, Rem. Sens-Basel 11 (2) (2019) 169.
[39] P. Shang, Y. Lu, Y. Du, R. Jaffe, R. Findlay, A. Wynn, Climatic and watershed controls of dissolved organic matter variation in streams across a gradient of agricultural land use, Sci. Total Environ. 612 (2018) 1442–1453.
[40] M. Groeneveld, D. Kothawala, L. Tranvik, Seasonally variable interactions between dissolved organic matter and mineral particles in an agricultural river, Aquat. Sci. 85 (2022) 2.
[41] D. Graeber, J. Gelbrecht, M. Pusch, C. Anlanger, D. von Schiller, Agriculture has changed the amount and composition of dissolved organic matter in Central European headwater streams, Sci. Total Environ. 438 (2012) 435–446.
[42] D. Elizabeth, C. Lara, M. Cornejo-D'Ottone, J. Nimptsch, M. Aguayo, B. Broitman, Contrasting land-uses in two small river basins impact the colored dissolved organic matter concentration and carbonate system along a river-coastal ocean continuum, Sci. Total Environ. 806 (2022) 150435.
[43] A. Stubbins, R. Spencer, H. Chen, P. Hatcher, K. Mopper, P. Hernes, V. Mwamba, A. Mangangu, J. Wabakanghanzi, J. Six, Illuminated darkness: molecular signatures of Congo River dissolved organic matter and its photochemical alteration as revealed by ultrahigh precision mass spectrometry, Limnol. Oceanogr. 55 (4) (2010) 1467–1477.
[44] D. Gao, L. Zhang, J. Liu, B. Peng, Z. Fan, W. Dai, P. Jiang, E. Bai, Responses of terrestrial nitrogen pools and dynamics to different patterns of freeze-thaw cycle: a meta-analysis, Glob. Change Biol. 24 (2018) 2377–2389.
[45] P. Raymond, J. Saiers, W. Sobczak, Hydrological and biogeochemical controls on watershed dissolved organic matter transport: pulse-shunt concept, Ecology 97 (1) (2016) 5–16.
[46] Z. Wen, K. Song, Y. Zhao, J. Du, J. Ma, Influence of environmental factors on spectral characteristic of chromophoric dissolved organic matter (CDOM) in Inner Mongolia plateau, China, Hydrol. Earth Syst. Sci. 20 (2016) 787–801.
[47] V. Mangal, N. Stock, C. Gueguen, Molecular characterization of phytoplankton dissolved organic matter (DOM) and sulfur components using high resolution Orbitrap mass spectrometry, Anal. Bioanal. Chem. 408 (7) (2016) 1891–1900.
[48] X. Liu, Y. Zhang, Y. Yin, Wind and submerged aquatic vegetation influence bio-optical properties in large shallow Lake Taihu, China, J. Geophys. Res. Biogeosci. 118 (2013) 713–727.
[49] Z. Wen, K. Song, Y. Shang, Y. Zhao, C. Fang, L. Lyu, Differences in the distribution and optical properties of DOM between fresh and saline lakes in a semi-arid area of Northern China, Aquat. Sci. 80 (2018) 22.
[50] M. Denis, L. Jeanneau, P. Petitjean, M. Anaëlle, G. Gérard, New molecular evidence for surface and sub-surface soil erosion controls on the composition of stream DOM during storm events, Biogeosci. Discuss. (2017) 1–26.