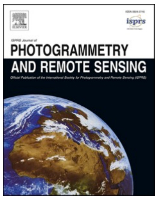
Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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



# A novel framework for river organic carbon retrieval through satellite data and machine learning

Shang Tian[a,b,c,d,1], Anmeng Sha[a,b,c,d,1], Yingzhong Luo[a], Yutian Ke[e], Robert Spencer[f], Xie Hu[g], Munan Ning[h], Yi Zhao[a], Rui Deng[a], Yang Gao[i], Yong Liu[a], Dongfeng Li[a,b,c,d,*]

[a] *Key Laboratory for Water and Sediment Sciences, Ministry of Education, College of Environmental Sciences and Engineering, Peking University, Beijing 100871, China*
[b] *State Environmental Protection Key Laboratory of All Materials Flux in River Ecosystems, Peking University, Beijing 100871, China*
[c] *Institute of Carbon Neutrality, Peking University, Beijing 100871, China*
[d] *Institute of Tibetan Plateau, Peking University, Beijing 100871, China*
[e] *Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA*
[f] *Department of Earth, Ocean and Atmospheric Science, Florida State University, Tallahassee, FL, USA*
[g] *College of Urban and Environmental Sciences, Peking University, Beijing 100871, China*
[h] *School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen, Guangdong Province, China*
[i] *Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China*

ARTICLE INFO

*Keywords:*
Particulate organic carbon, Dissolved organic carbon
Remote sensing, Machine learning
River

ABSTRACT

Rivers transport large amounts of carbon, serving as a critical link between terrestrial, coastal, and atmospheric biogeochemical cycles. However, our observations and understanding of long-term river carbon dynamics in large-scale remain limited. Integrating machine learning with remote sensing offers an effective approach for quantifying organic carbon (OC) from space. Here, we develop the Aquatic-Organic Carbon (Aqua-OC), a dynamic machine learning retrieval framework designed to estimate reach-scale river OC using nearly half a century of analysis-ready Landsat archives. We first integrate a globally representative river OC dataset, comprising 299,330 measurements of dissolved organic carbon (DOC) and 101,878 measurements of particulate organic carbon (POC). This dataset is then used to evaluate the performance of four machine learning methods, i.e., random forest (RF), extreme gradient boosting (XGBoost), Support vector regression (SVR), and deep neural network (DNN), using an optical water type classification strategy. We further leverage multimodal input features to enhance the Aqua-OC framework and OC retrieval accuracy by considering various factors related to OC sources and environmental conditions. The results demonstrate that the Aqua-OC can effectively estimate DOC ($R^2 = 0.68$, RMSE = 2.88 mg/L, Bias = 2.63 %, Error = 12.52 %) and POC ($R^2 = 0.76$, RMSE = 1.76 mg/L, Bias = 6.31 %, Error = 21.36 %). Additionally, the Mississippi River Basin case study demonstrates Aqua-OC's capability to map nearly four decades of reach-scale OC changes at a basin scale. This study provides a generalized method for satellite-based river OC retrieval at fine spatial and long-term temporal scales, thus offering an effective tool to quantify the rivers' role in the global carbon cycle.

## 1. Introduction

Rivers serve are the essential conduits between continents, oceans, and the atmosphere, playing key roles in the storage, transport, and transformation of carbon, thereby influencing global carbon cycles profoundly (Battin et al. 2023; Hilton and West 2020; Li et al. 2021; Regnier et al. 2022; Spencer and Raymond 2024; Zhang et al. 2023). However, our understanding and observations of river carbon are severely limited by the scarcity of in-field data, making it challenging to predict how global change affect the spatial distribution and temporal dynamics of river carbon across various scales (Battin et al. 2023). Recently, a consortium of hydrologists and biogeochemists have

[*] Corresponding author at: A215, College of Environmental Science and Engineering, Peking University, Beijing 100871, China.
*E-mail address:* dongfeng.li@pku.edu.cn (D. Li).
[1] These authors contribute equally to this article.

https://doi.org/10.1016/j.isprsjprs.2025.01.028
Received 20 November 2024; Received in revised form 19 January 2025; Accepted 24 January 2025

developed a blueprint for a global River Observation System (RIOS) to enhance our ability for observing and quantifying this crucial component of river organic carbon (OC) in the global carbon cycle (Battin et al. 2023; Dean and Battin 2024).

To achieve the overarching goals of RIOS, satellite networks are a powerful tool for providing spatio-temporal OC products over extended time series and mapping river OC dynamics from regional and global scales. In particular, Landsat satellites offer fine spatial resolution, short revisit cycles of less than a month, and a mission duration of over four decades (Wulder et al. 2022), making them an exceptional tool for understanding historical trends in river OC. Landsat data have been successfully applied in mapping other water quality parameters such as suspended sediment concentration (SSC) and Chlorophyll-a (Chl-a) in long-term series and the global scale (Dethier et al. 2022; Guo et al. 2022; Maciel et al. 2023; Wang et al. 2025).

River OC, comprising dissolved organic carbon (DOC) and particulate organic carbon (POC), has been better understood through recent advancements in remote sensing, greatly expanding our knowledge of its spatio-temporal dynamics in oceans and inland waters (Fichot et al. 2023; Griffin et al. 2018; Harkort and Duan 2023; Liu et al. 2019; Liu et al. 2021). However, existing methods tend to focus on local OC estimations and neglect the optical properties of water, failing to achieve long-term time series and river reach-scale reconstruction of OC dynamics. DOC includes colored dissolved organic matter (CDOM), which is detectable by satellite sensors in the visible bands (Griffin et al. 2018; Liu et al. 2024; Xu et al. 2018). Remote sensing techniques have recently emerged as a significant alternative to in-situ measurements, enabling the capture of CDOM dynamics over decades (Griffin et al. 2018; Kutser et al. 2005; Li et al. 2017; Zhang et al., 2024). However, it remains considerable uncertainties to estimate DOC from CDOM: (i) CDOM is not the only component of DOC, with the relative amount of non-chromophoric DOM varying among different rivers (Spencer et al. 2012), and the relationship of CDOM/DOC can be weak or exhibit seasonal variations in some water bodies (Mann et al. 2016); (ii) CDOM-based DOC retrieval approaches have been most successful when CDOM dominates the water composition. However, high SSC or Chl-a can interfere with the optical signal of CDOM, complicating accurate retrieval of DOC (Duan et al. 2014; Herrault et al. 2016). Overall, the variability in biogeochemical condition (e.g., DOC, SSC, and Chl-a levels) limits the applicability of empirical CDOM retrieval algorithms across different sites due to the distinct statistical relationships between CDOM and satellite data.

The most commonly used approach for retrieving POC is the SSC-based model (Huang et al. 2017; Matsuoka et al. 2022; Wang et al. 2020), but it has several limitations: (i) successful retrieval of POC from SSC depends on a stable POC:SSC ratio (Lin et al. 2018), which can vary based on the composition of particles in different optical water types; (ii) using the near-infrared (NIR) band is effective for measuring POC in highly turbid waters, but the turbidity in areas like some Arctic rivers is typically very low (Zhang et al. 2022); and (iii) river POC can be sourced from the exogenous (terrestrial erosion) and the autogenous (river biological production) (Marwick et al. 2015). Thus, SSC-based POC retrieval methods can only quantify the source from terrestrial erosion but typically not capture aquatic biomass.

The optical water type (OWT) classification strategy and machine learning modelling are expecting to solve the above-mentioned limitations. Spectral remote sensing reflectance is a critical optical property for deriving optical and biogeochemical properties of rivers that include Chl-a, CDOM absorption coefficient (the key element to retrieve DOC) and particulate back-scattering coefficient (the key element to retrieve POC) (Lee et al. 2002; Lee et al. 2015; O'Reilly et al. 1998). For POC, different sources of living and non-living fractions, seasonal changes of POC dynamics, and sources mixing (autochthonous production, terrestrial loading, sediment mixing) impact the spatio-temporal variability of water optical properties (Grey et al. 2001; Gu et al. 2006; Martineau et al. 2004). OWT classifies complex optical types, thus helping to explain the composition of OC and enhancing the retrieval capability of the model (Lin et al. 2018). Furthermore, machine learning has the potential to overcome the nonlinear, multicollinear, and heteroscedastic relationship between satellite images and water quality parameters, making it a powerful tool to quantify water quality parameters (includes DOC and POC) from satellite data across open ocean, coastal water, and inland water (Bonelli et al. 2022; Guo et al. 2024; Harkort and Duan 2023; Liu et al. 2021; Tian et al. 2023).

This study presents the Aqua-OC framework for retrieving long-term OC dynamics across river networks (Fig. 1) and aims to: (i) integrate a globally representative river OC dataset, comprising 299,330 measurements of DOC and 101,878 measurements of POC over 1984–2020, and thus contributing the river carbon research community; (ii) develop a dynamic OC retrieval framework based on machine learning methods and OWT classification strategy, providing a globally applicable solution for estimating long-term and river reach-scale OC; and (iii) evaluate the ability of Aqua-OC framework to reveal the spatio-temporal patterns of OC in Mississippi River and analyze the underlying drivers of OC changes. To our best knowledge, the Aqua-OC marks the initial effort to estimate river OC to align with the RIOS blueprint, expecting to enhance the understanding of rivers' role in the global carbon cycle.

## 2. Data

### 2.1. In-situ OC data

Enhancing the integration of in-situ data with the Landsat archive necessitates a robust repository of OC observations to elevate the possibility of achieving spatio-temporal collocation between satellite and field samples. In this study, we integrate a comprehensive global OC dataset to consider OWTs and ensure enough data to develop machine learning models. The data used for OC retrieval model development are sourced from:

1) AquaSat: The AquaSat merges several established public datasets encompassing the continental United States, such as the Water Quality Portal, LAGOS NE, and the Landsat archive. This integration comprises over 28,000 pairings of DOC observations with atmospherically corrected surface reflectance data from Landsat missions spanning the period from 1984 to 2019 (Ross et al. 2019).
2) GRQA: The GRQA initiative seeks to enhance water quality data coverage by consolidating and harmonizing five national, continental, and global datasets (Virro et al. 2021). During the GRQA compilation process, observation data from the five sources is standardized into a common format, and the associated metadata is harmonized. Outliers are identified and flagged, time series characteristics are calculated, and duplicate observations from sources with spatial overlap were detected. The final dataset includes 265,252 DOC measurements and 98,045 POC measurements.
3) MOREPOC: Ke et al. (2022) introduce MOREPOC (MOdern River archivEs of Particulate Organic Carbon), a novel, openly accessible, georeferenced global database focused on POC within SSC. This database encompasses information on POC levels gathered from 233 sites spanning 121 significant river systems worldwide. Encompassing 3,053 POC records, this database is designed to serve the earth system community as a foundational resource for constructing comprehensive and quantitative models related to the transport, transformation, and ultimate destiny of terrestrial POC. In MOREPOC, five different water depth profiles sampling techniques are compiled to measure POC content and composition. We only use POC data from surface sampling technology ("type = SS").

### 2.2. Matchups between in situ data and satellite data

Due to the highly customizable of sampling time and location of public dataset, we implement strict quality control to matchup inte-

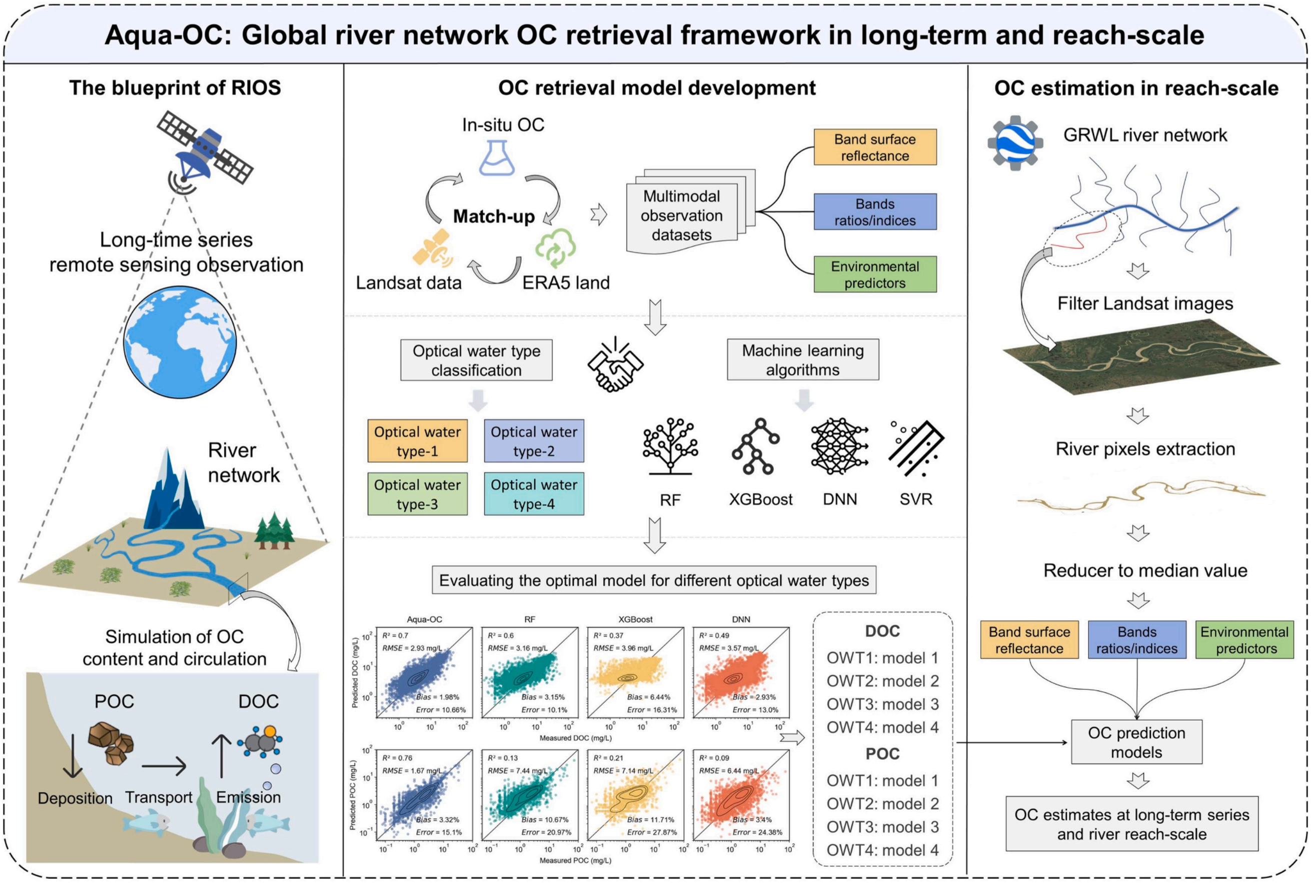


**Fig. 1.** The overarching methodological of the Aqua-OC framework. The RIOS blueprint enhances the ability to observe and quantify the role of rivers in the global carbon cycle. To develop Aqua-OC, over 400,000 organic carbon (OC) measurements and multimodal input features are used to evaluate various machine learning methods and identify the optimal approach for different optical water types (OWTs). Aqua-OC has been deployed on the Google Earth Engine (GEE) platform to enable long-term series and river reach-scale OC retrieval.

grated OC datasets and Landsat observations:

1) We use four widely used water surface extraction methods (Dynamic Surface Water Extent (DSWE) (Jones 2019), Modified Normalized Difference Water Index (MNDWI) (Xu 2006), the Arid Region Water Detection Rule (ARWDR) (Zhou et al. 2021), and Automated Water Extraction Index (AWEI) (Feyisa et al. 2014)) to ensure the sampling location is water pixel to exclude the error or wrong sampling point coordinates record. If the pixel of sampling point is identified by all water surface extraction method, this sample is proved to be effective.
2) We only use the Landsat-5/7 data and exclude Landsat-8 to matchup the in-situ measurements to avoid the problem of inconsistency of Landsat sensors in the four decades timespan.
3) To minimize errors caused by water movement and the influence of temporal variability in OC, we constrain the time difference between in situ measurements and Landsat overpass to be within ±1 days. Our time window is stricter than other studies that used larger offsets of 3–10 days to maximize matchup quantity (Kloiber et al. 2002; Olmanson et al. 2008).
4) We extract a 3 × 3 pixels window to exclude the effect of possible adjacency effects from the field sample sites and use the pixel quality_flags assessment bit index to eliminate low-quality pixels. The low-quality pixels are eliminated by the QA_PIXEL band bit mask technology is: 'cloud', 'snow', and 'cloud shadow'.
5) If more than half of the pixels in the window have remained after applying the 'pixel quality flags bit index' and the coefficient of variation (CV) among the remaining pixels is less than 0.15, the mean values of these valid retrievals are used.

Finally, we matchup a total of 46,020 OC measurements, including 36,862 DOC data and 9,158 POC data. The matchups have good temporal and spatial coverage (spatially basically cover the global scale and temporally span across 1984 to 2020, Fig. 2). Statistical histogram of OC is depicted in Fig. 3. The DOC concentration shows a normal distribution and the POC concentration shows a bimodal distribution.

### 2.3. Meteorological and environmental data

In this study, we incorporate seven environmental variables—high vegetation Leaf Area Index (LAI), average LAI for low vegetation, daily air temperature, wind speed, surface net solar radiation, evaporation over inland waters, and daily precipitation—as inputs into our machine learning model to estimate DOC concentration. These environmental datasets are sourced from ECMWF Reanalysis v5 (ERA5) Land product. Notably, ERA5-Land has demonstrated heightened reliability and accuracy in comparison to gauge or satellite-based products (Harkort and Duan 2023). We leverage ERA5-Land data spanning from 1984 to 2023. The rules of matchup between ERA5-Land data and in-situ data follow the description in Section 2.2.

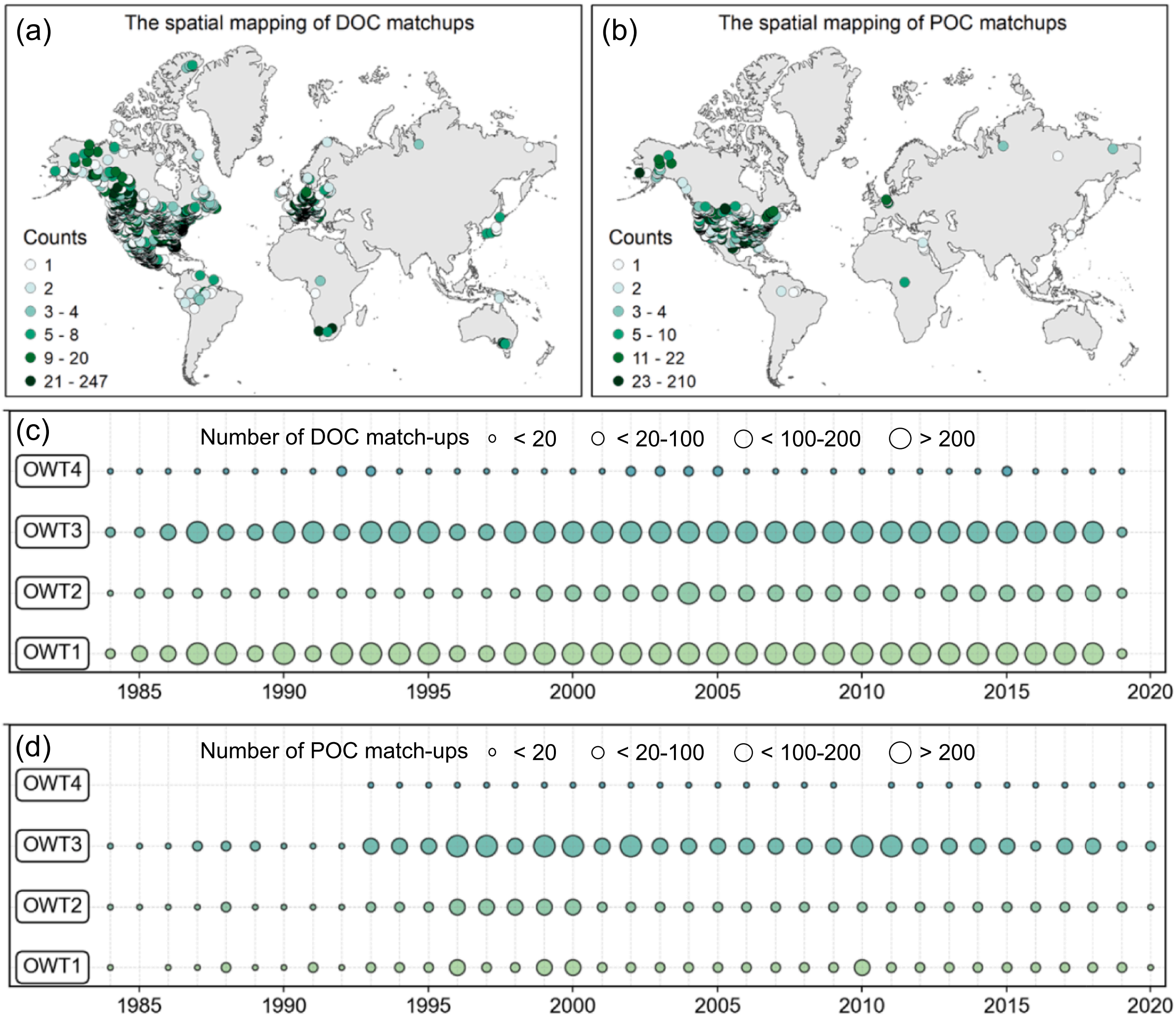


**Fig. 2.** The spatio-temporal mapping of OC matchups. (a, b) The spatial mapping of integrated OC matchups. (c, d) The temporal distribution of integrated OC matchups across 1984 to 2020.

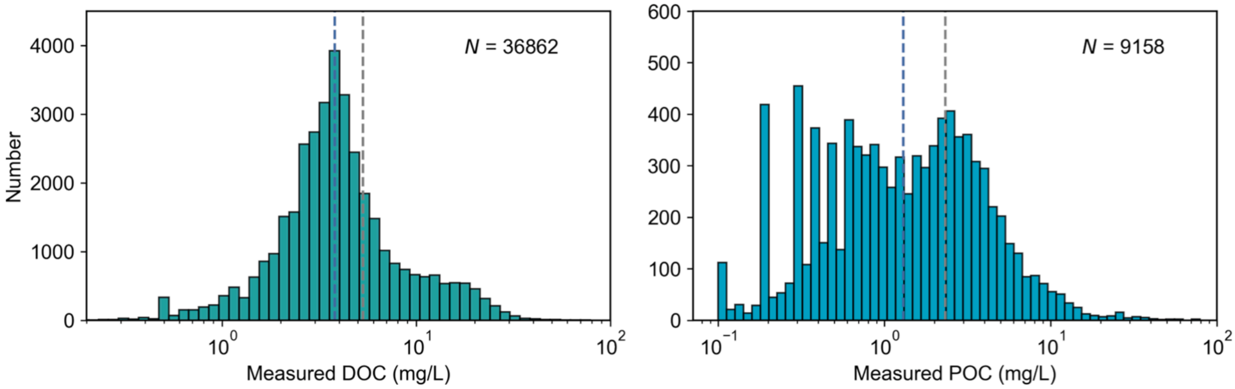


**Fig. 3.** Histogram of OC concentrations of all matchups. The dash line of blue and gray represent the mean and median values, respectively. The mean and median of DOC are 5.3 mg/L and 3.8 mg /L, respectively, while the mean and median of POC are 2.3 mg/L and 1.3 mg/L, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

## 3. Development of Aqua-OC framework

### 3.1. Machine learning methods

We choose four prominent machine learning algorithms to identify the most representative method from diverse machine learning families, including tree-based algorithms, kernel methods, and neural networks, for predicting OC. The machine learning algorithms are Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), and Deep Neural Network (DNN). The DNN is implemented in

Tensorflow (*version 2.14.0*) and all other machine learning models are implemented in package scikit-learn (*version 1.3.1*) in Python 3.8. Random samples are generated from the integrated OC datasets with 7:2:1 split ratio for training set, testing set and independent set, respectively. The independent set do not participate in the training and calibration processes of modelling and are specifically used to evaluate model generalization. We implement 10-fold cross-validation to optimize hyperparameters and assess whether variations in dataset partitioning affect model performance. The summary of hyper-parameters used for tuning of these models is given in Table S1. We use the SHapley Additive exPlanations (SHAP) to quantify features importance.

RF is an ensemble learning method that builds multiple independent decision trees using random subsets of features. These random trees collectively form a "Random Forest" with the model prediction derived from the average of all outputs from the individual trees. Compared to other machine learning methods, RF is straightforward to interpret while delivering strong performance, making it a popular choice in water quality research (Wang et al. 2022a).

XGBoost is a boosting ensemble algorithm introduced by Chen and Guestrin (2016), consisting of multiple regression trees and optimizing the gradient boosting decision tree. Each training round builds on the results of the previous one. The algorithm incorporates a regularization term to control model complexity, helping to prevent overfitting, reduce computation time, and achieve the optimal solution efficiently.

SVR is effective for handling small sample sizes and nonlinear problems, making it commonly used in remote sensing research, such as multiple water quality parameters retrieval (Tian et al. 2023). It is based on the theory of structural risk minimization, which constructs an optimal classification surface in the feature space, enabling the learner to achieve global optimization.

DNN is a feed-forward artificial neural network (ANN) with multiple hidden layers, which are fully connected between the input and output layer. Many complex regression and classification problems require multi-hidden layer DNNs for better data representation, deeper feature extraction, and improved accuracy in learning patterns from input data. The DNN model has shown effectiveness in various regression problems because of its robustness and flexibly (Wang et al. 2022b).

### 3.2. Optical water type classification

We adopt a robust approach to generate the k-means clustering based on four spectral bands (B1-B4) to and exclude the SWIR bands (B5-B7) due to less or no information of water properties (Dethier et al. 2020; Dethier et al. 2019). Specifically, the k-means algorithm requires a predetermined number of clusters. We utilize the silhouette coefficient (with the maximum value indicating the best clustering effect) to determine the optimal number of clusters. We implemented k-means clustering using the module from Scikit-learn v1.3.1.

The silhouette coefficient of sample *i* is calculated as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{1}$$

$$a(i) = ave\{a(1), a(2), a(3), \cdots\cdots, a(k)\} \tag{2}$$

$$b(i) = min\{b(i_1), b(i_2), b(i_3), \cdots\cdots, b(i_l)\} \tag{3}$$

where $a(i)$ is the intra-cluster dissimilarity, which is the distance from sample *i* to other samples within the cluster to which it belongs. $b(i)$ is the minimum inter-cluster dissimilarity, which is the average distance from sample *i* to all points within clusters that do not include it. The value of $s(i)$ ranges from −1 to 1. The silhouette coefficient ($s$) for the clustering result is the mean of the silhouette coefficients for all samples, calculated as:

$$s = \frac{\sum_{i=1}^{n} S_{(i)}}{n} \tag{4}$$

Employing the k-means unsupervised method to group samples spectrally, the distance metric utilized in this clustering is defined as follows:

$$d(x, c) = 1 - \frac{xc'}{\sqrt{(xx')(cc')}} \tag{5}$$

where $x$ represents the reference spectrum, and $c$ represents the target spectrum.

### 3.3. Specific input features for DOC and POC

We employ multimodal input features by considering various factors related to OC sources and environmental conditions. For DOC, the model inputs contain spectral bands surface reflectance, variable bands indices and environmental predictors. Phytoplankton growth influences variations in DOC, as phytoplankton release extracellular DOC and generate significant amounts of DOC during cell degradation (Zhang et al. 2009). So, we also use Chl-a related indices and empirical algorithms to predict DOC. Recent study demonstrates that incorporating environmental predictors into models significantly improves models' performance, highlighting the crucial role of environmental processes in modeling DOC variations in inland waters at larger scales. Harkort and Duan (2023) provide a detailed description of the reasons for selecting environmental predictive factors. The variable importance of DOC retrieval model across different OWTs shows in Fig. 4.

For POC, the model inputs contain spectral bands surface reflectance, variable bands indices and constituents correlated to POC. Considering of the complex source of POC in the river system can be divided into exogenous (sediment and terrestrial loading) and autogenous (mainly produced by produced plankton) sources (Marwick et al. 2015). We also use plankton and sediment relative bands indices to predict POC (Liu et al. 2019). The variable importance of POC retrieval model across different OWTs is shown in Fig. 4.

### 3.4. Model evaluation

We employ four statistical indicators to comprehensively evaluate model performance, including $R^2$, Root Mean Squared Error (RMSE), Median Symmetric Accuracy ("Error"), and Symmetric Signed Percentage Bias ("Bias"). $R^2$ and RMSE represent the proportion of variance and statistical error in the overall data that the model can explain, which are commonly used in machine learning. "Error" is defined as a symmetric percentage error that equally penalizes both over- and under-estimation, with lower values indicating better performance and perfect accuracy represented by 0 %. "Bias" refers to a percentage bias that also maintains symmetry between over- and under-estimation, where values closer to zero reflect better performance; positive values indicate over-estimation, while negative values indicate under-estimation.

$$R^2 = 1 - \sum_{i=1}^{N} \frac{(X_{r\_i} - X_{m\_i})^2}{(X_{r\_i} - \overline{X_m})^2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_r - X_m)^2} \tag{7}$$

$$Error = 100 \times (e^{median\left(\left|\log_{10}\frac{X_r}{X_m}\right|\right)} - 1) \tag{8}$$

$$MdLQ = median(\log_{10}\frac{X_r}{X_m}) \tag{9}$$

$$Bias = 100 \times sign(MdLQ) \times (e^{|MdLQ|} - 1) \tag{10}$$

In Eqs. (6)–(10), $X_m$ represents the in-situ measured data, and $X_r$ represents the retrieved data. Although our data is normally distributed for DOC, the model can still be affected by extreme values. Both metrics

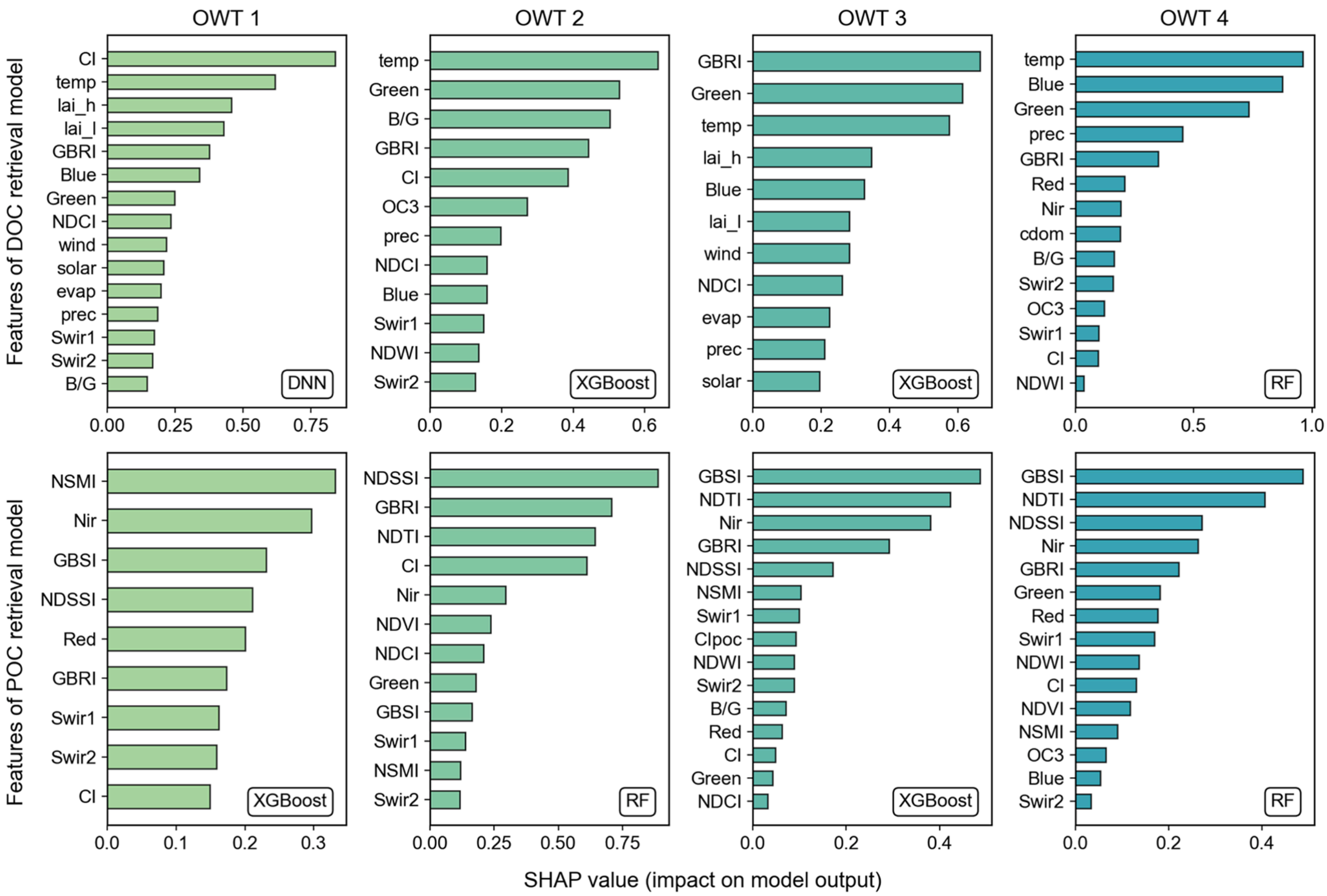


**Fig. 4.** Feature importance in the machine learning algorithms for predicting OC across four OWTs. The indices, formulas, and references of the features are listed in Table S4.

of Error and Bias are zero-centered and symmetric and designed to address the widely documented drawbacks of other commonly used statistical measures (e.g., RMSE and $R^2$ are asymmetric with respect to over-forecasting and under-forecasting, constrained to be positive, and not resistant to outliers) (Morley et al. 2018). Therefore, comprehensive metrics better reflect the overall performance of the model.

## 4. Application of Aqua-OC at the basin scale

### 4.1. Definition of river reaches

We define the boundaries of rivers based on the summary statistics version of the Global River Widths from Landsat (GRWL) database (Allen and Pavelsky 2018). GRWL utilizes Landsat archives to precisely delineate river boundaries, aligning seamlessly with the objectives of our study. The division of river reaches is based on a simplified vector version of grouping and merging, aiming to maximize the representation of dynamic river features. Since some river reaches may not represent permanent water bodies throughout the 40-year time series, we exclude reaches where the frequency of missing values for a river reach exceeds 30 % over the four decades time series. We have finalized 451 reaches for Mississippi River basin (study area shows in section of 5.3.1). Based on buffer zones created by doubling the river width along the reach centerline, we use our images pre-processing scheme to extract the spectral bands surface reflectance and calculate spectral indices. Then we apply the Aqua-OC framework to river reaches to obtain the annual average OC.

### 4.2. Satellite images pre-processing

Landsat archive products courtesy of the USGS provides excellent opportunity for scientific research organizations to reconstruct historical long-term (1984–2023) water quality parameters (WQPs). We use Landsat atmospherically corrected tier 1 images through Google Earth Engine (GEE). The following processing steps are implemented:

1) Landsat surface reflectance products have been selected (LANDSAT/LT05/C02/T1_L2, LANDSAT/LE07/C02/T1_L2), where Landsat 5 and 7 are atmospheric corrected by the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) (Schmidt et al. 2013).
2) We exclude the Landsat-8 data to avoid of inconsistency of Landsat sensors in a course of four decades timespan (Roy et al. 2016).
3) By utilizing the internal cloud cover metadata field in Landsat, we filter images with cloud coverage below 30 %.
4) We exclude low-quality pixels from Landsat Collection-2 using the pixel quality_flags assessment bit index. This process identifies and removes pixels affected by: ‘cloud’, ‘snow’, ‘cloud shadow’.
5) Water surface extraction used DSWE method developed by the USGS to classify water pixels of Landsat images in GEE.
6) Based on the river vector that is reduced to vectors by extracting the water surface, we remove small areas such as lake pixels based on whether the water area features intersect with GRWL river centerlines.

### 4.3. Model application

We match OWT and specific Aqua-OC algorithm for each river reach.

Based on buffer zone created by doubling the river width (provided by GRWL database) along the reach centerline, we use our images preprocessing scheme to process images data within the buffer zone. To match each OWT-specific Aqua-OC algorithm, we determined the OWT of the reach based on the spectral angle and then use specific Aqua-OC algorithm. Specifically, we match the target spectrum $r^{*}(\lambda_i)$ (spectrum composed of the average surface reflectance of each band in the buffer zone) and reference spectrum $r^{ref}(\lambda_i)$ (spectrum composed of average surface reflectance of each band of three OWTs) to calculate the spectral angle. The smallest spectral angle to the $r^{ref}(\lambda_i)$ is designated as the most likely OWT. The spectral angle ($\alpha$) between $r^{*}(\lambda_i)$ and $r^{ref}(\lambda_i)$ is calculated as follows:

$$\alpha = cos^{-1}\frac{\sum_{i=1}^{n}\{r^{*}(\lambda_i)r^{ref}(\lambda_i)\}}{\sqrt{\sum_{i=1}^{n}\{r^{*}(\lambda_i)\}^2}\sqrt{\sum_{i=1}^{n}\{r^{ref}(\lambda_i)\}^2}} \quad (11)$$

### 4.4. Analyzing long-term trends

We used the Mann-Kendall non-parametric trend test (version 1.4.3 of pymannkendall package in Python 3.8) to analyze changes in annual OC for each river reach, using α = 0.1 to test for statistical significance. Mann-Kendall test is robust to missing values and serial dependence. Nevertheless, we are careful to avoid over-interpreting the sparse data from the early years of the satellite record. Therefore, we count the number of valid observations for each reach over four decades period, and the trend test is only applied to reaches with more than 10 valid observations. For example, if a reach with more than 10 valid observations in four decades time series, trend analysis is applied.

## 5. Result

### 5.1. Cluster results

Cluster analysis differentiates subsets based on the shape and magnitude of surface reflectance, resulting in four optimal clusters. These conditions are not unique to specific rivers, lakes, or regions, nor are they solely due to differences between freshwater and marine waters; they reflect average conditions influenced by the optical properties of the water column, ultimately depending on the absorption and scattering characteristics of in-water constituents, such as phytoplankton and non-phytoplankton particles. Fig. 5 displays surface reflectance spectra colored by OWT, revealing distinct differences in mean surface reflectance spectra for each OWT. Generally, OWT reflectance spectra exhibit unique characteristics, with clear variations in spectral shape and magnitude of surface reflectance across defined OWTs, indicating that the classification scheme effectively captures the unique spectral traits of global in-situ measurements. Fig. 6 shows the distribution of optical constituent components, DOC and POC, for each OWT group. The highest median DOC concentration of 3.79 mg/L is found in OWT 2, while OWT 3 has the highest concentration of POC of 1.8 mg/L. The median value for each constituent further confirms that OWTs are not derived from a simple OC concentration threshold but instead result from a mixed combination of each optical constituent.

### 5.2. Model performance

We test the performance of machine learning algorithms to estimate POC and DOC concentrations in different OWT groups. For DOC, Algorithm performance is highly variable across the four tested models. For OWT 1 (Fig. 7), DNN shows excellent ability than other methods with an $R^2$ of 0.64, RMSE of 2.32 mg/L, Bias of 1.72 %, and Error of 9.78 %. Although inferior to DNN, XGBoost also shows the good performance with an $R^2$ of 0.41, RMSE of 2.96 mg/L, Bias of 3.04 %, and Error of 11.19 %. In both the DNN and XGBoost models, some deviation values near the measured concentrations are less than 1 mg/L. However, these low deviations are consistent across all four methods, indicating that the issue may be related to the quality of the in-situ data rather than errors within the machine learning methods. For OWT 2, XGBoost outperforms other methods with $R^2$ of 0.65, RMSE of 3.96 mg/L, Bias of 4.68 %, and Error of 14.51 %. However, XGBoost exhibits a slight over-estimation of lower DOC values, particularly those below 1 mg/L. For OWT 3, XGBoost also outperforms other methods with $R^2$ of 0.70, RMSE of 2.05 mg/L, Bias of −3.15 %, and Error of 16.55 %. For OWT 4, RF effectively estimates DOC, achieving $R^2$ of 0.62, RMSE of 2.41 mg/L, Bias of 3.77 %, and Error of 12.76 %. The 10-fold cross-validation confirms the robustness of models across different OWTs (Table S2). The model performance in independent set as shown in Fig. S1.

In evaluating the model performance of POC in OWT 1 (Fig. 8), XGBoost stand out among the methods, achieving an $R^2$ of 0.70, RMSE of

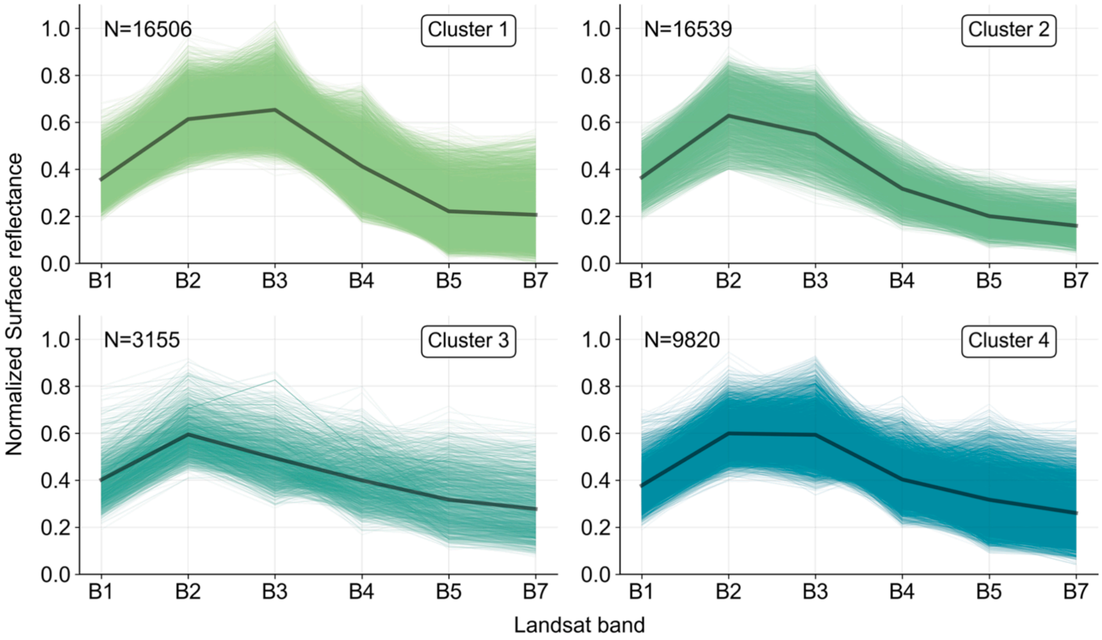


**Fig. 5.** Normalized surface reflectance data of OC matchups sorted into the four clusters from the k-means cluster analysis; grey lines denote mean reflectance.

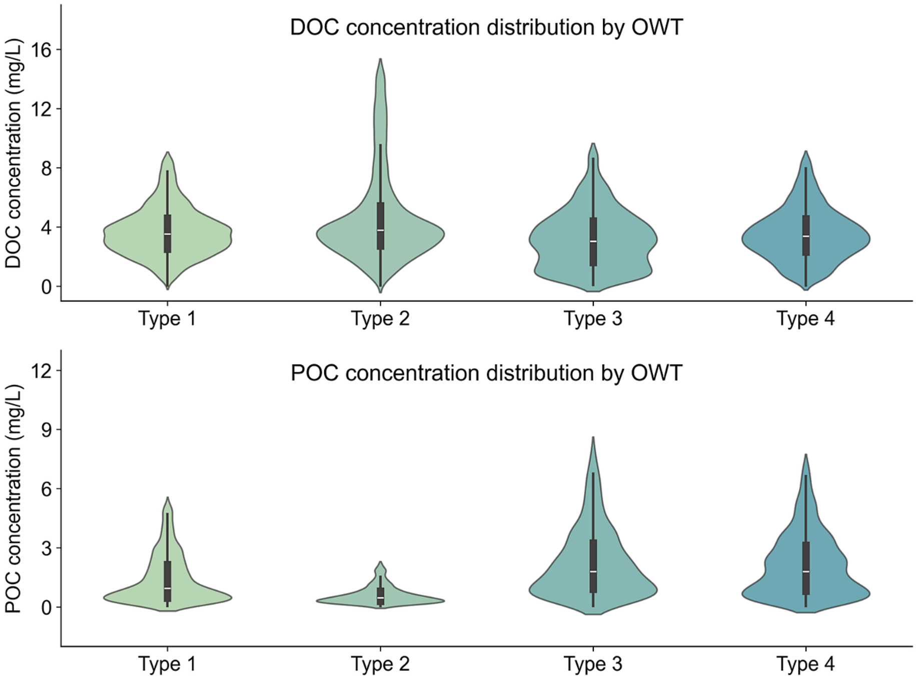


**Fig. 6.** In-situ OC values distribution for OWT groups. Within the violins, white lines show the median value, thick lines the interquartile range, and whiskers extend to a maximum of 1.5 times the interquartile range.

1.26 mg/L, Bias of 10.83 %, and Error of 23.87 %. In OWT 2, RF demonstrates superior performance compared to other methods with an $R^2$ of 0.77, RMSE of 0.84 mg/L, Bias of −2.79 %, and Error of 21.29 %. Moving on to OWT 3, XGBoost performs well over the other methods with an $R^2$ of 0.72, RMSE of 3.06 mg/L, Bias of 10.46 %, and Error of 22.93 %. In the case of OWT 4, RF effectively estimates POC with an $R^2$ of 0.70, RMSE of 2.08 mg/L, Bias of 5.83 %, and Error of 18.87 %. Overall, all models struggle to predict lower POC concentrations (below 1 mg/L) across all OWTs, although our Aqua-OC retrieval framework demonstrates comparatively better performance. The Aqua-OC framework displays a tighter clustering of data points along the identity line, with higher POC concentration values closely aligning with the identity line except for a few points. The 10-fold cross-validation confirms the robustness of the models across different OWTs (Table S2). The model performance in independent set as shown in Fig. S2.

### 5.3. Aqua-OC application to large-scale river OC estimates

#### 5.3.1. Study area

The Mississippi River Basin (MRB) originates at Lake Itasca in northwestern Minnesota, USA, meandering through the south-central region of North America. It traverses 31 states in the United States and two provinces in Canada before eventually flowing into the Gulf of Mexico (Fig. 9). Covering a drainage area of 3.24 million $km^2$, it ranks as the third-largest drainage basin in the world. The southeastern region of the MRB has a subtropical monsoon climate, characterized by hot, rainy summers and mild, drier winters. In contrast, the rest of the basin experiences a temperate continental climate, with less precipitation and greater temperature variations between winter and summer (Yin et al. 2023). Over the past few decades, the MRB has been significantly impacted by rapid human activities, which have substantially altered the magnitude, annual variations, and decadal trends of OC concentration and flux.

#### 5.3.2. Spatial-temporal variations of OC

Large spatial variations in river OC concentration are found in the MRB (Fig. 10). In general, the mean DOC concentration in MRB is 5.4 ± 2.2 mg/L (mean ±1 standard deviation, same hereafter), and DOC concentrations are relatively high in the central and northern regions and low in the northwestern and southern regions. High DOC concentration distribution in the middle MRB is associated with the expanded cropland, which exports a large amount of DOC through cultivation-intensified degradation of the organic matter (Ren et al. 2016). The reforestation in the upper and lower Ohio basins is believed to have contributed to an increase in DOC export due to higher DOC leaching in forest land compared to other systems (Delprat et al. 1997; Pandey and Pandey 2013). In contrast, the mean POC concentration in MRB is 2.2 ± 1.2 mg/L, and POC concentration is relatively high in the central and southern area and low in the northern and eastern part of the basin. The spatial variability of POC consist with the spatial distribution of SSC provided by Gardner et al. (2023).

The long-term (1984–2023) trend of river DOC and POC in MRB shows significant differences (Fig. 10). In marked contrast to the significant increasing trend of DOC observed in approximately 24 % of river reaches (108 out of 451, $p < 0.1$), POC shows widespread significant decreases, with around 43 % of river reaches (194 out of 451, $p < 0.1$) exhibiting the declining trend. Previous studies contend that on a decadal scale over the MRB, the changes in historical OC concentration have been mainly due to direct anthropogenic drivers, i.e., land use change, intensive land management practices, and massive damming (Raymond and Cole 2003; Raymond et al. 2008). Human activities, such as fertilization, irrigation, embanking, and tillage, are the dominant factor contributing to the long-term increase in river DOC concentration (Raymond et al. 2008; Shen et al. 2021). In contrast, the construction of dams appears to be the dominant driver behind the widespread decreases observed in POC.

Overall, Aqua-OC demonstrates exceptional capability in mapping long-term series and reach-scale OC across large basins. Its ability to capture the spatio-temporal variations of river network OC provides invaluable insights into the dynamic processes governing river OC dynamics. This deeper understanding of river OC patterns facilitates a more comprehensive analysis of the underlying drivers and ecological implications of these variations.

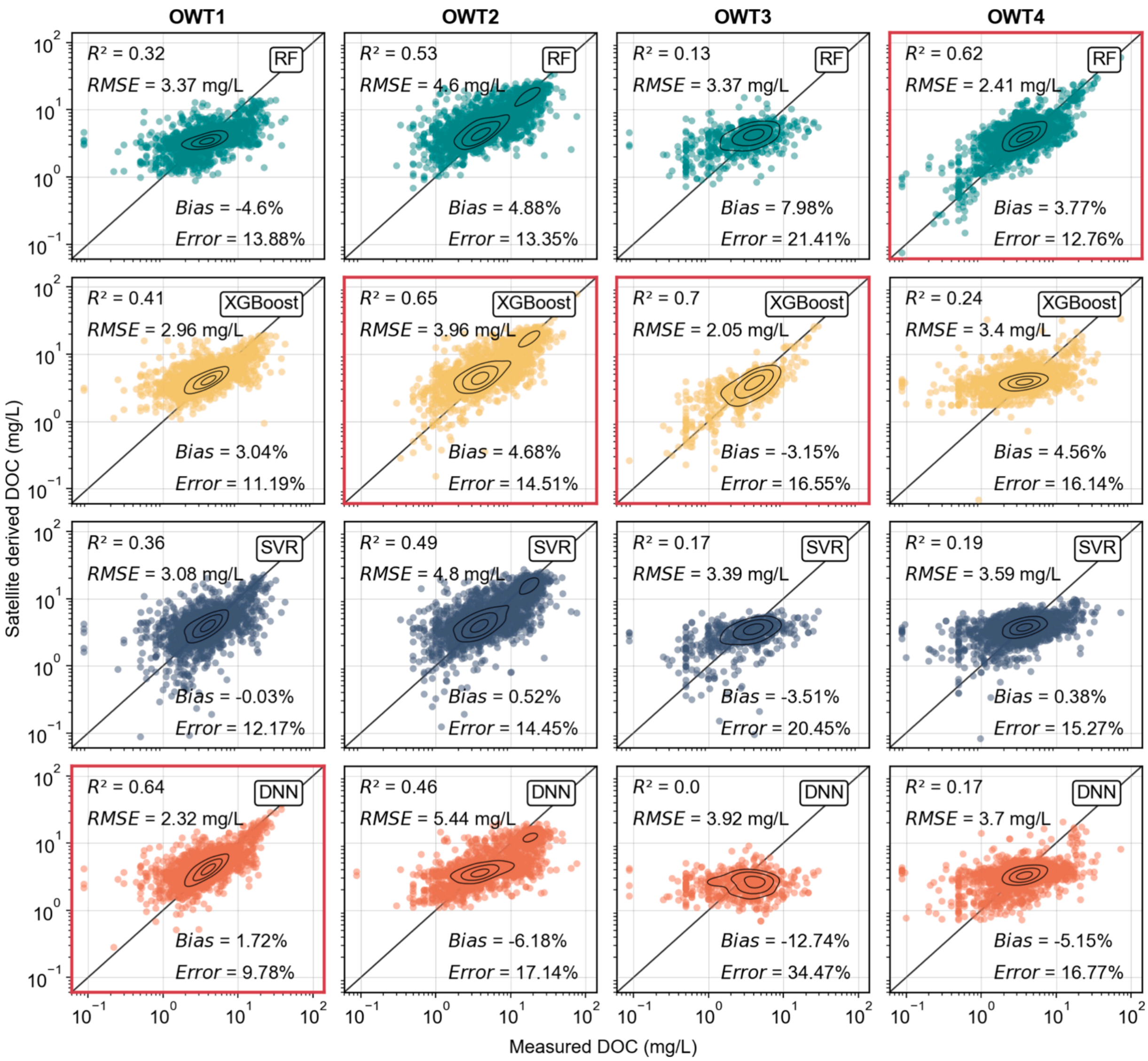


**Fig. 7.** Scatter plots show the performance of the four machine learning algorithms in retrieving DOC across four OWTs for testing set. The red box represents the best-performing model in different OWT. The contour lines correspond with density estimates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

## 6. Discussion

### 6.1. The first practice for blueprint of RIOS

A consortium of hydrologists and biogeochemists convened at the University of Bristol on June 14–15, 2023, to explore the potential structure of a global RIOS (Battin et al. 2023; Dean and Battin 2024). The researchers established overarching, core science, and community objectives, concluding with the importance of a RIOS for quantifying and predicting rivers' contributions to the global carbon cycle. RIOS will integrate data from river sensor networks, satellite imagery, and mathematical models to estimate carbon dynamics associated with river ecosystem metabolism. Within the RIOS blueprint, satellite technology serves as a cornerstone for providing long-term historical data to inform the nested organization of river OC dynamics and their spatio-temporal heterogeneity. Although satellites provide opportunities and successfully apply in quantifying the river network topology, surface and inundation area, water storage, discharge, and suspended sediments (Allen and Pavelsky 2018; Dethier et al. 2022; Frasson et al. 2021; Lin et al. 2021; Yamazaki et al. 2019), satellite imagery still leaves substantial data gaps in ecosystem processes and carbon biogeochemistry, especially in river OC dynamics. To our best knowledge, the Aqua-OC framework is the first practice of RIOS, aiming to provide a living observational network to determine long-term trends in river carbon dynamics. Furthermore, we provide an opportunity to enrich the global river OC data and so that the research community could conduct further analysis about carbon transformations and fluxes across the land–ocean continuum and understand the processes driving by anthropogenic and climatic disturbances.

### 6.2. Aqua-OC empowers understanding the role of rivers in global carbon cycling

Leveraging OWT strategy and a data-driven machine learning retrieval framework, Aqua-OC offers high accuracy, global portability, and the ability to reconstruct long-term and reach-scale OC dynamics. These advantages position Aqua-OC to make significant contributions to assessing the role of rivers in the global carbon cycle. Firstly, the accuracy and portability of OC retrieval are demonstrably reliable. By assigning the best performing algorithm (i.e., lowest RMSE, Error and Bias) to a particular water type, the dynamic switching strategy (that means by assigning the best performing algorithm to a particular OWT)

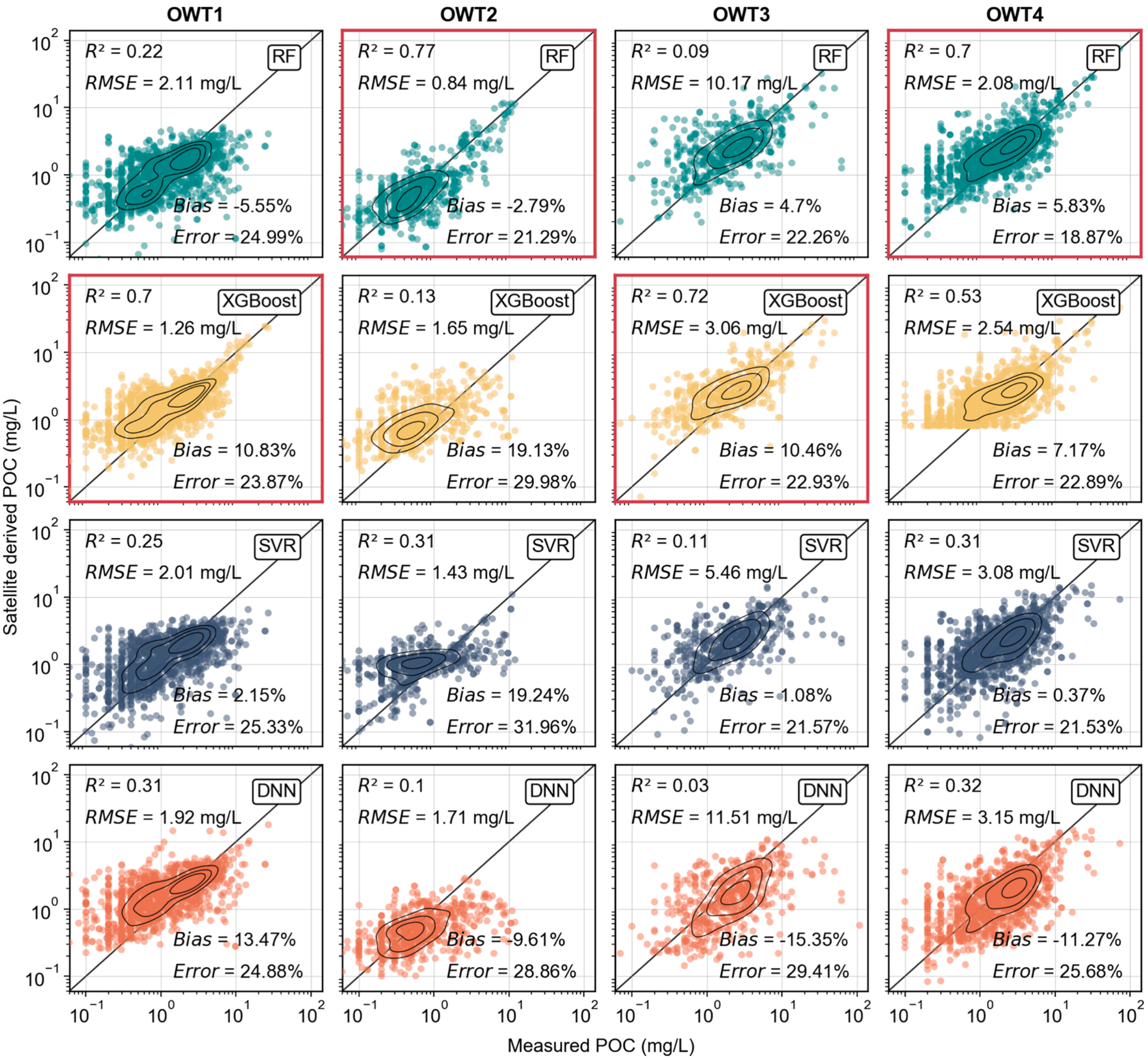


**Fig. 8.** Scatter plots show the performance of the four machine learning algorithms in retrieving POC across four OWTs for testing set. The red box represents the best-performing model in different OWT. The contour lines correspond with density estimates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is the core of the Aqua-OC framework. The dynamic retrieval framework is based on global covered dataset and can capture the OC features across different OWT, thus improving the retrieval accuracy (Fig. 11). For DOC, the overall $R^2$, RMSE, Bias and Error have improved for the dynamic switching framework, from 0.37 to 0.60, 3.16–3.96 mg/L, 2.93 %–6.44 %, and 10.11 %–16.32 % improve to 0.68, 2.88 mg/L, 2.63 %, and 12.52 %. For POC, the overall $R^2$, RMSE, Error and Bias have improved from 0.09 to 0.21, 6.44–7.44 mg/L, 3.40 %–11.71 %, and 20.97 %–27.87 % to 0.76, 1.76 mg/L, 6.31 %, and 21.36 %. The improvement in model accuracy is attributed to the classification of OWTs and the reliability of input parameters. Different OWTs represent unique water optical properties, characterizing the composition and concentration range of OC. In addition to spectral bands surface reflectance, we have incorporated numerous band indices and empirical algorithms as input parameters to comprehensively consider variables affecting OC variations. For instance, we have included POC-related internal and external source parameters (SSC and Chl-a) and utilized environmental variables to predict DOC. Furthermore, the integration of a massive OC database (over 400,000 measurements) enables the full potential of machine learning.

Secondly, the Aqua-OC framework provides significant advantages to the river carbon research community, including long-term series dating back to 1984, extensive global coverage, and spatially explicit estimations at the river reach scale, enabling multifaceted assessments of rivers' role in the global carbon cycle. The field of river carbon research is rapidly evolving and gaining momentum (Karlsson 2024; McClelland et al. 2016; Regnier et al. 2013; Repasch et al. 2021; Spencer and Raymond 2024), yet numerous gaps remain to be filled, including (i) estimating the carbon flux from terrestrial and biospheric sources into river networks and its subsequent delivery to the ocean; (ii) understanding the transport processes of OC in terrestrial river networks disturbed by anthropogenic and climatic factors; (iii) bridging observations from local to regional to global scales; and (iv) extending temporal observations from event to decadal scales (Battin et al. 2023; Bauer et al. 2013; Bianchi et al. 2024; Dean and Battin 2024; Regnier et al. 2022). Aqua-OC possesses the potential to address these objectives. For instance, by enabling high-precision mapping of OC in major rivers worldwide, Aqua-OC can provide insights into the global OC flux transported from rivers to the ocean. Additionally, its ability to reconstruct OC at the reach-scale over four decades enables long-term analysis of OC transport processes and dynamics in terrestrial river networks, shedding light on the underlying drivers of these changes. Overall,

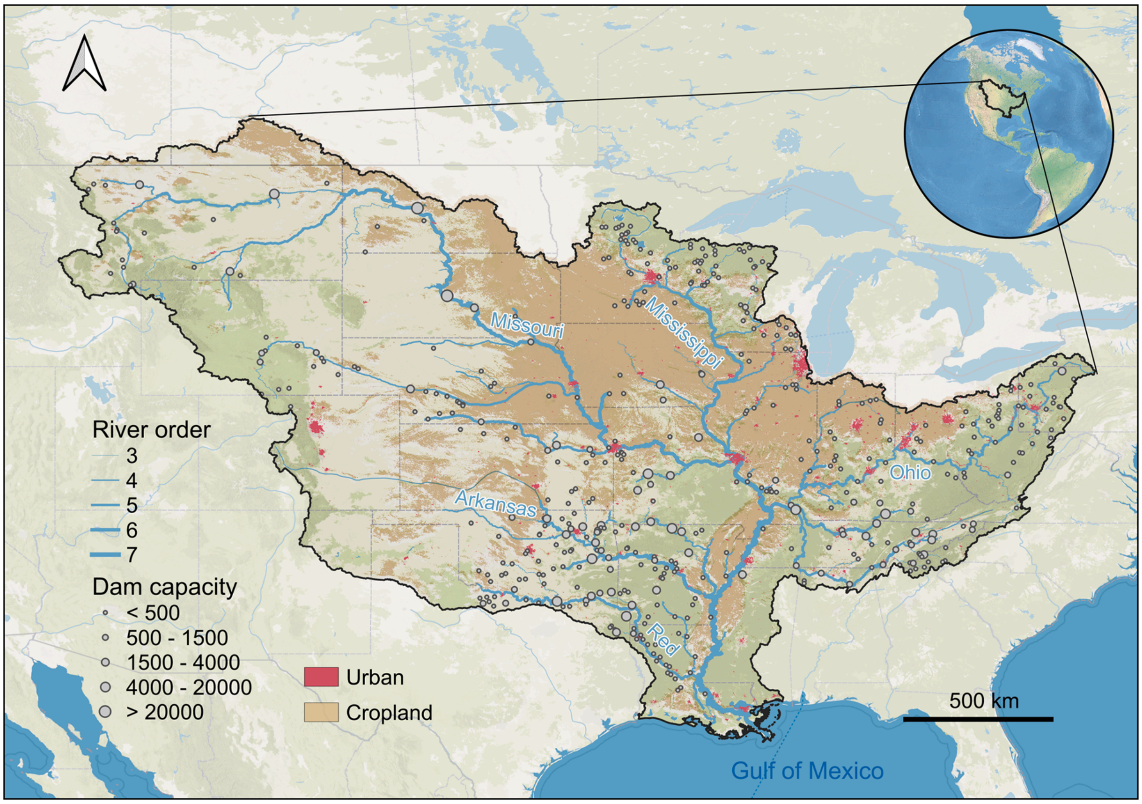


**Fig. 9.** Sketch map of the Mississippi River basin. Definition of river order following the Strahler ordering system. The land cover data from the MODIS Land Cover Type Product (https://lpdaac.usgs.gov/products/mcd12q1v006/). Dam data are sourced from the Global Reservoir and Dam Database (GRanD, version 1.3) (Lehner et al. 2011).

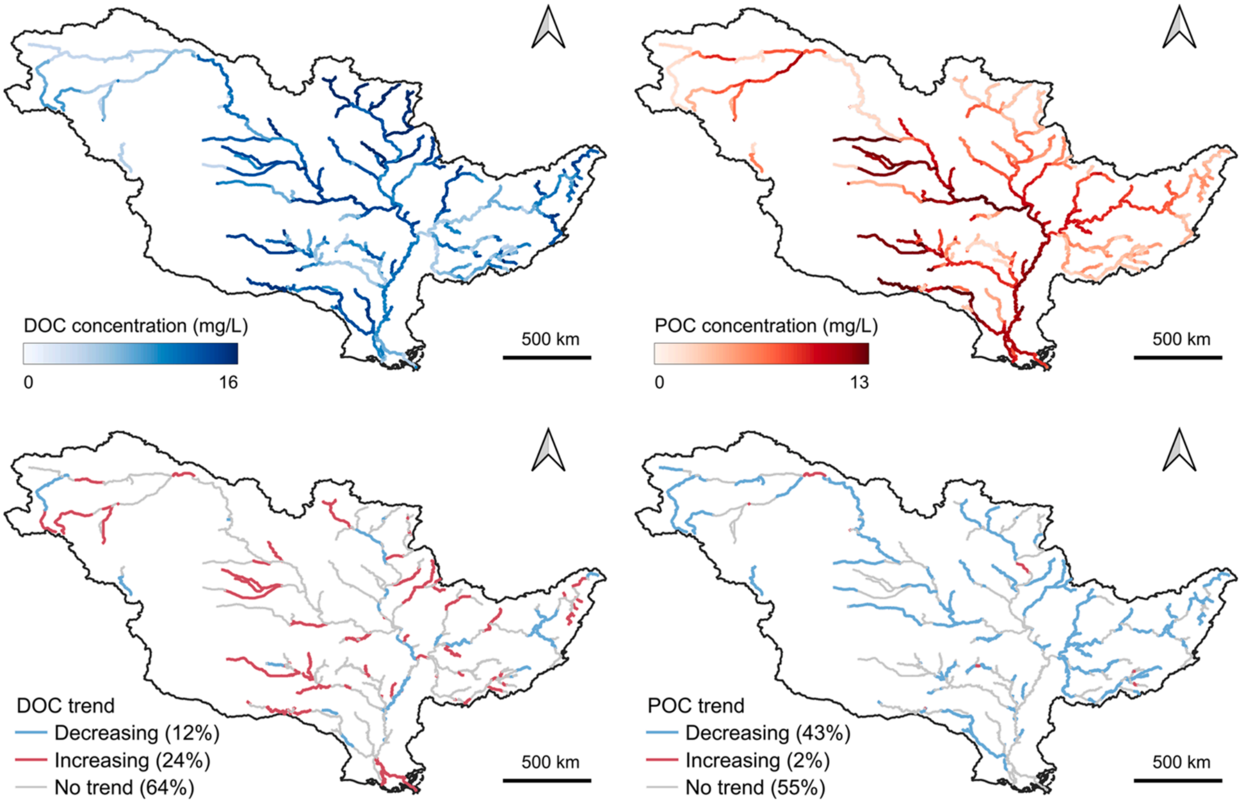


**Fig. 10.** Map of trends and concentration in mean annual OC from 1984 to 2018 showing the percentage of river reaches affected by decreasing, increasing, or no significant trend.

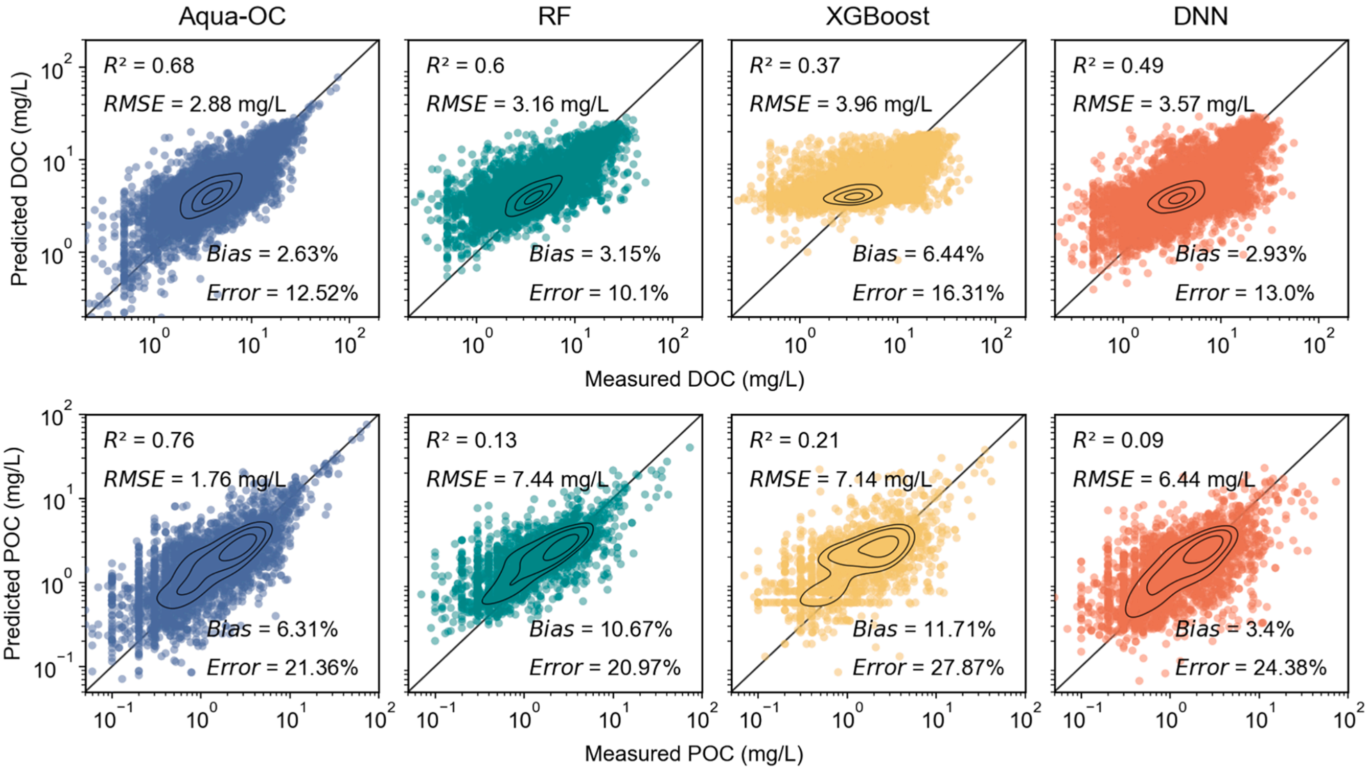


**Fig. 11.** Scatter plots show the performance of the Aqua-OC compared with other machine learning algorithms. The contour lines correspond with density estimates.

Aqua-OC represents a significant advancement in quantifying river OC, enhancing our ability to observe and predict changes in this crucial component of rivers in the global carbon cycle.

### 6.3. Model uncertainty

Land-based atmospheric correction strategy and the errors of cloud masking algorithms may be the limitations of Aqua-OC. While many aquatic-based atmospheric correction algorithms have been developed to address the challenges of accurately extracting water-leaving reflectance in aquatic remote sensing, these methods are not universal. Their effectiveness can vary significantly based on different atmospheric processors and specific OWTs, which can greatly influence the prediction outcomes of water quality mapping (Pahlevan et al. 2021). Machine learning models can yield improved predictions even without atmospheric correction by circumventing the secondary errors that arise from neglecting the suitable application range of particular atmospheric correction algorithms (Medina-Lopez 2020; Toming et al. 2020). Furthermore, the normal atmospheric correction algorithm for Landsat satellites performs as well as aquatic-based correction algorithms (e.g., ACOLITE, SeaDAS) (Kuhn et al. 2019), and have been successfully applied in aquatic remote sensing over regional to global extents using data-driven approaches (Dethier et al. 2022; Dethier et al. 2023; Gardner et al. 2023; Li et al. 2024). A small fraction of low-quality pixels is affected by the bottom reflectance of optically shallow water and thin clouds are not representative of the true surface reflectance of river reach. For example, according to map the reach-scale band surface reflectance, the high reflectance pixels increase towards the shorelines of the river (Fig. 12a). The increased uncertainties are primarily attributed to the mixed pixels or bottom reflectance of optically shallow water. Additionally, low-quality pixels that are not identified by the cloud masking algorithm may exhibit high reflectance values that differ from other river pixels (Fig. 12a). We extract the median surface reflectance to minimize the impact of uncertain outliers. The results indicate that our median extraction strategy more accurately represents the true reflectance characteristics of river reaches, which closely aligns with the mean calculated after removing outliers of low-quality pixels (Fig. 12b, Table S3).

The other source of uncertainty arises from the in-situ data. Our integrated OC dataset (GRQA and AquaSat) combines existing public datasets, leading to considerable variability in the techniques, methods, and interpretations of OC measurements. Despite our efforts to implement strict quality controls for matching OC data, discrepancies in sampling techniques, analytical inconsistencies, and their corresponding calibrations remain unavoidable. We find the range of predicted values in the low-value range is relatively dispersed, a common phenomenon for modeling that use public datasets (Dethier et al. 2020; Gardner et al. 2023; Harkort and Duan 2023; Maciel et al. 2023). However, the extent to which these differences impact algorithm development and validation remains unclear. Given the large sample size (over 400,000 samples) used in this research and the additional quality controls implemented, minor variations in sampling methods are unlikely to significantly affect model development and evaluation.

## 7. Conclusions

This study introduces the Aquatic-Organic Carbon (Aqua-OC), a novel and dynamic OC retrieval framework that leverages state-of-the-art machine learning techniques and an optical water type (OWT) classification strategy. We further integrate a comprehensive global dataset of more than 400,000 OC observations and demonstrate the ability of Aqua-OC framework in estimating long-term series and reach-scale OC across various OWTs. The results show that Aqua-OC is a reliable tool for estimating DOC ($R^2 = 0.68$, RMSE = 2.88 mg/L, Bias = 2.63 %, Error = 12.52 %) and POC ($R^2 = 0.76$, RMSE = 1.76 mg/L, Bias = 6.31 %, Error = 21.36 %). Furthermore, we apply the Aqua-OC framework in the Mississippi River Basin and find the substantial increases in DOC (24 %; 108 out of 451 river reaches) and declines in POC (43 %; 194 out of 451 river reaches) over 1984–2023 in response to agriculture expansion and damming. The observed patterns in the

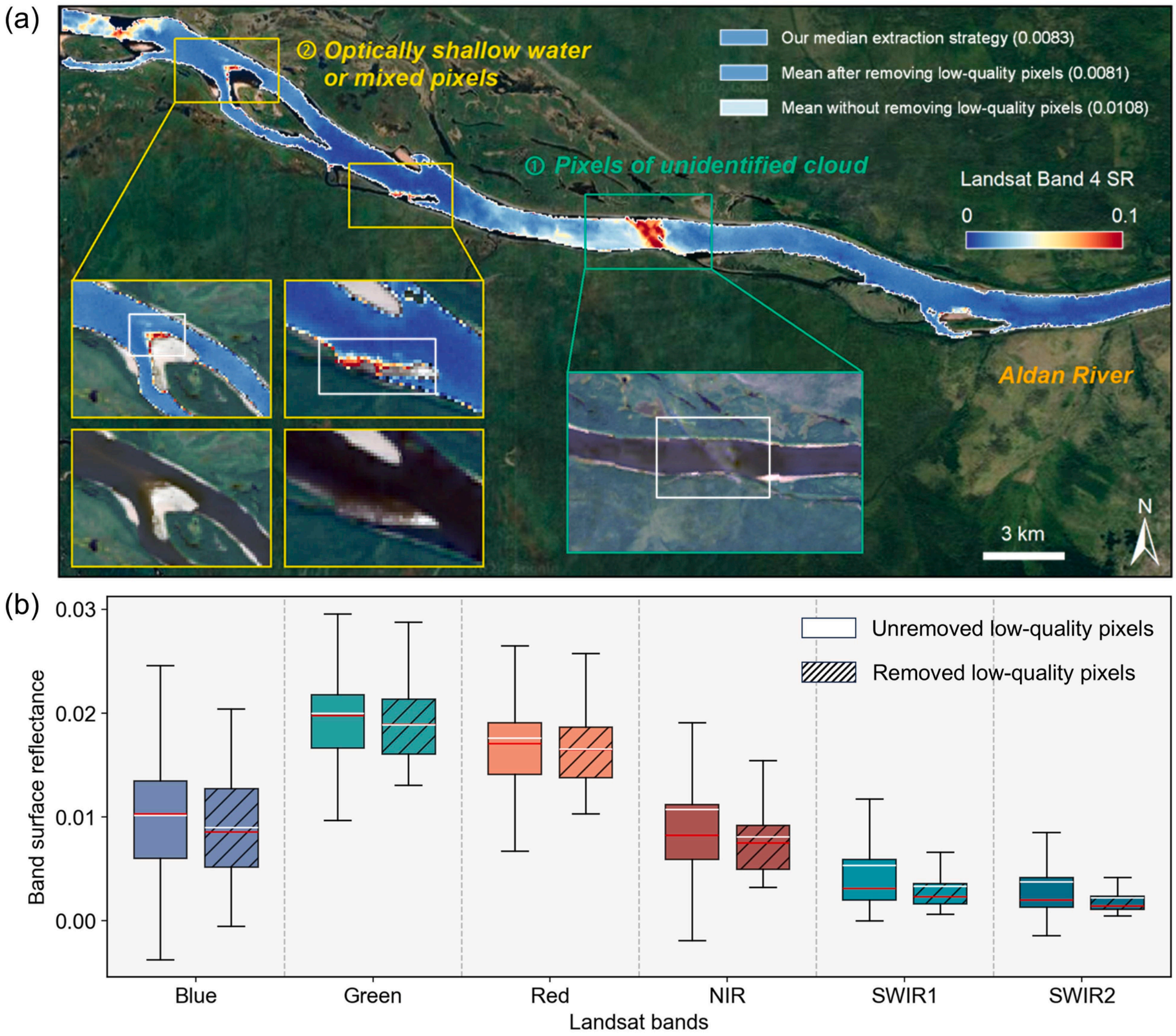


**Fig. 12.** (a) The case of Aldan River shows the median extraction strategy could be effective minimize the uncertainty caused by low-quality pixels. The reflectance values of river reach obtained using the median extraction strategy (0.0083) are essentially the same as the mean reflectance values derived after removing outliers (0.0081). (b) Surface reflectance distribution of bands with and without low-quality pixels removed. On boxes, the red and white center line shows the median and mean values, the whiskers denote the full range (min and max), and box limits indicate the 25th and 75th percentiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Mississippi River Basin align with previous findings, indicating that Aqua-OC provides reasonable OC estimates. While locally tuned models may offer higher accuracy, Aqua-OC addresses key challenges in river remote sensing and has the potential to reveal large-scale trends in river OC over extended periods, seasons, and at reach-scale. Overall, this study marks the first application of a global River Observation System (RIOS) and has great potential to advance river carbon observation over contrasting spatial scales and deepen our understanding of rivers' role in the global carbon cycle.

## CRediT authorship contribution statement

**Shang Tian:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Investigation, Formal analysis, Data curation, Conceptualization. **Anmeng Sha:** Visualization, Validation, Software, Resources, Formal analysis, Data curation, Conceptualization. **Yingzhong Luo:** Visualization, Validation, Methodology, Investigation, Data curation. **Yutian Ke:** Visualization, Methodology, Investigation. **Robert Spencer:** Writing – review & editing, Funding acquisition, Data curation, Conceptualization. **Xie Hu:** Validation, Supervision, Software, Resources. **Munan Ning:** Validation, Methodology. **Yi Zhao:** Writing – review & editing, Software, Conceptualization. **Rui Deng:** Visualization. **Yang Gao:** Writing – review & editing, Validation, Supervision, Formal analysis, Conceptualization. **Yong Liu:** Writing – review & editing, Visualization, Validation, Supervision, Resources. **Dongfeng Li:** Writing – review & editing, Investigation, Funding acquisition, Formal analysis, 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.

## Acknowledgments

This study is supported by the National Natural Science Foundation of China (Grant No. 52479055), National Key R&D Program of China (Grant No. 2023YFC3209900), the Fundamental Research Funds for the Central Universities, Peking University (7100604495), Alibaba DAMO Academy Young Fellow Award, AI for Science (AI4S)-Preferred Program of Peking University and the Excellent Young Scientists Fund of the National Natural Science Foundation of China. We thank Jan Karlsson, Tom Battin and Marisa Repasch for the earlier fruitful discussions. We appreciate the careful and thoughtful comments of two anonymous reviewers, which helped improve this paper substantially.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.isprsjprs.2025.01.028.

## References

Allen, G.H., Pavelsky, T.M., 2018. Global extent of rivers and streams. Science. 361, 585–588. https://doi.org/10.1126/science.aat0636.

Battin, T.J., Lauerwald, R., Bernhardt, E.S., Bertuzzo, E., Gener, L.G., Hall Jr, R.O., Hotchkiss, E.R., Maavara, T., Pavelsky, T.M., Ran, L., 2023. River ecosystem metabolism and carbon biogeochemistry in a changing world. Nature. 613, 449–459. https://doi.org/10.1038/s41586-022-05500-8.

Bauer, J.E., Cai, W.-J., Raymond, P.A., Bianchi, T.S., Hopkinson, C.S., Regnier, P.A., 2013. The changing carbon cycle of the coastal ocean. Nature. 504, 61–70. https://doi.org/10.1038/nature12857.

Bianchi, T.S., Mayer, L.M., Amaral, J.H., Arndt, S., Galy, V., Kemp, D.B., Kuehl, S.A., Murray, N.J., Regnier, P., 2024. Anthropogenic impacts on mud and organic carbon cycling. Nat. Geosci. 1–11. https://doi.org/10.1038/s41561-024-01405-5.

Bonelli, A.G., Loisel, H., Jorge, D.S., Mangin, A., d'Andon, O.F., Vantrepotte, V., 2022. A new method to estimate the dissolved organic carbon concentration from remote sensing in the global open ocean. Remote Sens. Environ. 281, 113227. https://doi.org/10.1016/j.rse.2022.113227.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Dean, J.F., Battin, T.J., 2024. Future directions for river carbon biogeochemistry observations. Nat. Water. 2, 219–222. https://doi.org/10.1038/s44221-024-00207-8.

Delprat, L., Chassin, P., Linères, M., Jambert, C., 1997. Characterization of dissolved organic carbon in cleared forest soils converted to maize cultivation. Eur. J. Agron. 7, 201–210. https://doi.org/10.1016/S1161-0301(97)00046-4.

Dethier, E., Renshaw, C., Magilligan, F., 2020. Toward improved accuracy of remote sensing approaches for quantifying suspended sediment: Implications for suspended-sediment monitoring. J. Geophys. Res.: Earth Surf. 125, e2019JF005033. https://doi.org/10.1029/2019JF005033.

Dethier, E.N., Sartain, S.L., Lutz, D.A., 2019. Heightened levels and seasonal inversion of riverine suspended sediment in a tropical biodiversity hot spot due to artisanal gold mining. Proc. Natl. Acad. Sci. 116, 23936–23941. https://doi.org/10.1073/pnas.1907842116.

Dethier, E.N., Renshaw, C.E., Magilligan, F.J., 2022. Rapid changes to global river suspended sediment flux by humans. Science. 376, 1447–1452. https://doi.org/10.1126/science.abn7980.

Dethier, E.N., Silman, M., Leiva, J.D., Alqahtani, S., Fernandez, L.E., Pauca, P., Çamalan, S., Tomhave, P., Magilligan, F.J., Renshaw, C.E., 2023. A global rise in alluvial mining increases sediment load in tropical rivers. Nature. 620, 787–793. https://doi.org/10.1038/s41586-023-06309-9.

Duan, H., Feng, L., Ma, R., Zhang, Y., Loiselle, S.A., 2014. Variability of particulate organic carbon in inland waters observed from MODIS Aqua imagery. Environ. Res. Lett. 9, 084011. https://doi.org/10.1088/1748-9326/9/8/084011.

Feyisa, G.L., Meilby, H., Fensholt, R., Proud, S.R., 2014. Automated water extraction index: a new technique for surface water mapping using landsat imagery. Remote Sens. Environ. 140, 23–35. https://doi.org/10.1016/j.rse.2013.08.029.

Fichot, C.G., Tzortziou, M., Mannino, A., 2023. Remote sensing of dissolved organic carbon (DOC) stocks, fluxes and transformations along the land-ocean aquatic continuum: advances, challenges, and opportunities. Earth Sci. Rev. 242, 104446. https://doi.org/10.1016/j.earscirev.2023.104446.

Frasson, R.P.d.M., Durand, M.T., Larnier, K., Gleason, C., Andreadis, K.M., Hagemann, M., Dudley, R., Bjerklie, D., Oubanas, H., Garambois, P.A., 2021. Exploring the factors controlling the error characteristics of the Surface Water and Ocean Topography mission discharge estimates. Water Resour. Res. 57, e2020WR028519. https://doi.org/10.1029/2020WR028519.

Gardner, J., Pavelsky, T., Topp, S., Yang, X., Ross, M.R., Cohen, S., 2023. Human activities change suspended sediment concentration along rivers. Environ. Res. Lett. 18, 064032. https://doi.org/10.1088/1748-9326/acd8d8.

Grey, J., Jones, R.I., Sleep, D., 2001. Seasonal changes in the importance of the source of organic matter to the diet of zooplankton in Loch Ness, as indicated by stable isotope analysis. Limnol. Oceanogr. 46, 505–513. https://doi.org/10.4319/lo.2001.46.3.0505.

Griffin, C., McClelland, J., Frey, K., Fiske, G., Holmes, R., 2018. Quantifying CDOM and DOC in major Arctic rivers during ice-free conditions using Landsat TM and ETM+ data. Remote Sens. Environ. 209, 395–409. https://doi.org/10.1016/j.rse.2018.02.060.

Gu, B., Chapman, A.D., Schelske, C.L., 2006. Factors controlling seasonal variations in stable isotope composition of particulate organic matter in a softwater eutrophic lake. Limnol. Oceanogr. 51, 2837–2848. https://doi.org/10.4319/lo.2006.51.6.2837.

Guo, H., Tian, S., Huang, J.J., Zhu, X., Wang, B., Zhang, Z., 2022. Performance of deep learning in mapping water quality of Lake Simcoe with long-term Landsat archive. ISPRS J. Photogramm. Remote Sens. 183, 451–469. https://doi.org/10.1016/j.isprsjprs.2021.11.023.

Guo, H., Huang, J.J., Zhu, X., Tian, S., Wang, B., 2024. Spatiotemporal variation reconstruction of total phosphorus in the Great Lakes since 2002 using remote sensing and deep neural network. Water Res. 255, 121493. https://doi.org/10.1016/j.watres.2024.121493.

Harkort, L., Duan, Z., 2023. Estimation of dissolved organic carbon from inland waters at a large scale using satellite data and machine learning methods. Water Res. 229, 119478. https://doi.org/10.1016/j.watres.2022.119478.

Herrault, P.-A., Gandois, L., Gascoin, S., Tananaev, N., Le Dantec, T., Teisserenc, R., 2016. Using high spatio-temporal optical remote sensing to monitor dissolved organic carbon in the Arctic River Yenisei. Remote Sens. 8, 803. https://doi.org/10.3390/rs8100803.

Hilton, R.G., West, A.J., 2020. Mountains, erosion and the carbon cycle. Nat. Rev. Earth & Environ. 1, 284–299. https://doi.org/10.1038/s43017-020-0058-6.

Huang, C., Jiang, Q., Yao, L., Li, Y., Yang, H., Huang, T., Zhang, M., 2017. Spatiotemporal variation in particulate organic carbon based on long-term MODIS observations in Taihu Lake, China. Remote Sens. 9, 624. https://doi.org/10.3390/rs9060624.

Jones, J.W., 2019. Improved automated detection of subpixel-scale inundation—revised dynamic surface water extent (DSWE) partial surface water tests. Remote Sens. 11, 374. https://doi.org/10.3390/rs11040374.

Karlsson, J., 2024. Emergent responses shape the coupled carbon cycle in a changing Arctic. Nat. Water 1–2. https://doi.org/10.1038/s44221-024-00250-5.

Ke, Y., Calmels, D., Bouchez, J., Quantin, C., 2022. MOdern river archivEs of particulate organic carbon: MOREPOC. Earth Syst. Sci. Data. 14, 4743–4755. https://doi.org/10.5194/essd-14-4743-2022.

Kloiber, S.M., Brezonik, P.L., Olmanson, L.G., Bauer, M.E., 2002. A procedure for regional lake water clarity assessment using Landsat multispectral data. Remote Sens. Environ. 82, 38–47. https://doi.org/10.1016/S0034-4257(02)00022-6.

Kuhn, C., de Matos Valerio, A., Ward, N., Loken, L., Sawakuchi, H.O., Kampel, M., Richey, J., Stadler, P., Crawford, J., Striegl, R., 2019. Performance of Landsat-8 and Sentinel-2 surface reflectance products for river remote sensing retrievals of chlorophyll-a and turbidity. Remote Sens. Environ. 224, 104–118. https://doi.org/10.1016/j.rse.2019.01.023.

Kutser, T., Pierson, D.C., Kallio, K.Y., Reinart, A., Sobek, S., 2005. Mapping lake CDOM by satellite remote sensing. Remote Sens. Environ. 94, 535–540. https://doi.org/10.1016/j.rse.2004.11.009.

Lee, Z., Carder, K.L., Arnone, R.A., 2002. Deriving inherent optical properties from water color: a multiband quasi-analytical algorithm for optically deep waters. Appl. Opt. 41, 5755–5772. https://doi.org/10.1364/AO.41.005755.

Lee, Z., Wei, J., Voss, K., Lewis, M., Bricaud, A., Huot, Y., 2015. Hyperspectral absorption coefficient of "pure" seawater in the range of 350–550 nm inverted from remote sensing reflectance. Appl. Opt. 54, 546–558. https://doi.org/10.1364/AO.54.000546.

Lehner, B., Liermann, C.R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., 2011. High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. Front. Ecol. Environ. 9, 494–502. https://doi.org/10.1890/100125.

Li, D., Lu, X., Overeem, I., Walling, D.E., Syvitski, J., Kettner, A.J., Bookhagen, B., Zhou, Y., Zhang, T., 2021. Exceptional increases in fluvial sediment fluxes in a warmer and wetter High Mountain Asia. Science. 374, 599–603. https://doi.org/10.1126/science.abi9649.

Li, J., Yu, Q., Tian, Y.Q., Becker, B.L., 2017. Remote sensing estimation of colored dissolved organic matter (CDOM) in optically shallow waters. ISPRS J. Photogramm. Remote Sens. 128, 98–110. https://doi.org/10.1016/j.isprsjprs.2017.03.015.

Li, J., Wang, G., Song, C., Sun, S., Ma, J., Wang, Y., Guo, L., Li, D., 2024. Recent intensified erosion and massive sediment deposition in Tibetan Plateau rivers. Nat. Commun. 15, 722. https://doi.org/10.1038/s41467-024-44982-0.

Lin, J., Lyu, H., Miao, S., Pan, Y., Wu, Z., Li, Y., Wang, Q., 2018. A two-step approach to mapping particulate organic carbon (POC) in inland water using OLCI images. Ecol. Indic. 90, 502–512. https://doi.org/10.1016/j.ecolind.2018.03.044.

Lin, P., Pan, M., Wood, E.F., Yamazaki, D., Allen, G.H., 2021. A new vector-based global river network dataset accounting for variable drainage density. Sci. Data. 8, 28. https://doi.org/10.1038/s41597-021-00819-9.

Liu, D., Bai, Y., He, X., Pan, D., Chen, C.-T.-A., Li, T., Xu, Y., Gong, C., Zhang, L., 2019. Satellite-derived particulate organic carbon flux in the Changjiang River through different stages of the Three Gorges Dam. Remote Sens. Environ. 223, 154–165. https://doi.org/10.1016/j.rse.2019.01.012.

Liu, H., Li, Q., Bai, Y., Yang, C., Wang, J., Zhou, Q., Hu, S., Shi, T., Liao, X., Wu, G., 2021. Improving satellite retrieval of oceanic particulate organic carbon concentrations using machine learning methods. Remote Sens. Environ. 256, 112316. https://doi.org/10.1016/j.rse.2021.112316.

Liu, Z., Mo, S., Liu, F., Korshin, G., Yan, M., 2024. A novel algorithm to extrapolate ultraviolet absorption of chromophoric dissolved organic matter from remote sensing ocean color. ISPRS J. Photogramm. Remote Sens. 214, 104–118. https://doi.org/10.1016/j.isprsjprs.2024.05.026.

Maciel, D.A., Pahlevan, N., Barbosa, C.C., Martins, V.S., Smith, B., O'Shea, R.E., Balasubramanian, S.V., Saranathan, A.M., Novo, E.M., 2023. Towards global long-term water transparency products from the Landsat archive. Remote Sens. Environ. 299, 113889. https://doi.org/10.1016/j.rse.2023.113889.

Mann, P.J., Spencer, R.G., Hernes, P.J., Six, J., Aiken, G.R., Tank, S.E., McClelland, J.W., Butler, K.D., Dyda, R.Y., Holmes, R.M., 2016. Pan-Arctic trends in terrestrial dissolved organic matter from optical measurements. Front. Earth Sci. 4, 25. https://doi.org/10.3389/feart.2016.00025.

Martineau, C., Vincent, W.F., Frenette, J.J., Dodson, J.J., 2004. Primary consumers and particulate organic matter: isotopic evidence of strong selectivity in the estuarine transition zone. Limnol. Oceanogr. 49, 1679–1686. https://doi.org/10.4319/lo.2004.49.5.1679.

Marwick, T.R., Tamooh, F., Teodoru, C.R., Borges, A.V., Darchambeau, F., Bouillon, S., 2015. The age of river-transported carbon: a global perspective. Global Biogeochem. Cycles. 29, 122–137. https://doi.org/10.1002/2014GB004911.

Matsuoka, A., Babin, M., Vonk, J.E., 2022. Decadal trends in the release of terrigenous organic carbon to the Mackenzie delta (Canadian Arctic) using satellite ocean color data (1998–2019). Remote Sens. Environ. 283, 113322. https://doi.org/10.1016/j.rse.2022.113322.

McClelland, J., Holmes, R., Peterson, B., Raymond, P., Striegl, R., Zhulidov, A., Zimov, S., Zimov, N., Tank, S., Spencer, R., 2016. Particulate organic carbon and nitrogen export from major Arctic rivers. Global Biogeochem. Cycles. 30, 629–643. https://doi.org/10.1002/2015GB005351.

Medina-Lopez, E., 2020. Machine learning and the end of atmospheric corrections: A comparison between high-resolution sea surface salinity in coastal areas from top and bottom of atmosphere sentinel-2 imagery. Remote Sens. 12, 2924. https://doi.org/10.3390/rs12182924.

Morley, S.K., Brito, T.V., Welling, D.T., 2018. Measures of model performance based on the log accuracy ratio. Space Weather. 16, 69–88. https://doi.org/10.1002/2017SW001669.

Olmanson, L.G., Bauer, M.E., Brezonik, P.L., 2008. A 20-year Landsat water clarity census of Minnesota's 10,000 lakes. Remote Sens. Environ. 112, 4086–4097. https://doi.org/10.1016/j.rse.2007.12.013.

O'Reilly, J.E., Maritorena, S., Mitchell, B.G., Siegel, D.A., Carder, K.L., Garver, S.A., Kahru, M., McClain, C., 1998. Ocean color chlorophyll algorithms for SeaWiFS. J. Geophys Res. Oceans. 103, 24937–24953. https://doi.org/10.1029/98JC02160.

Pahlevan, N., Mangin, A., Balasubramanian, S.V., Smith, B., Alikas, K., Arai, K., Barbosa, C., Bélanger, S., Binding, C., Bresciani, M., 2021. ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters. Remote Sens. Environ. 258, 112366. https://doi.org/10.1016/j.rse.2021.112366.

Pandey, U., Pandey, J., 2013. Impact of DOC trends resulting from changing climatic extremes and atmospheric deposition chemistry on periphyton community of a freshwater tropical lake of India. Biogeochemistry. 112, 537–553. https://doi.org/10.1007/s10533-012-9747-7.

Raymond, P.A., Cole, J.J., 2003. Increase in the export of alkalinity from North America's largest river. Science. 301, 88–91. https://doi.org/10.1126/science.1083788.

Raymond, P.A., Oh, N.-H., Turner, R.E., Broussard, W., 2008. Anthropogenically enhanced fluxes of water and carbon from the Mississippi River. Nature. 451, 449–452. https://doi.org/10.1038/nature06505.

Regnier, P., Friedlingstein, P., Ciais, P., Mackenzie, F.T., Gruber, N., Janssens, I.A., Laruelle, G.G., Lauerwald, R., Luyssaert, S., Andersson, A.J., 2013. Anthropogenic perturbation of the carbon fluxes from land to ocean. Nat. Geosci. 6, 597–607. https://doi.org/10.1038/ngeo1830.

Regnier, P., Resplandy, L., Najjar, R.G., Ciais, P., 2022. The land-to-ocean loops of the global carbon cycle. Nature. 603, 401–410. https://doi.org/10.1038/s41586-021-04339-9.

Ren, W., Tian, H., Cai, W.J., Lohrenz, S.E., Hopkinson, C.S., Huang, W.J., Yang, J., Tao, B., Pan, S., He, R., 2016. Century-long increasing trend and variability of dissolved organic carbon export from the Mississippi River basin driven by natural and anthropogenic forcing. Global Biogeochem. Cycles. 30, 1288–1299. https://doi.org/10.1002/2016GB005395.

Repasch, M., Scheingross, J.S., Hovius, N., Lupker, M., Wittmann, H., Haghipour, N., Gröcke, D.R., Orfeo, O., Eglinton, T.I., Sachse, D., 2021. Fluvial organic carbon cycling regulated by sediment transit time and mineral protection. Nat. Geosci. 14, 842–848. https://doi.org/10.1038/s41561-021-00845-7.

Ross, M.R., Topp, S.N., Appling, A.P., Yang, X., Kuhn, C., Butman, D., Simard, M., Pavelsky, T.M., 2019. AquaSat: A data set to enable remote sensing of water quality for inland waters. Water Resour. Res. 55, 10012–10025. https://doi.org/10.1029/2019WR024883.

Roy, D.P., Kovalskyy, V., Zhang, H., Vermote, E.F., Yan, L., Kumar, S., Egorov, A., 2016. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. Remote Sens. Environ. 185, 57–70. https://doi.org/10.1016/j.rse.2015.12.024.

Schmidt, G., Jenkerson, C.B., Masek, J., Vermote, E., Gao, F., 2013. Landsat ecosystem disturbance adaptive processing system (LEDAPS) algorithm description. US Geological Survey.

Shen, Z., Rosenheim, B.E., Törnqvist, T.E., Lang, A., 2021. Engineered continental-scale rivers can drive changes in the carbon cycle. AGU Adv. 2, e2020AV000273. https://doi.org/10.1029/2020AV000273.

Spencer, R.G., Butler, K.D., Aiken, G.R., 2012. Dissolved organic carbon and chromophoric dissolved organic matter properties of rivers in the USA. J. Geophys. Res.: Biogeosci. 117. https://doi.org/10.1029/2011JG001928.

Spencer, R.G., Raymond, P.A., 2024. Riverine DOM. In: Biogeochemistry of Marine Dissolved Organic Matter. Elsevier, pp. 657–691.

Tian, S., Guo, H., Xu, W., Zhu, X., Wang, B., Zeng, Q., Mai, Y., Huang, J.J., 2023. Remote sensing retrieval of inland water quality parameters using Sentinel-2 and multiple machine learning algorithms. Environ. Sci. Pollut. Res. 30, 18617–18630. https://doi.org/10.1007/s11356-022-23431-9.

Toming, K., Kotta, J., Uuemaa, E., Sobek, S., Kutser, T., Tranvik, L.J., 2020. Predicting lake dissolved organic carbon at a global scale. Sci. Rep. 10, 8471. https://doi.org/10.1038/s41598-020-65010-3.

Virro, H., Amatulli, G., Kmoch, A., Shen, L., Uuemaa, E., 2021. GRQA: global river water quality archive. Earth Syst. Sci. Data Discuss. 2021, 1–30. https://doi.org/10.5194/essd-13-5483-2021.

Wang, Z., Li, R., Chen, Z., Yao, Q., Gao, B., Xu, M., Yang, L., Li, M., Zhou, C., 2022b. The estimation of hourly PM2. 5 concentrations across China based on a Spatial and Temporal Weighted Continuous Deep Neural Network (STWC-DNN). ISPRS J. Photogramm. Remote Sens. 190, 38–55. https://doi.org/10.1016/j.isprsjprs.2022.05.011.

Wang, F., Wang, Y., Chen, Y., Liu, K., 2020. Remote sensing approach for the estimation of particulate organic carbon in coastal waters based on suspended particulate concentration and particle median size. Mar. Pollut. Bull. 158, 111382. https://doi.org/10.1016/j.marpolbul.2020.111382.

Wang, X., Wen, Z., Liu, G., Tao, H., Song, K., 2022a. Remote estimates of total suspended matter in China's main estuaries using Landsat images and a weight random forest model. ISPRS J. Photogramm. Remote Sens. 183, 94–110. https://doi.org/10.1016/j.isprsjprs.2021.11.001.

Wang, C., Zhou, C., Zhou, X., Duan, M., Yan, Y., Wang, J., Wang, L., Jia, K., Sun, Y., Wang, D., 2025. A universal method to recognize global big rivers estuarine turbidity maximum from remote sensing. ISPRS J. Photogramm. Remote Sens. 220, 509–523. https://doi.org/10.1016/j.isprsjprs.2025.01.002.

Wulder, M.A., Roy, D.P., Radeloff, V.C., Loveland, T.R., Anderson, M.C., Johnson, D.M., Healey, S., Zhu, Z., Scambos, T.A., Pahlevan, N., 2022. Fifty years of Landsat science and impacts. Remote Sens. Environ. 280, 113195. https://doi.org/10.1016/j.rse.2022.113195.

Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int. J. Remote Sens. 27, 3025–3033. https://doi.org/10.1080/01431160600589179.

Xu, J., Fang, C., Gao, D., Zhang, H., Gao, C., Xu, Z., Wang, Y., 2018. Optical models for remote sensing of chromophoric dissolved organic matter (CDOM) absorption in Poyang Lake. ISPRS J. Photogramm. Remote Sens. 142, 124–136. https://doi.org/10.1016/j.isprsjprs.2018.06.004.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P.D., Allen, G.H., Pavelsky, T.M., 2019. MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset. Water Resour. Res. 55, 5053–5073. https://doi.org/10.1029/2019WR024873.

Yin, S., Gao, G., Li, Y., Xu, Y.J., Turner, R.E., Ran, L., Wang, X., Fu, B., 2023. Long-term trends of streamflow, sediment load and nutrient fluxes from the Mississippi River Basin: Impacts of climate change and human activities. J. Hydrol. 616, 128822. https://doi.org/10.1016/j.jhydrol.2022.128822.

Zhang, T., Li, D., East, A.E., Walling, D.E., Lane, S., Overeem, I., Beylich, A.A., Koppes, M., Lu, X., 2022. Warming-driven erosion and sediment transport in cold regions. Nat. Rev. Earth & Environ. 3, 832–851. https://doi.org/10.1038/s43017-022-00362-0.

Zhang, T., Li, D., East, A.E., Kettner, A.J., Best, J., Ni, J., Lu, X., 2023. Shifted sediment-transport regimes by climate change and amplified hydrological variability in cryosphere-fed rivers. Sci. Adv. 9, eadi5019. https://doi.org/10.1126/sciadv.adi5019.

Zhang, Y., van Dijk, M.A., Liu, M., Zhu, G., Qin, B., 2009. The contribution of phytoplankton degradation to chromophoric dissolved organic matter (CDOM) in eutrophic shallow lakes: field and experimental evidence. Water Res. 43, 4685–4697. https://doi.org/10.1016/j.watres.2009.07.024.

Zhang, Z., Zhang, H., Jin, Y., Guo, H., Tian, S., Huang, J.J., Zhu, X., 2024. Analysing the spatiotemporal variation and influencing factors of Lake Chaohu's CDOM over the past 40 years using machine learning. Ecohydrology. 17, e2639. https://doi.org/10.1002/eco.2639.

Zhou, H., Liu, S., Hu, S., Mo, X., 2021. Retrieving dynamics of the surface water extent in the upper reach of Yellow River. Sci. Total Environ. 800, 149348. https://doi.org/10.1016/j.scitotenv.2021.149348.