Remote Sensing of Environment 339 (2026) 115388

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Towards a global assessment of coastal dissolved organic carbon

K. Toming [a,*], G. Kulk [b,c], R. Quast [d], R. Shevchuk [d], T. Kutser [a]

[a] Estonian Marine Institute, University of Tartu, Tallinn, Estonia
[b] Marine Processes and Observations, Plymouth Marine Laboratory, Plymouth, United Kingdom
[c] National Centre of Earth Observation, Plymouth Marine Laboratory, Plymouth, United Kingdom
[d] Brockmann Consult GmbH, Hamburg, Germany

ARTICLE INFO

Edited by: Menghua Wang

Keywords:
Dissolved organic carbon
Coastal waters
Remote sensing
Ocean colour
Global carbon cycle
Ocean colour climate change initiative
European space agency

ABSTRACT

Dissolved organic carbon (DOC) plays a crucial role in ecological and biogeochemical processes. Many regional satellite algorithms for DOC in coastal waters have been developed. However, there is currently no global algorithm capable of addressing the variability and complexity of DOC dynamics in coastal waters. We address this gap by developing a global DOC satellite retrieval algorithm for coastal waters by using daily, 4-km resolution data from the European Space Agency (ESA) Ocean Colour Climate Change Initiative (OC-CCI) from 1997 to 2023, combined with sea surface salinity (GLORYS12v1 database) and temperature (ESA SST-CCI, version 3.0 database). For model development, we matched these satellite datasets with *in situ* DOC concentration data from the CoastDOM v1 database. A pairwise correlation between different variables was used to identify the most relevant predictor variables of coastal DOC concentrations. Several statistical methods, including multiple linear regression (MLR), random forest regression (RF), and extreme gradient boosting (XGBoost), were tested. The best performance was achieved by a RF model using sea surface salinity and temperature, the remote sensing reflectance at 560 nm and total absorption at 412 nm, with cross-validation metrics of $R^2 = 0.91$, RMSE = 0.52 mg C L^{-1}, and MAPE = 13.01%, and independent validation on unseen data giving $R^2 = 0.83$, RMSE = 0.64 mg C L^{-1}, and MAPE = 23.15%. Although the developed algorithm showed high performance, the relatively coarse resolution of OC-CCI poses challenges, as it may fail to resolve sharp DOC gradients in dynamic coastal waters such as river plumes and estuaries, potentially reducing accuracy in those areas. Still, OC-CCI offers climate-quality data for a longer period of time compared to individual ocean-colour sensors. Expanding *in situ* observations, especially in underrepresented areas, will further enhance model accuracy and applicability. The results help to understand carbon dynamics in coastal ecosystems and offer a robust tool for satellite-based assessments of DOC in coastal waters globally.

1. Introduction

Knowledge of the coastal carbon cycle and its budget is key in understanding and predicting the impact of climate change. Coastal waters are dynamic and productive areas between terrestrial and ocean systems, and regulate nutrient flows, support marine biodiversity, and provide important ecosystem services, such as fisheries and carbon sequestration (Bauer et al., 2013; Campbell et al., 2022; Dunne et al., 2007; Fennel et al., 2019; Gattuso et al., 1998; Muller-Karger et al., 2005). Although coastal waters make up a relatively small portion of the total area of the ocean (7–11%), they have a considerable impact on the global carbon cycle and play a major role in climate regulation. These regions are also considered critical in achieving emission reductions necessary for fulfilling a variety of Sustainable Development Goals and the Paris Agreement targets (Hoegh-Guldberg et al., 2019). Unfortunately, coastal environments are strongly influenced by human activities, including urbanisation, pollution, and climate change, which have significant effects on the biogeochemical processes of coastal waters (Regnier et al., 2013, 2022).

One of the key components of coastal biogeochemistry is Dissolved Organic Carbon (DOC) (Hansell and Carlson, 2013). DOC serves as an essential component of the marine as well as the global carbon cycle, and improving global carbon budget estimates and predicting ecosystem responses to climate change depends on an awareness of the spatial and temporal variability of DOC in coastal waters. Additionally, DOC plays a crucial role in microbial food webs (Dafner and Wangersky, 2002), reduces light penetration, which limits photosynthesis by phytoplankton and macro-algae (Martin et al., 2021), and its decomposition releases CO_2,

* Corresponding author.
E-mail address: kaire.toming.001@ut.ee (K. Toming).

https://doi.org/10.1016/j.rse.2026.115388
Received 27 May 2025; Received in revised form 27 November 2025; Accepted 23 March 2026
Available online 8 April 2026

therefore contributing to seawater acidification (Semiletov et al., 2016).

Although much is known about carbon pools and fluxes in the open ocean (Brewin et al., 2021; Kulk et al., 2020; Siegel et al., 2023), numerous uncertainties concerning the coastal carbon cycle have remained, both regionally and globally. Barrón and Duarte (2015) estimated DOC concentration in coastal waters based on 3510 *in situ* measurements worldwide, reporting a mean depth-averaged DOC concentration of 285.7 ± 5.7 μmol C L^{-1} (mean ± standard error), but acknowledging that the *in situ* measurements were unevenly distributed, with most measurements collected in the Northern Hemisphere and fewer studies on the coasts of Africa, South America, Australia, and Asia. Lønborg et al. (2024), in CoastDOM v1, compiled *in situ* observations from 1978 to 2022 and across all continents, with a total of 62,338 individual data points. They showed that the average DOC concentration in coastal waters was 182 ± 314 μmol C L^{-1} (mean ± standard deviation). However, most data are from the Northern Hemisphere, with gaps from the Southern Hemisphere showing that traditional field measurements, while accurate, are often limited in spatial coverage as well as in temporal resolution.

Satellite remote sensing can provide high spatial and temporal resolution data and, therefore, support systematic carbon monitoring at local, regional, and global levels (Campbell et al., 2022). Satellite data have been used to monitor ocean carbon pools and fluxes by direct and indirect approaches, with some combining both (Brewin et al., 2021). Direct methods involve determining if a certain carbon fraction has an effect on the water colour (visible electromagnetic radiation) that the remote sensing instruments are detecting. Indirect techniques, on the other hand, relate the carbon pool or flux to "visible" proxies like retrieved variables of ocean colour, thermal radiometry, microwave radiometry, and satellite altimetry. These methods can be mechanistic, empirical, or statistical (Brewin et al., 2021). Most of the DOC pool does not have any effect on water colour and consequently cannot be mapped with remote sensing directly. Still, a small component of the DOC pool is chromophoric, which can be directly monitored by ocean-colour remote sensing. In many cases, there is a correlation between the concentration of DOC and its optically active component, Coloured Dissolved Organic Matter (CDOM) (Mannino et al., 2008), and it allows the retrieval of DOC from CDOM absorption or from remote sensing reflectance ($R_{rs(\lambda)}$). Therefore, several models have been developed to estimate DOC in coastal waters at a local scale by using ocean-colour remote-sensing data and its relationship with CDOM (Cao and Tzortziou, 2024; Liu et al., 2019; Mannino et al., 2008) or $R_{rs(\lambda)}$ (Cao and Tzortziou, 2021).

Strong correlations between CDOM absorption and DOC have been observed in many coastal regions (Ferrari et al., 1996; Fichot and Benner, 2012; Vodacek et al., 1997). Unfortunately, the relationship is often regionally and seasonally highly variable due to terrestrial inflows, land use, and the proportion of autochthonous DOC in the overall DOC pool (Harvey et al., 2015; Kowalczuk et al., 2010). This makes it challenging to use a single model at a global scale. Several regional DOC algorithms can be found in the literature that primarily focus on relating satellite-derived $R_{rs(\lambda)}$ to the absorption of coloured dissolved organic matter (a_{CDOM}), and subsequently deriving DOC *via* empirical relationships with a_{CDOM} (Cao et al., 2018; Cao and Tzortziou, 2024; Cherukuru et al., 2016, 2021; le Fouest et al., 2018; Liu et al., 2014; Mannino et al., 2008, 2016; Pan and Wong, 2015; Qiong et al., 2011; Tehrani et al., 2013; Tuzcu Kokal et al., 2024; Zhang et al., 2024), but no global algorithm that would be able to address the variability and complexity of DOC dynamics in coastal waters on a broader scale is currently available (also see section 4 Discussion).

In the present study, we aim to address this gap by developing a global DOC retrieval algorithm for coastal waters by using daily 4-km resolution data from the European Space Agency (ESA) Ocean Colour Climate Change Initiative (OC-CCI; Sathyendranath et al., 2023) from 1997 to 2023. Good spatial resolution data are needed in capturing the complexity of coastal waters, where significant gradients in DOC can occur over small spatial and short temporal scales. Although the 4-km resolution of OC-CCI may not detect all fine-scale spatial features, it still provides a robust framework for developing a global algorithm that balances spatial coverage and computational feasibility. In this study, we combine optical and environmental variables to advance DOC modeling in coastal waters beyond a region-specific approach. Since *in situ* measurements remain essential for developing and validating remote sensing models, the *in situ* DOC concentration data from the CoastDOM database (Lønborg et al., 2024) is used. Using the *in situ* and remote sensing observations, we tested three different methods, namely Multiple Linear Regression (MLR), Random Forest regression (RF), and Extreme Gradient Boosting regression (XGBoost), to develop a satellite-based algorithm for estimating DOC concentrations in coastal waters at the global scale.

2. Materials and methods

2.1. Study area and in situ data

The study focuses on global coastal waters (Fig. 1), which for the *in situ* data are defined according to the Coastal and Marine Ecological Classification Standard (CMECS, 2012), as areas with bottom depths less than 200 m. *In situ* data were used to develop the satellite-based model and validate the estimates of coastal DOC concentrations. The source of the *in situ* data of DOC concentration, as well as SST and SSS, was the CoastDOM database by Lønborg et al. (2024). *In situ* DOC data from the years between 1997 and 2023 of up to 1-m depth were used (in total 18,810 measurements) and *in situ* data located in areas with depths less than 200 m were included in further analysis (in total 17,833 measurements).

2.2. Satellite data

OC-CCI v6.0 data (1997–2023) were used to obtain daily ocean-colour remote sensing data (Sathyendranath et al., 2023). OC-CCI is a merged and bias-corrected time series product, that includes ocean-colour observations from MERIS, MODIS-AQUA, SeaWiFS, VIIRS and Sentinel-3 OLCI (with v6.0 being band-shifted to MERIS). In addition to $R_{rs(\lambda)}$, OC-CCI products include inherent optical properties (IOPs), like absorption of dissolved and detrital materials (a_{dg}), phytoplankton absorption (a_{ph}), total absorption (a_{tot}) and particulate backscattering (b_{bp}) at six wavelengths (412, 443, 490, 510, 560 and 665 nm). It also includes the light diffuse attenuation coefficient at 490 nm (K_{d_490}), and chlorophyll-*a* concentration (chl-a). The K_{d_490} can be related to the absorption and light attenuation by all water constituents, including CDOM (Paavel et al., 2011). At the same time, chl-a serves as an indicator of phytoplankton biomass and associated autochthonous DOC production in coastal waters. Therefore, all these available variables were considered in algorithm development. Additionally, the OC-CCI dataset provides pixel-by-pixel uncertainty estimates for each wavelength and information on water type classifications (Jackson et al., 2017). For the satellite data, coastal waters were defined by OC-CCI water classes 12–14 (Jackson et al., 2017). Pixels from classes 1–11 (representative of open ocean environments) were removed from further analyses.

In addition, SST from ESA SST-CCI (daily temporal resolution; ~5 km spatial resolution; SST-CCI) and SSS from GLORYS12V1 (daily temporal resolution; ~8 km spatial resolution; SSS-GLORYS) were used for algorithm development (Embury et al., 2024; Good and Embury, 2024; Jean-Michel et al., 2021). Salinity indicates freshwater influence, which is closely tied to the distribution and variability of DOC and CDOM in coastal waters (Bowers et al., 2000; Fichot and Benner, 2011), while temperature affects microbial degradation and primary production, both of which influence DOC dynamics in coastal waters. Therefore, in addition to the OC-CCI products, SST and SSS were also considered in the algorithm development. To ensure consistency across datasets, SST-CCI (~5 km) and SSS-GLORYS (~8 km) were resampled using nearest-neighbour interpolation to align spatially with the 4-km

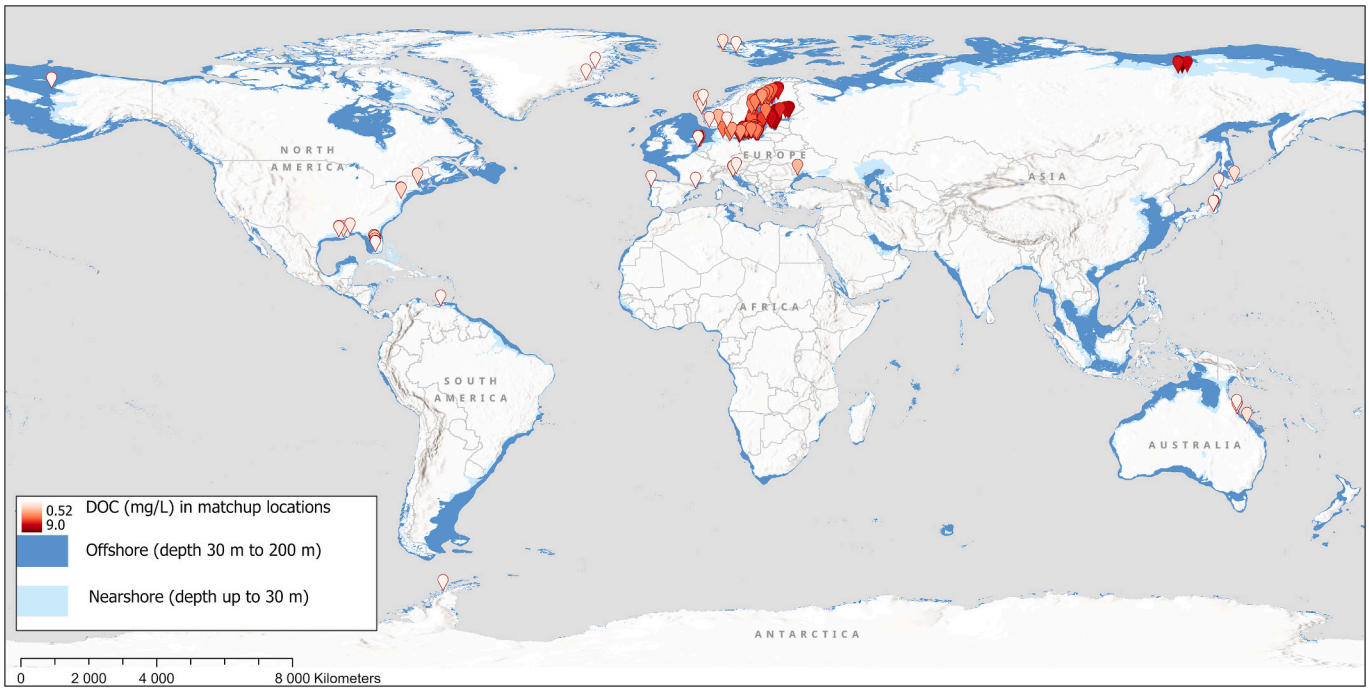

Fig. 1. Spatial distribution of match-up data points with *in situ* Dissolved Organic Carbon (DOC, mg C L^{-1}) concentrations.

grid of the OC-CCI data. SST-CCI and SSS-GLORYS data correlated strongly with *in situ* temperature and salinity measurements ($r = 0.97$ and $r = 0.99$, respectively; Fig. S1). The source of *in situ* data of SSS and SST was CoastDOM database by Lønborg et al., 2024 (see section 2.1). While the SSS-GLORYS dataset showed a strong overall correlation with *in situ* SSS, higher variance was observed at higher salinity levels (Fig. S1). This could reflect known regional biases in the GLORYS12v1 product, with higher biases observed in the Arctic and western tropical Pacific Ocean, near the Amazon River plume and around Indonesia (*i.e.*, in comparison with data from the World Ocean Atlas; Drévillon et al., 2023). Yet, the global average bias in SSS-GLORYS remains close to zero (Drévillon et al., 2023). The higher error in SSS in specific regions could lead to higher uncertainty in the predicted DOC, but we note that error in other predictor variables, such as those from ocean colour data, are likely to dominate. To account for these and other sources of errors in OC-CCI, SST-CCI, and SSS-GLORYS datasets, uncertainties were propagated through the DOC retrieval algorithm using Monte Carlo simulations (see Section 2.7).

2.3. Match-up analyses

The nearest neighbour method from Python's xarray package (Hoyer and Hamman, 2017) was used to match satellite data with *in situ* DOC measurements. For quality control, a 3 × 3 pixel box around the nearest pixel was extracted, and at least 4 valid observations per 9 pixels and a coefficient of variation less than 0.2 were needed to pass the quality control. Outliers were identified and removed using the Interquartile Range (IQR) method with a 1.5 threshold.

Strict quality control was applied during model training to ensure that all candidate input variables met the same filtering criteria. Only common match-up data points where all variables passed the quality control were retained ensuring consistency across all variables during model development.

After identifying the best-performing model, an independent validation dataset was compiled. In this step, the same quality control thresholds were used, but applied only to the variables used in the final model. As a result, it was possible to include additional match-up data that had been previously excluded due to other variables having no match-up or failing quality control. All variables in the validation dataset needed to pass the same quality control as the training set, including the coefficient of variation threshold and IQR-based outlier filtering. The resulting independent validation dataset comprised 452 match-ups and provided broad spatial and seasonal coverage. Importantly, the training and validation datasets were non-overlapping, and the training set was not re-adjusted after regressor selection prioritising maximum data quality for training and broader coverage for validation. Frequency histograms of the validation dataset are presented in Fig. S2. These new match-up data were used as an independent validation dataset to assess the performance of the final model.

2.4. Feature selection

Band combinations are usually more effective compared to single bands in detecting the optical properties of water. Therefore, fifteen formulas (Toming et al., 2024) based on two- or three-band combinations (Table 1) were tested with various OC-CCI spectral bands (resulted in a total of 1170 band combinations).

Table 1
Formulas based on two- or three-band combinations that were tested with various OC-CCI spectral bands. B represents the water-leaving reflectance and index a, b, or c denotes different OC-CCI spectral bands (6 bands in different options).

Formula
$B_a + B_b$
$B_a - B_b$
B_a / B_b
$B_a * B_b$
$B_a + B_b + B_c$
$B_a + B_b * B_c$
$(B_a + B_b) * B_c$
$(B_a - B_b) * B_c$
$(B_a + B_b) / B_c$
$B_a * B_b / B_c$
$(B_a - B_b) / (B_a + B_b)$
$(B_a / B_b) * (B_a / B_b)$
$B_a / B_b - B_a / B_c$
$B_a - (B_b + B_c) / 2$
$B_a / (B_b + B_c)$

Table 2
Combinations of regressors used in the Dissolved Organic Carbon (DOC) satellite-retrieval models. Abbreviations: Remote sensing reflectance (Rrs(λ)), Inherent Optical Properties (IOP), Sea Surface Salinity (SSS), and Sea Surface Temperature (SST).

Number of Regressors	Combinations of regressors
2 regressors	SSS + SST
	SSS + $R_{rs(\lambda)}$
	SSS + $R_{rs(\lambda)}$ band combination
	SSS + IOP
	SST + $R_{rs(\lambda)}$
	SST + $R_{rs(\lambda)}$ band combination
	SST + IOP
	IOP + $R_{rs(\lambda)}$
	IOP + $R_{rs(\lambda)}$ band combination
3 regressors	SSS + SST + $R_{rs(\lambda)}$
	SSS + SST + $R_{rs(\lambda)}$ band combination
	SSS + SST + IOP
	SSS + $R_{rs(\lambda)}$ + IOP
	SSS + $R_{rs(\lambda)}$ band combination + IOP
	SST + $R_{rs(\lambda)}$ + IOP
	SST + $R_{rs(\lambda)}$ band combination + IOP
4 regressors	SSS + SST + $R_{rs(\lambda)}$ + IOP
	SSS + SST + $R_{rs(\lambda)}$ band combination + IOP

Before model development, a visual inspection of pair-wise correlation between different variables (see section 3.2) was used to identify the most relevant predictor variables of coastal DOC concentrations.

Variables and band combinations with moderate to strong correlations with DOC concentrations were considered for further model development if the correlation coefficient was $r \geq 0.7$ for $R_{rs(\lambda)}$ band combinations, and $r \geq 0.5$ for absorption-related variables (including chl-a and K_{d_490}). Additionally, SSS, SST and all 6 $R_{rs(\lambda)}$ and a_{dg} were included for further model development.

2.5. Model development

Using the selected features (see section 3.2), MLR, RF and XGBoost were used to develop an algorithm for predicting global DOC concentrations in coastal waters. MLR is an easily interpretable regression model that estimates the linear relationship between a dependent variable and multiple independent variables. RF is an ensemble learning method that creates multiple decision trees to handle complex nonlinear relationships and make accurate predictions (Breiman, 2001). XGBoost is also an ensemble learning method that effectively handles complex relationships, and it is based on gradient boosting that enhances model performance through iterative optimisation (Chen and Guestrin, 2016). All three models have previously been used in environmental studies, including for predicting the DOC concentrations in the global open ocean (Laine et al., 2024).

Due to the strong correlations between $R_{rs(\lambda)}$ and IOPs, careful manual selection of input variables was done to avoid issues related to multicollinearity, particularly for the MLR model. For the RF and XGBoost models, multicollinearity is less critical. However, avoiding the inclusion of too many variables or highly collinear ones can help to optimise model performance, avoid overfitting and reduce unnecessary complexity. Consequently, the models were configured to include either 2, 3 or 4 regressors (Table 2).

The Scikit-learn package, version 1.5.2 (Pedregosa et al., 2012), was used to perform development and accuracy evaluation of the MLR and RF models and the XGBoost package, version 2.1.2 (Chen and Guestrin, 2016), was used for the XGBoost model. All predictor variables were standardised using z-score standardisation, and model development and evaluation were done using 5-fold cross-validation (CV). For hyperparameter optimisation of the RF and XGBoost models, the Optuna package, version 4.1.0 (Akiba et al., 2019), was employed. All computations were conducted in Python 3.10.

2.6. Quality assessment

The selection of the best model for retrieving DOC in coastal waters was based on multiple statistical accuracy metrics, including the coefficient of determination (R^2), the root-mean-squared error (RMSE) and the mean absolute percentage error (MAPE). R^2 is the squared correlation between the measured and predicted values (the closer to 1, the

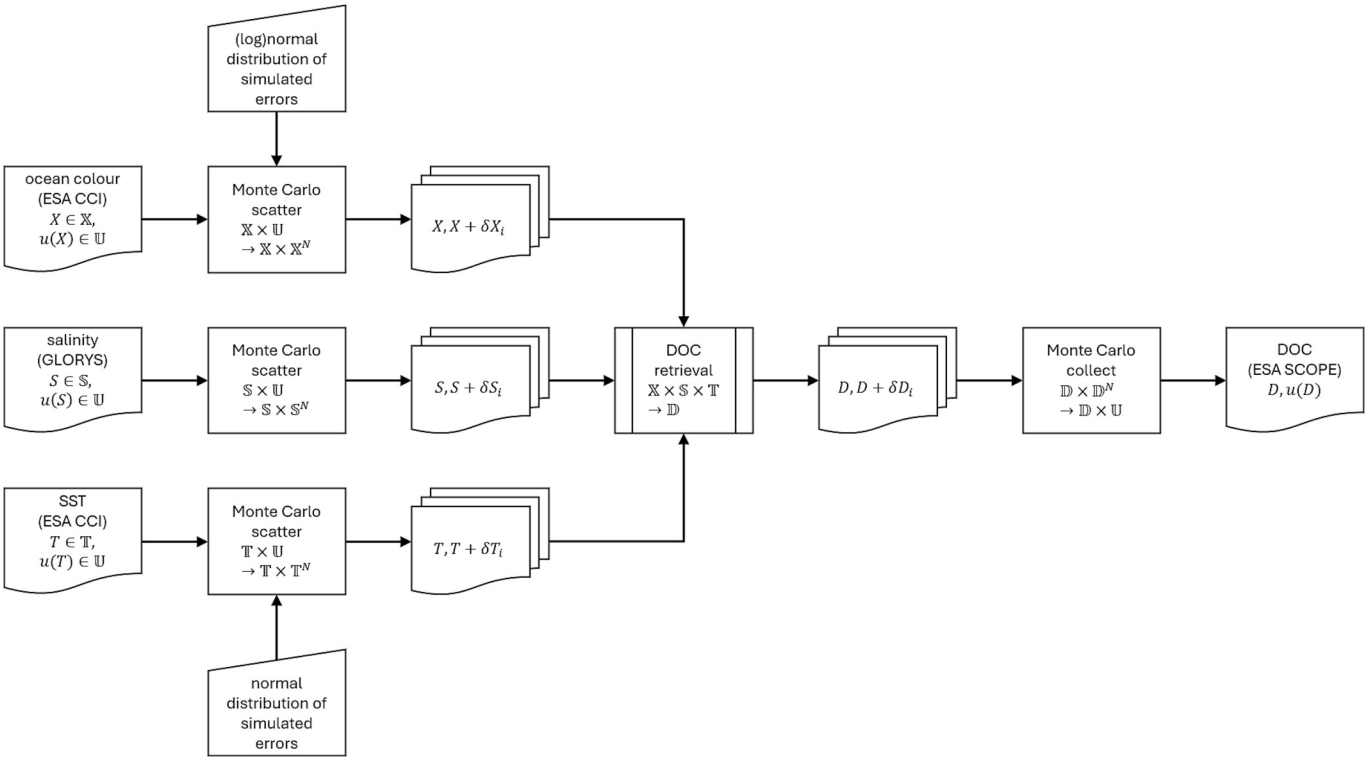

Fig. 2. Illustration of uncertainty propagation by Monte Carlo simulation of errors domains of OC-CCI, SST-CCI and SSS-GLORYS data denoted by $\mathbb{X}$, $\mathbb{T}$ and $\mathbb{S}$, respectively; domain of uncertainties denoted by $\mathbb{U}$; range of estimated DOC denoted by $\mathbb{D}$.

better the model) and it was calculated using Eq. 1:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \quad (1)$$

where $\hat{y}$ is the predicted value, y is the measured value, $\bar{y}$ is the mean value of measured y values, and n is the number of measurements. RMSE is the mean difference between the measured and predicted values (the lower the value, the better the model) and it was calculated using Eq. 2:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y} - y_i)^2} \quad (2)$$

where $\hat{y}$ is the predicted value, y is the measured value, and n is the number of measurements. MAPE is the mean absolute percentage error (the lower the value, the better the model) and it was calculated using Eq. 3:

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (3)$$

where $\hat{y}$ is the predicted value, y is the measured value, and n is the number of measurements.

The performance metrics were then normalised to a [0,1] or [1,0] range, depending on the metric. The R^2 was normalised so that higher values corresponded to better performance, while the other metrics were normalised so that lower values, which indicate better model performance, received higher normalised scores. Each model's final score was calculated by summing the normalised values, and the model with the highest score was selected as the best one.

After training and identifying the optimal model, an extended match-up dataset was created (see Section 2.3). This extended dataset was used as an independent validation dataset to evaluate the performance of the best-performing model using the same statistical accuracy metrics (R^2, MAPE, RMSE) that were used to select the best model.

2.7. Uncertainty propagation

Uncertainties in OC-CCI, SST-CCI and SSS-GLORYS datasets were propagated by means of Monte Carlo simulation of uncorrelated random errors (Quast and Shevchuk, 2025). Errors in OC-CCI datasets were simulated by random draws from normal or lognormal error distributions encoded by OC-CCI per-datum estimates of bias and root mean squared errors obtained from comparison to *in situ* measurements (Jackson et al., 2017). Errors in SST-CCI datasets were simulated by random draws from a normal distribution defined by per-datum total standard uncertainty associated with remotely sensed SST. Random errors in the SSS-GLORYS dataset were drawn from a normal distribution adopting a global per-datum standard uncertainty of 0.1 PSU. Monte Carlo simulations generated an ensemble of $N = 10$ possible variants per datum, which were sequentially processed by the DOC estimation algorithm, in addition to the nominal DOC estimation. Eventually, per-datum DOC was obtained by computing standard uncertainty, *i.e.*, the standard deviation of DOC estimates from the nominal estimate. Fig. 2 illustrates the uncertainty propagation workflow.

3. Results

3.1. Match-ups

The final, quality-controlled, match-up dataset consisted of 970 observations for each variable (Fig. 1). The majority of the match-up data were from the Northern Hemisphere, similar to the full *in situ* dataset. This spatial bias reflects the uneven distribution of available coastal *in situ* observations and represents a limitation in extrapolating the model to less-represented regions such as the Southern Hemisphere and tropical coasts. Nevertheless, the dataset still covered a wide range of

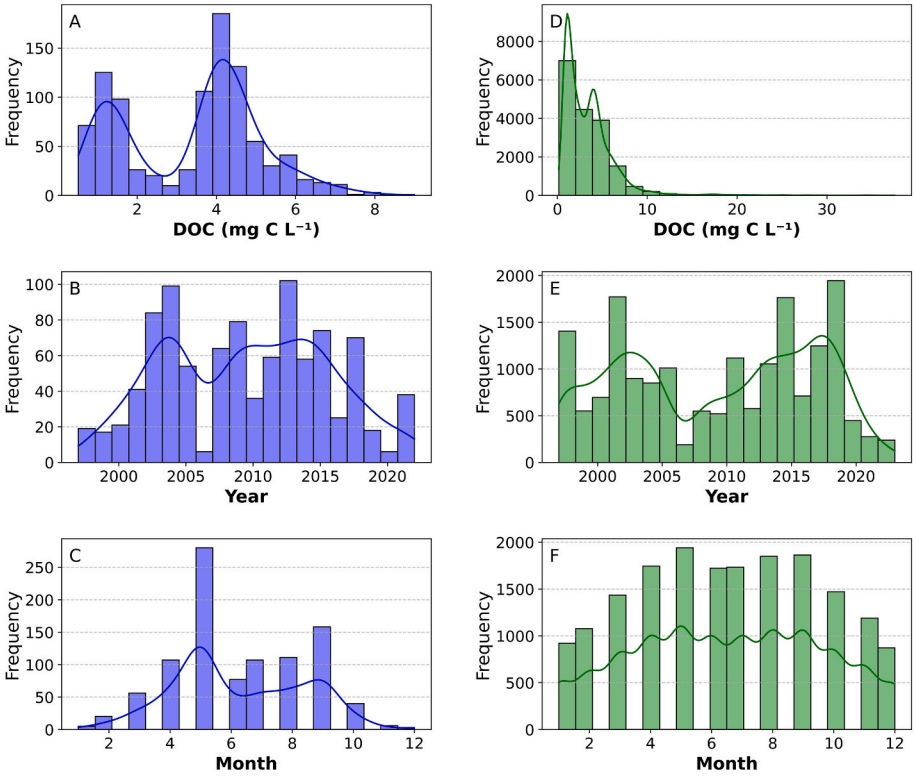

Fig. 3. Frequency histograms of the match-up dataset (A – C; N = 970) and whole in situ dataset (up to 1 m, D – F; N = 17,813). A and D – in situ Dissolved Organic Carbon concentration (DOC, mg C L^{-1}); B and E – years of DOC measurements and C and F – months of DOC measurements.

conditions, ensuring a diverse representation of the data (Table S1). DOC values ranged from 0.52 mg C L^{-1} to 9.00 mg C L^{-1}. The median DOC concentration was 3.92 mg C L^{-1}, and the mean was 3.41 mg C L^{-1} (±1.71 mg C L^{-1} standard deviation). Based on the interquartile range (IQR), 25% of the observations had DOC concentrations below 1.53 mg C L^{-1}, and 75% of the observations were less than 4.50 mg C L^{-1}. Similarly, the frequency of the used *in situ* DOC concentrations had two distinct peaks at approximately 1.0 mg C L^{-1} and 4.0 mg C L^{-1} (Fig. 3A). These peaks mainly reflect differences between DOC concentrations in the Baltic Sea (a relatively closed waterbody with strong river influence) and coastal areas that are more open to the ocean (Fig. 1). The match-ups between *in situ* and OC-CCI data are more frequent during the years between 2003 and 2005 and 2008–2015 (Fig. 3B), with most occurring between April and October and a notable peak in June

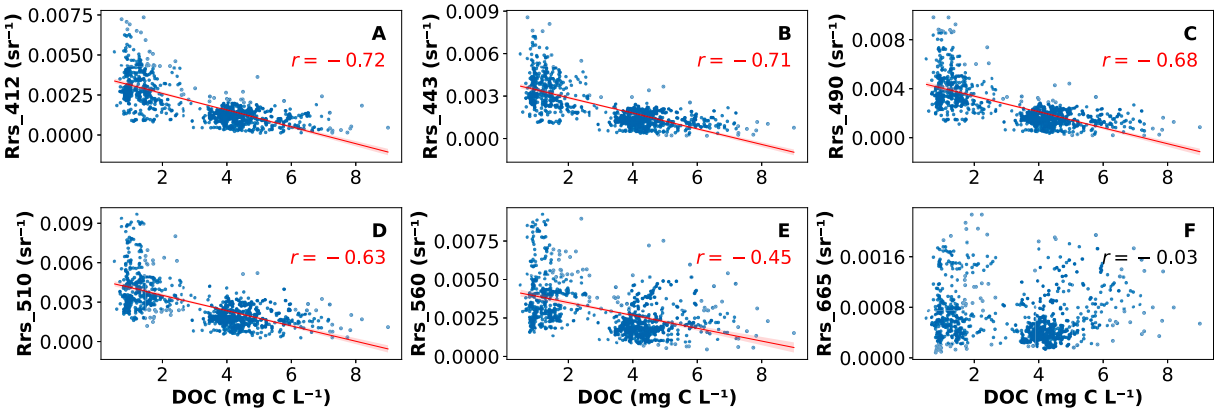

Fig. 4. Scatter plots showing the correlations (r) between in situ Dissolved Organic Carbon concentration (DOC, mg C L^{-1}) and Remote sensing reflectance ($R_{rs(\lambda)}$, sr^{-1}). Red regression line shows the best-fit relationship and the shaded area around the line shows the 95% confidence interval of the regression. The p-values < 0.001 are marked red.

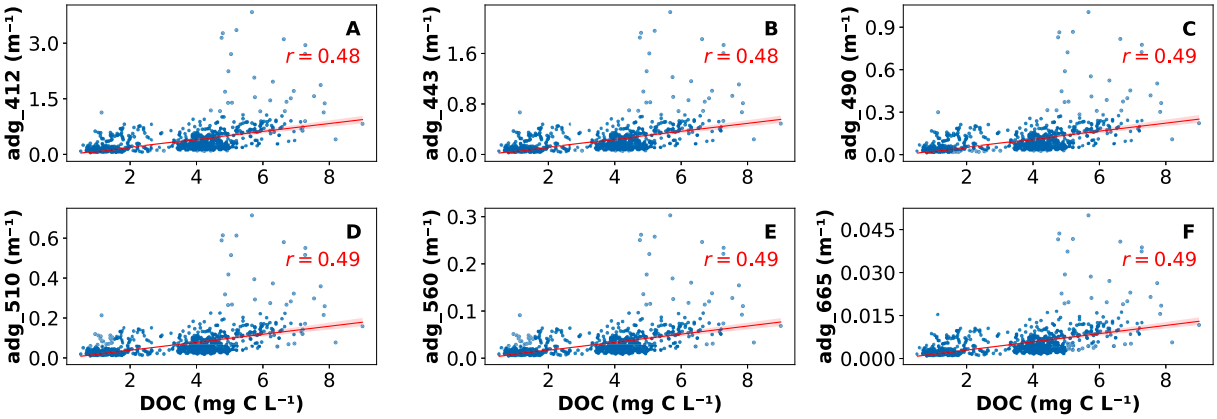

Fig. 5. Scatter plots showing the correlations (r) between in situ Dissolved Organic Carbon concentration (DOC, mg C L^{-1}) and absorption of dissolved and detrital materials (a_{dg}, m^{-1}). The red regression line shows the best-fit relationship, and the shaded area around the line shows the 95% confidence interval of the regression. The p-values < 0.001 are marked red.

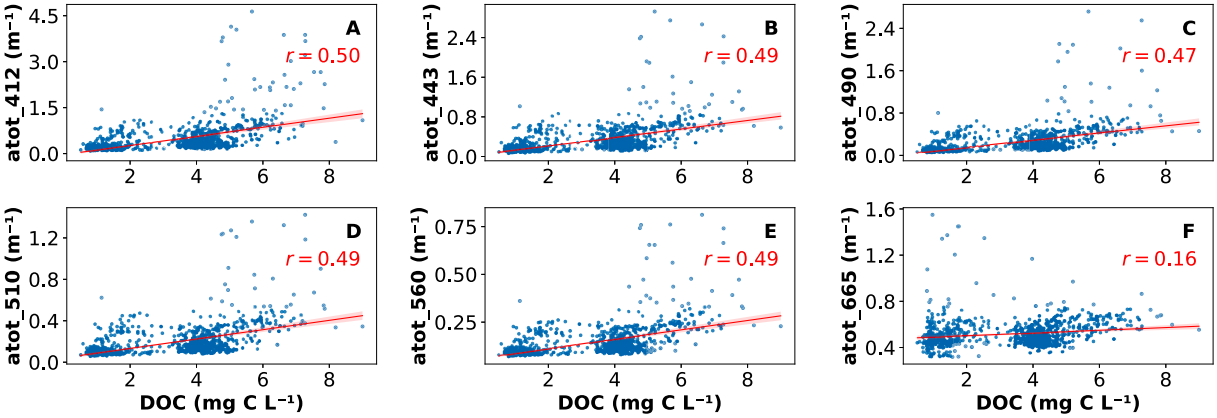

Fig. 6. Scatter plots showing the correlations (r) between in situ Dissolved Organic Carbon concentration (DOC, mg C L^{-1}) and total absorption (a_{tot}, m^{-1}). The red regression line shows the best-fit relationship, and the shaded area around the line shows the 95% confidence interval of the regression. The p-values < 0.001 are marked red.

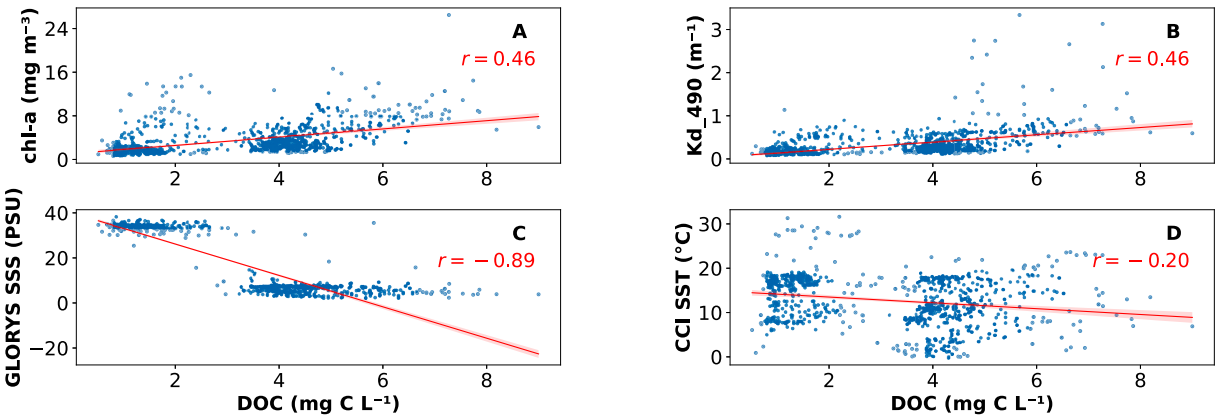

Fig. 7. Scatter plots showing the correlations (r) between in situ Dissolved Organic Carbon concentration (DOC, mg C L^{-1}) and chlorophyll-a (chl-a), diffuse attenuation coefficient at 490 nm (K_{d_490}, m^{-1}), Sea Surface Salinity (SSS) from GLORYS12V1 (SSS-GLORYS, PSU) and Sea Surface Temperature from ESA SST-CCI (SST-CCI, °C). The red regression line shows the best-fit relationship, and the shaded area around the line shows the 95% confidence interval of the regression. The p-values < 0.001 are marked red.

(Fig. 3C). This seasonal pattern is typical to satellite data, as its availability is heavily influenced by atmospheric conditions (cloud cover) and ice cover at higher latitudes between November and March. Fig. 3D–F show frequency distributions of the full *in situ* dataset (depths up to 1 m, $N = 17{,}813$) that is generally consistent with frequencies of the match-up dataset, supporting the representativeness of the match-up subset.

3.2. Feature selection, model development and uncertainty

The feature selection analysis showed that the correlation between *in situ* DOC concentration and $R_{rs(\lambda)}$ was strongest for wavelengths at 412, 443, and 490 nm ($r = -0.72$ to $r = -0.68$; Fig. 4). The correlation between *in situ* DOC and a_{dg} was moderate and consistent across all wavelengths ($r = 0.48$ to $r = 0.49$; Fig. 5) due to exponential decrease of a_{dg}. The correlation between *in situ* DOC concentration and a_{ph} was somewhat weaker than with a_{dg} ($r = 0.07$ to $r = 0.47$; Fig. S3). The correlation between *in situ* DOC concentration and a_{tot} was similar to that of a_{ph}, ranging from $r = 0.16$ to $r = 0.50$ (Fig. 6). In contrast to the absorption variables, all b_{bp} products showed negligible correlations with *in situ* DOC concentration (Fig. S4). Consequently, the b_{bp} products were excluded from further analyses. There was a moderate positive correlation between *in situ* DOC concentration and chl-*a* as well as with K_{d_490} ($r = 0$ 0.46; Fig. 7). *In situ* DOC concentrations showed a strong

Table 3
List of variables tested in the Dissolved Organic Carbon (DOC) prediction model. Abbreviations: Remote sensing reflectance ($R_{rs(\lambda)}$, sr^{-1}), absorption of dissolved and detrital materials (a_{dg}, m^{-1}), total absorption (a_{tot}, m^{-1}), Sea Surface Salinity (SSS) from GLORYS12V1 (SSS-GLORYS, PSU) and Sea Surface Temperature from ESA SST-CCI (SST-CCI, °C).

Category	Variable
Salinity and temperature	SSS-GLORYS
	SST-CCI
Remote sensing reflectance	$R_{rs(412)}$
	$R_{rs(443)}$
	$R_{rs(490)}$
	$R_{rs(510)}$
	$R_{rs(560)}$
	$R_{rs(665)}$
	$R_{rs(665)}$ - $R_{rs(412)}$
	$R_{rs(665)}$ - $R_{rs(443)}$
	$R_{rs(665)}$ - $R_{rs(490)}$
	$R_{rs(412)}$ * $R_{rs(490)}$ / $R_{rs(510)}$
	$R_{rs(443)}$ * $R_{rs(490)}$ / $R_{rs(560)}$
	$R_{rs(443)}$ * $R_{rs(510)}$ / $R_{rs(560)}$
	$R_{rs(490)}$ * $R_{rs(443)}$ / $R_{rs(510)}$
	$R_{rs(510)}$ * $R_{rs(412)}$ / $R_{rs(560)}$
	$R_{rs(490)}$ - ($R_{rs(510)}$ + $R_{rs(665)}$) / 2
	$R_{rs(665)}$ - ($R_{rs(412)}$ + $R_{rs(490)}$) / 2
	$R_{rs(665)}$ - ($R_{rs(490)}$ + $R_{rs(443)}$) / 2
	$R_{rs(665)}$ - ($R_{rs(443)}$ + $R_{rs(412)}$) / 2
	$R_{rs(665)}$ - ($R_{rs(510)}$ + $R_{rs(412)}$) / 2
	$R_{rs(665)}$ - ($R_{rs(510)}$ + $R_{rs(443)}$) / 2
	$R_{rs(665)}$ - ($R_{rs(510)}$ + $R_{rs(490)}$) / 2
IOPs	a_{dg_665}
	a_{dg_412}
	a_{dg_560}
	a_{dg_443}
	a_{dg_510}
	a_{dg_490}
	a_{tot_412}

Table 4
The best satellite-retrieval models and their performance metrics for Multiple Linear Regression (MLR), Random Forest (RF) and XGBoost across combinations with 2, 3, or 4 regressors. Highlighted in blue is the overall best-performing model.

Model	Regressors	R^2	RMSE (mg C L^{-1})	MAPE (%)
MLR_4	SSS-GLORYS SST-CCI $R_{rs(490)}$ − ($R_{rs(510)}$ + $R_{rs(665)}$) / 2 a_{tot_412}	0.85	0.67	17.45
MLR_3	SSS-GLORYS SST-CCI a_{tot_412}	0.84	0.68	17.72
MLR_2	SSS-GLORYS a_{tot_412}	0.84	0.68	18.06
RF_4	SSS-GLORYS SST-CCI $R_{rs(560)}$ a_{tot_412}	0.91	0.52	13.01
RF_3	SSS-GLORYS $R_{rs(560)}$ a_{tot_412}	0.90	0.55	13.62
RF_2	SSS-GLORYS a_{dg_443}	0.88	0.60	14.36
XG_BOOST_4	SSS-GLORYS SST-CCI $R_{rs(560)}$ a_{tot_412}	0.90	0.54	13.70
XG_BOOST_3	SSS-GLORYS SST-CCI $R_{rs(560)}$ − ($R_{rs(443)}$ + $R_{rs(412)}$) / 2	0.89	0.57	14.57
XG_BOOST_2	SSS-GLORYS a_{tot_412}	0.86	0.64	17.29

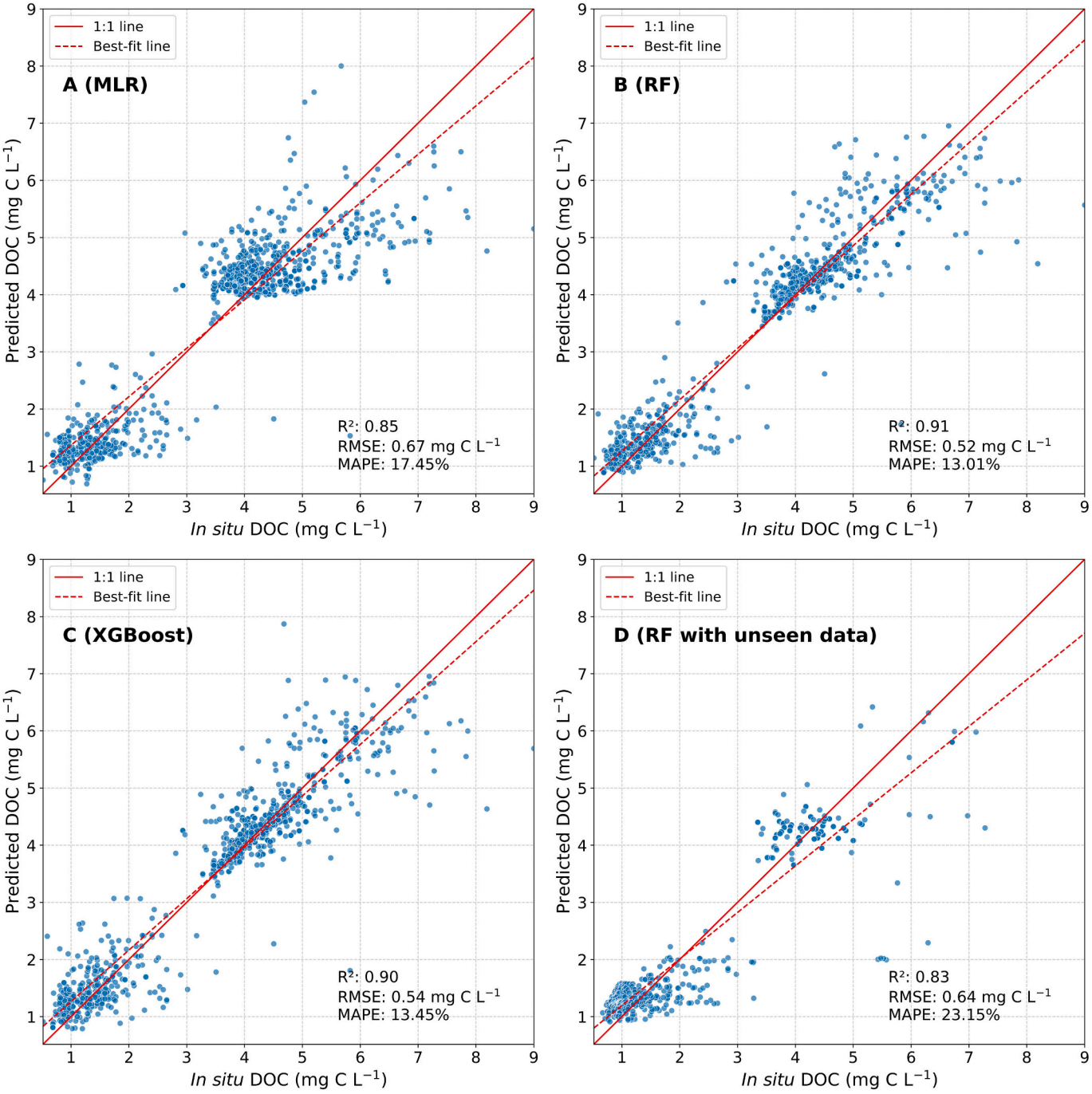

Fig. 8. Estimates of satellite-based Dissolved Organic Carbon (DOC) against in situ DOC concentrations for the best-performing models: A-C) Observed (in situ) versus predicted DOC in mg C L^{-1} from 5-fold cross-validation of the Multiple Linear Regression (MLR), the Random Forest Regression (RF) and the XGBoost Regression (XGBoost) models, respectively, and D) Observed (in situ) versus predicted DOC (mg C L^{-1}) using the Random Forest model applied to an independent unseen validation dataset. The best-fit line (dashed), the ideal 1:1 line (solid) and the accuracy metrics (R^2, RMSE, and MAPE) of each model are shown.

negative correlation with SSS ($r = -0.89$; Fig. 7). There was a moderate negative correlation between DOC concentration and SST ($r = -0.20$; Fig. 7).

Following this analysis, all variables and variable combinations included as one of the 2, 3 or 4 regressors (see paragraph 2.7) in the satellite-retrieval model development are shown in Table 3. All $R_{rs(\lambda)}$ bands (412–665 nm) and multiple band combinations were systematically tested in combination with environmental and IOP variables to identify the best predictor sets. Although $R_{rs(\lambda)}$ in the blue wavelengths (412–443 nm) showed the strongest individual correlation with *in situ* DOC, the combination that yielded the best overall model performance in cross-validation indicated that predictive strength depended on the combined effect of complementary variables rather than on correlation strength alone.

The best satellite-retrieval models and their performance metrics across configurations with 2, 3, or 4 regressors are shown in Table 4. For the best-performing MLR models, a_{tot_412} and SSS were included as features in each model, regardless of the number of regressors. The SST was

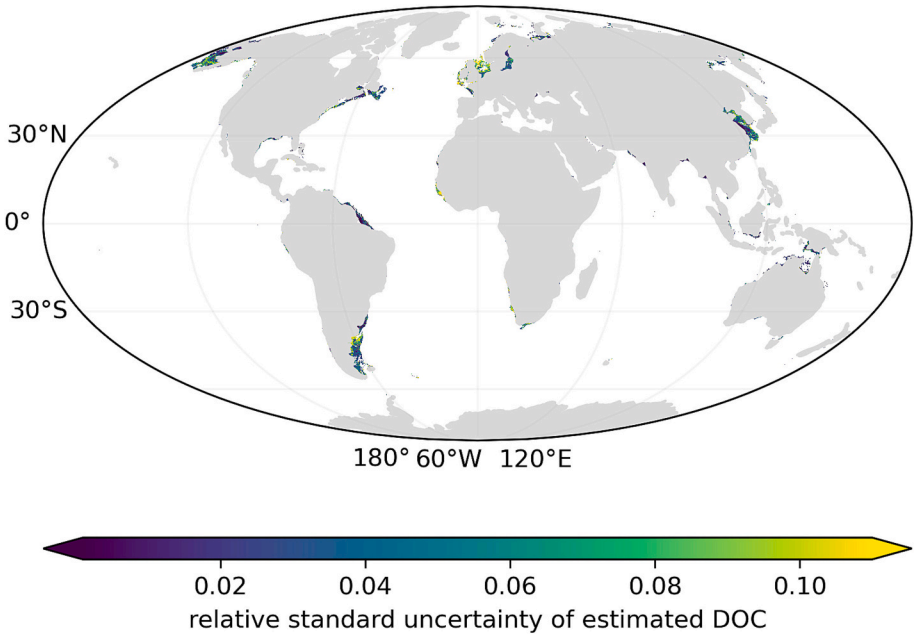

Fig. 9. Example of relative standard uncertainty of estimated DOC (April 2019).

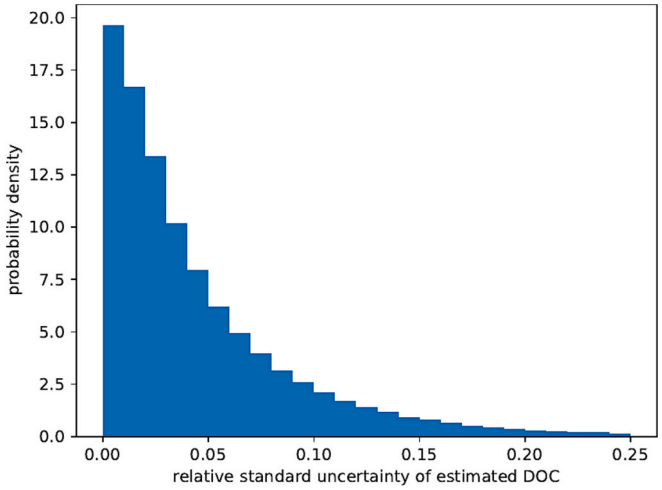

Fig. 10. Example of the distribution of relative standard uncertainty of estimated DOC (April 2019).

an additional feature in the 3- and 4-regressors MLR models, whereas the 4-regressor model also included a band ratio based on $R_{rs(\lambda)}$. For the best-performing RF models, SSS was again an important feature in all models. While a_{dg_443} was a feature in the 2-regressor RF model, the a_{tot_412} and the $R_{rs(560)}$ were features in the 3- and 4- regressors models, with SST being the additional feature in the 4-regressor RF model. For the best-performing XGBoost models, again SSS was a feature in all models regardless of the number of regressors. The SST featured in the 3- and 4-regressor XGBoost models, while the a_{tot_442} featured in the 2- and 4-regressor models. The 4-regressor XGB model also included the $R_{rs(560)}$ and the 3-regressor model contained a band ratio based on the $R_{rs(\lambda)}$ at three wavebands. The parameterisation of the best-performing satellite-retrieval models is shown in Table S2, with MLR equations and Optuna hyperparameter tuning results for the RF and XGBoost models. The corresponding estimates of DOC of the best-performing MLR, RF and XGboost models against *in situ* DOC concentrations are shown in Fig. 8A–C. The MLR model showed slightly lower R^2 and higher RMSE and MAPE compared to the RF and XGBoost models, which showed very similar performance. The model with 4 regressors (SSS-GLORYS, SST-CCI, $R_{rs(560)}$, and a_{tot_412}) using RF was the optimal model ($R^2 = 0.91$, RMSE = 0.52 mg C L^{-1} and MAPE = 13.01%) for the generation of coastal DOC concentrations from satellite observations. To further validate this best-performing model, an extended match-up dataset ($N = 452$), containing only the variables included in the optimal model, was used as an independent validation set and the results are shown in Fig. 8D ($R^2 = 0.83$, RMSE = 0.64 mg C L^{-1}, and MAPE = 23.15%).

The relative standard uncertainty of DOC estimates ranged from <1% to >11%. Figs. 9 and 10 illustrate examples of the geographical and statistical distribution of relative DOC standard uncertainty.

3.3. Mapping of global coastal DOC using the developed model—an example

We used the best-performing RF model with 4-regressors (SSS-GLORYS, SST-CCI, $R_{rs(560)}$, and a_{tot_412}) to map satellite estimates of monthly DOC in global coastal regions for two randomly selected months—April and September 2019 (Fig. 11A–B). Some coastal regions lack DOC estimates in the global maps (*e.g.*, along the coasts of North America, South America, and eastern Africa), this is primarily due to the spatial masking criteria used to define coastal waters (bathymetry shallower than 200 m, and water classes 12–14 from the OC-CCI dataset), which represent optically coastal environments. In areas where water classes 12–14 are not present within the ≤200 m depth zone, no DOC estimates were generated. Since global-scale maps limit the visibility of such narrow coastal zones, a zoomed-in view of the Baltic Sea for April 2019 is shown in Fig. 11C and a zoomed-in view of the Rio de la Plata area for September 2019 is presented in Fig. 11D. The maps demonstrate that the model can map DOC concentrations globally using data from OC-CCI, SST-CCI and GLORYS, with relatively high DOC concentrations in the Baltic Sea and other river-influenced areas such as the Rio de la Plata.

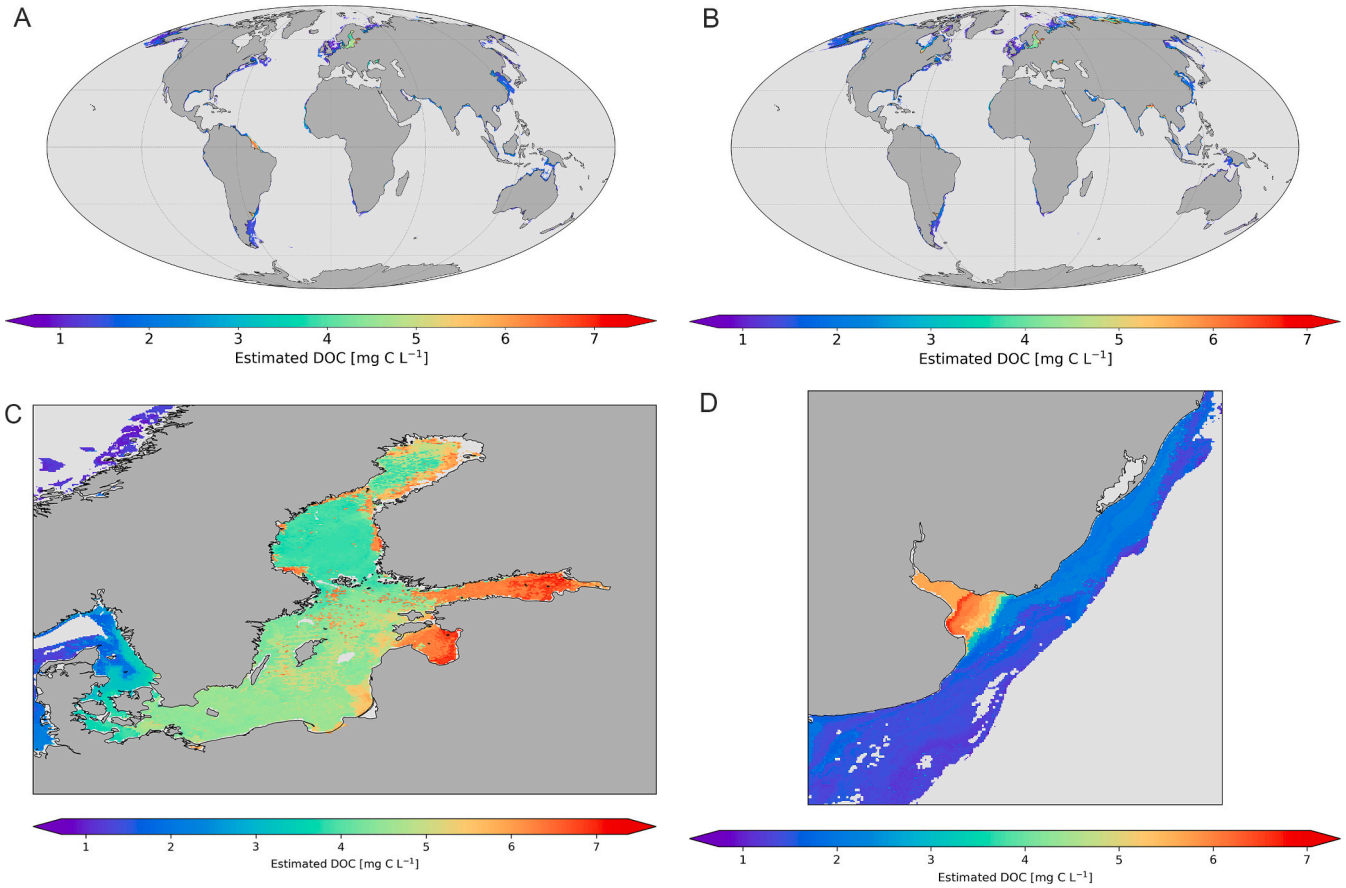

Fig. 11. A. Monthly global coastal dissolved organic carbon (DOC) for April 2019. The RF model with 4 regressors (SSS-GLORYS, SST-CCI, $R_{rs(560)}$, and atot_412) was used for the estimation. **B.** Monthly global coastal dissolved organic carbon (DOC) for September 2019. The RF model with 4 regressors (SSS-GLORYS, SST-CCI, $R_{rs(560)}$, and atot_412) was used for the estimation. **C.** Monthly coastal dissolved organic carbon (DOC) for April 2019 in the Baltic Sea. The RF model with 4 regressors (SSS-GLORYS, SST-CCI, $R_{rs(560)}$, and atot_412) was used for the estimation. **D.** Monthly coastal dissolved organic carbon (DOC) for September 2019 in the Rio de la Plata area. The RF model with 4 regressors (SSS-GLORYS, SST-CCI, $R_{rs(560)}$, and atot_412) was used for the estimation.

4. Discussion

4.1. Match-ups and in situ DOC observations

Most of the carbon transported from land to the sea is processed in near-coastal waters, with up to 80% buried in sediments, never reaching the deep ocean carbon pool (Bauer et al., 2013). Therefore, estimating DOC in coastal waters is needed for understanding the global carbon cycle. Although satellite remote sensing offers better spatial and temporal coverage, the development and accuracy estimation of remote sensing-based DOC retrieval models needs high-quality *in situ* data. Therefore, *in situ* data are critical for validating and calibrating satellite-derived estimates (Bailey and Werdell, 2006). The most comprehensive global *in situ* DOC dataset for coastal waters is part of the CoastDOM database (Lønborg et al., 2024). It includes over 62,000 individual DOC measurements providing an invaluable resource for broader DOC assessments in coastal waters, as well as for validating remote sensing products. However, despite the large volume of available *in situ* data, the use of this dataset in satellite validation is somewhat constrained by sampling depth and timing. After filtering surface layer data and performing a match-up analyses, only approximately 1000 of the 62,000 measurements, *i.e.*, less than 2%, were available to use for model development and validation. This low match-up number is partly due to the strict temporal and spatial criteria and quality control measures applied, but it also reflects the limited availability of cloud-free ocean colour observations at daily resolution, particularly in coastal regions. Data gaps caused by persistent cloud cover are less pronounced when using temporally aggregated products (*e.g.*, 5-day, 8-day, or monthly composites).

Despite this, the mean DOC concentration of the *in situ* data used in the current study (3.41 ± 1.71 mg C L^{-1} (standard deviation), *i.e.*, 284.2 ± 142.5 μmol C L^{-1}) was very close to the mean DOC concentration estimated by Barrón and Duarte (2015) (285.7 ± 5.7 μmol C L^{-1} standard error). However, the mean DOC concentration was higher than the mean reported in the CoastDOM database (182 ± 314 standard deviation μmol C L^{-1}) (Lønborg et al., 2024), but still within the range of their reported standard deviation. The higher mean is expected as only surface layer values were used in the current study. In general, DOC concentrations are higher in surface layers due to terrestrial runoff and photochemical processes that predominantly occur near the surface (Ding et al., 2019; Shinomura et al., 2005). Furthermore, the spatial bias in the *in situ* dataset (largely dominated by samples from the Northern Hemisphere) presents a limitation that may affect the performance of DOC retrieval models in less-represented areas, such as the Southern Hemisphere and tropical coastal regions. However, the dataset still covers a broad range of environmental conditions that provide a strong basis for model development.

In situ radiometry is generally preferred for algorithm development because it avoids errors related to atmospheric correction and ensures that optical and biogeochemical measurements are taken from the same location at the same time (Bailey and Werdell, 2006). However, one of the main challenges in developing and training models is the lack of simultaneous radiometric and biogeochemical data, especially for models that require large training data sets, *e.g.*, machine learning

models. In contrast, satellite-derived R_{rs} provides extensive spatial and temporal coverage, enabling consistent global analysis and model application across a wide range of environments. Since *in situ* radiometric measurements are still rare, especially those made simultaneously with dissolved organic carbon measurements, this study relied on R_{rs} data obtained from satellites. Strict quality control and matchup selection criteria were used to help to minimise potential errors.

Nevertheless, to optimise the use of *in situ* data for remote sensing applications, greater spatial and temporal overlap between field campaigns and satellite overpasses is needed. Additionally, more *in situ* data, especially from areas not well represented in this study (Fig. 1) are needed to reduce the uncertainty in global DOC algorithms. Therefore, it is highly encouraged that CDOM, DOC and other carbon fractions together with radiometric measurements are included in monitoring programmes and plans of scientific cruises.

4.2. Predicting DOC in coastal waters using satellite remote sensing

Several models have been developed to predict DOC in coastal waters using remote sensing data (Table 5). Most of these models have been developed for specific regions and require validation if used in other coastal areas (Cao et al., 2018; Cao and Tzortziou, 2024; Cherukuru et al., 2016, 2021; le Fouest et al., 2018; Liu et al., 2014; Mannino et al., 2008, 2016; Pan and Wong, 2015; Qiong et al., 2011; Tehrani et al., 2013; Tuzcu Kokal et al., 2024; Zhang et al., 2024). The most common approach follows two-steps: (1) Relating a_{CDOM} or the spectral slope of a_{CDOM} to $R_{rs(\lambda)}$ at specific wavelengths or $R_{rs(\lambda)}$ band ratios, and (2) Deriving DOC from a_{CDOM} using empirical relationships. The a_{CDOM} and DOC often have a strong relationship in coastal waters, but it tends to vary across regions and seasons (Fichot and Benner, 2011; Harvey et al., 2015). Using the a_{CDOM} spectral slope can improve DOC retrieval accuracy in those cases (Cao et al., 2018; Harvey et al., 2015; Mannino et al., 2016). However, there are also other approaches. For example, machine learning methods like deep neural networks (DNN) have been used to retrieve DOC concentrations directly from $R_{rs(\lambda)}$ data (Zhang et al., 2024), and variables like chl-a concentrations and SST have also been used to improve DOC estimations (Liu et al., 2014; Pan and Wong, 2015). Zhang et al. (2024) proposed a novel DNN-based DOC retrieval method using spatial interpolation to expand sparse *in situ* data into a virtual training dataset. This approach significantly increased sample diversity and model robustness, resulting in very high retrieval accuracy (RMSE = 0.08 mg C L^{-1}). Although our approach differs methodologically, their work underscores the growing potential of machine learning for DOC estimation, especially in data-limited regions. However, there are important limitations to consider. The model was developed and tested in a single, optically constrained coastal region (Jiangsu) with

Table 5
Overview of models for satellite-based DOC estimations in coastal waters.

Type of model	Description	Error Metrics	Sensor and its spatial resolution	Study area	Reference
Empirical model	$R_{rs(\lambda)}$ was used to retrieve a_{CDOM} at various wavelengths and a_{CDOM} spectral slope in the 275–295 nm ($S_{275–295}$). DOC concentrations were obtained from a_{CDOM} and $S_{275–295}$. DOC = $a_{CDOM\ (300)}$/(exp(−15.05–33.95* $S_{275–295}$) + exp.(−1.502–104.3*$S_{275–295}$)	Mean absolute percent difference (MAPD) =18% (MERIS Envisat) and MAPD = 33% (MODIS Aqua)	MODIS Aqua (1 km); MERIS-Envisat (300 m)	The Chesapeake Bay and Delaware Bay estuaries and coastal waters of the Middle Atlantic Bight, the northern Gulf of Mexico, USA	(Cao et al., 2018)
Empirical model	The non-linear relationship between the DOC-specific $a_{CDOM(300)}$ and $S_{275–295}$ was used.	MAPD = 10%	Sentinel-3, OLCI	The Long Island Sound-tidal estuary of the North Atlantic Ocean, USA	(Cao and Tzortziou, 2024).
Empirical model	DOC = 327.5*exp.(−3.638*$R_{rs(412)}$/$R_{rs(488)}$)	R^2 = 0.86	MODIS Aqua (1 km)	Moreton Bay, Australia	(Cherukuru et al., 2016).
The Inversion model (three-component)	1) a regional spectral optical library of specific IOPs; 2) a forward model to simulate total absorption, total backscattering and backscattering albedo; 3) an inversion mechanism to derive biogeochemical parameters through spectral matching of remote sensing and forward-modelled backscattering albedo.	Bias = 0.96 Mean Average Error (MAE) = 1.36 μmol C L^{-1}	MODIS Aqua (1 km)	Sarawak coastal waters, Malaysia	(Cherukuru et al., 2021).
Empirical model	If chl-a < 0.8 mg^{-1} m^3, the relationship between DOC and a_{CDOM} was used and if chl-a > 0.8 mg^{-1} m^3, the DOC was retrieved using a multilinear model that incorporated a_{CDOM} and chl-a.	Average Relative Error = 10.2%–24.8%	MODIS Aqua (1 km)	East China Sea, China	(Liu et al., 2014)
Empirical model	The a_{CDOM} was correlated with *in situ* $R_{rs(\lambda)}$ band ratios, and DOC was then derived from a_{CDOM}.	MAPD = 9.3 ± 7.3%	SeaWiFS (1.1 km), MODIS Aqua (1 km)	Middle Atlantic Bight, USA	(Mannino et al., 2008)
Empirical model	R_{rs} at 443 and 547 nm were related to a_{CDOM} at specific wavelengths (412 nm, 443 nm, and 355 nm), and then a_{CDOM} was used to estimate DOC concentration.	MAPD = 11.1–18.8, RMSE = 15.2–25.0 μmol L^{-1}, R^2 = 0.42–0.90	SeaWiFS (1.1 km), MODIS Aqua (1 km)	Middle Atlantic Bight, USA	(Mannino et al., 2016)
Empirical model	$a_{CDOM(412)}$ was related to the $R_{rs(488)}$ and $R_{rs(555)}$ by the empirical relationship and then $a_{CDOM(412)}$ and SST were used to retrieve DOC.	R^2 = 0.917, RMSE = 9.8 μmol C L^{-1}, MAPD = 7.3%	SeaWiFS (1.1 km), MODIS Aqua (1 km)	Northern South China Sea Shelf-sea, China	(Pan and Wong, 2015)
Empirical model	Band ratio algorithms that use the seasonal relationships between a_{CDOM}- $R_{rs(\lambda)}$ and a_{CDOM}-DOC	R^2 = 0.4–0.72 RMSE = 26.69–44.22 μmol C L^{-1}	SeaWiFS (1.1 km), MODIS Aqua (1 km), and MERIS (300 m)	Gulf of Mexico on the Texas-Louisiana shelf, USA	(Tehrani et al., 2013)
Empirical model	1) the algorithm estimates the a_{CDOM} at 250 nm from the $R_{rs(490)}$/$R_{rs(665)}$ band ratio using a double exponential fit. 2) the algorithm retrieves DOC concentration from a_{CDOM} using a quadratic fit.	R^2 = 0.75, MAPE = 19.12%	Sentinel-2 MSI (10–60 m)	Plum Island Estuary, Massachusetts, USA	(Tuzcu Kokal et al., 2024)
Machine learning	A deep neural network (DNN) was used to retrieve DOC concentration from $R_{rs(\lambda)}$.	RMSE =0.08 mg C L^{-1}, MAPE = 4.13%, Relative Prediction Deviation (RPD) = 3.28	MODIS Terra (500 m)	Coastal waters of Jiangsu, China	(Zhang et al., 2024)

very limited *in situ* data, and its applicability to more complex or diverse environments remains uncertain. Additionally, the interpolation process assumes spatial continuity, which may not hold for all biogeochemical variables, potentially affecting retrieval accuracy. In contrast, approach of the current study is trained and validated, using observed *in situ* data, with regressors derived from remote sensing satellite observations, to capture a wide range of environmental variability.

Global DOC retrieval algorithms for the open ocean have been developed and validated (Aurin et al., 2018; Laine et al., 2024). However, these algorithms were optimised for open-ocean conditions. Coastal waters are optically more complex and spatially heterogeneous, and while recent *in situ* data-based syntheses now describe global patterns and stocks of coastal DOC (Lønborg et al., 2025), there remains a need for satellite-based DOC retrievals specifically developed and validated across the global coastal ocean. Hence, in the current study, we introduced an approach that enables daily, global-scale DOC estimation in coastal waters using the ESA OC-CCI dataset at 4-km resolution together with the ESA SST-CCI and GLORYS12V1 SSS. While the spatial resolution of the OC-CCI merged dataset is coarser than that of single sensors, like MODIS (1 km) or Sentinel-3 OLCI (300 m), which are used in regional algorithms (see Table 5), the OC-CCI data provide daily global coverage since 1997, which gives it a significant advantage for tracking DOC dynamics over time.

In the current study, various optical and environmental variables were included in the feature selection of the satellite-based DOC model. Most of the features showed a moderate to strong correlation with *in situ* DOC concentrations. Among $R_{rs(\lambda)}$, DOC correlated best with the blue wavelengths, which was expected since CDOM absorption decreases exponentially with increasing wavelength (Mannino et al., 2008; Maritorena et al., 2002). However, despite this stronger correlation, the best-performing model used $R_{rs(560)}$ in combination with SSS, SST, and atot_412. Water-leaving signals at blue wavelengths may be negligible in coastal and inland waters due to higher concentrations of CDOM and phytoplankton that both absorb strongly in the blue part of the spectrum. All bands were tested in the model selection process, and $R_{rs(560)}$ proved to be the most reliable predictor in the optimal set, *i.e.*, the model's performance was driven by the combined effect of multiple variables rather than the strength of any single-band correlation. a_{dg} includes absorption by both CDOM and detrital material. While CDOM absorption is directly related to DOC, the presence of detrital material weakens the correlation between adg and DOC, as detrital material is not directly related to DOC. Additionally, the variability of the detrital material depends largely on sediment resuspension and riverine inputs, which may not always coincide with the variability of DOC concentration (Aurin and Dierssen, 2012). The moderate relationship between a_{ph} and DOC concentrations suggests that although some of the DOC is formed by autochthonous organic matter from phytoplankton, other sources, such as allochthonous organic matter from land-based inputs, play a more important role in coastal waters (Matsuoka et al., 2012). As expected, negligible correlations were found between b_{bp} and DOC concentration since dissolved material primarily affects water absorption rather than backscattering (Stramski et al., 2004). The positive relationship between DOC and chl-*a* concentration reflects that phytoplankton exudation and decomposition contribute to the dissolved organic matter and thereby also affects the DOC concentration (Toming et al., 2013). The correlation with K_{d_490} suggested that light attenuation increases in water in conjunction with increasing DOC concentration, due to the absorption of CDOM in the blue part of the spectrum, which reduces water transparency and increases K_{d_490} values (Paavel et al., 2011). Among the analysed features, DOC concentration showed the strongest correlation with SSS. Lower salinity indicates freshwater inputs, such as river discharge, that are associated with DOC transported to coastal waters from land (Pan and Wong, 2015; Tehrani et al., 2013). The weak correlation between DOC concentration and SST suggests that temperature has an inconsistent effect on DOC variability, as temperature significantly affects the DOC concentration through multiple mechanisms. For example, warmer temperatures can enhance biological activity, including phytoplankton exudation and microbial decomposition of organic matter, which has a positive effect on DOC concentration. On the other hand, higher temperatures are often associated with more intense solar radiation, which enhances the photochemical degradation of DOC, decomposing it into smaller, more labile compounds or leading to its mineralisation into CO_2. These opposing processes may explain the relatively moderate overall correlation between DOC concentration and SST in coastal waters observed in our study.

In addition to the DOC relationships with predictor variables, the accuracy of DOC retrievals is also influenced by the characteristics of the satellite data products themselves. The present model benefits from the long-term, globally consistent coverage of OC-CCI, SST-CCI, and GLORYS12V1 datasets. However, the rather coarse spatial resolution of these satellite products introduces additional limitations in dynamic and heterogeneous coastal regions. Multi-kilometre pixels often integrate signals from optically distinct water masses, *e.g.*, in river plume areas, which can smooth or dilute the sub-pixel variability of DOC and its optical proxies (Aurin et al., 2013). Aurin et al. (2013) demonstrated that pixel sizes of 500 m or smaller are generally required to resolve more than 90% of optical variability in river plumes and coarser resolutions significantly underestimate the variability. Although match up criteria that include inter-pixel variability filters partially address this issue, residual biases are likely to remain in areas with strong horizontal gradients. Therefore, retrieval performance in such heterogeneous coastal waters (*e.g.*, Rio de la Plata, Fig. 11D) may be lower than in more homogeneous coastal environments. To improve predictions in highly heterogeneous areas, future work could use input data from higher spatial resolution satellites, like Sentinel-2 MSI or Sentinel-3 OLCI. This would help better resolve spatial variability and improve retrieval accuracy in those areas. However, the uncertainty analysis showed that for most coastal waters the uncertainty remains low (mostly below 5%), suggesting that the model can be effectively applied across diverse coastal environments.

Another potential source of error is the use of a combined multi-mission OC-CCI dataset, which may introduce temporal inconsistencies that could affect DOC retrievals, especially in coastal and high-latitude regions. Although the OC-CCI dataset uses advanced harmonisation methods, like sensor-based atmospheric correction, band-shifting, and gain bias correction (Sathyendranath et al., 2019), recent studies (*e.g.*, van Oostende et al., 2022) have shown that inter-mission discontinuities persist. These arise not only from differences in the radiometric properties of the sensors, but also from variable spatial and temporal coverage, especially in optically complex waters. Such inconsistencies can introduce artificial steps in time series and may bias retrievals in regions where sensor coverage has changed over time. The Temporal Gap Detection Method (TGDM) proposed by van Oostende et al. (2022) help to homogenise temporal coverage. However, it was not implemented in the current study as our main objective was model development, with a focus on match-ups rather than long-term trend analysis. TGDM effectively reduces temporal artefacts but does so by masking irregularly observed days, which can substantially reduce the number of valid observations and potentially exclude regions of high interest. However, if the model developed in this study is later applied for time-series or trend analysis, integrating TGDM or similar temporal homogenization approaches may help reduce artefacts and improve the robustness of DOC retrievals, particularly in coastal zones where coverage and atmospheric correction performance vary most.

5. Conclusions

This study developed and evaluated a novel approach to DOC concentrations in global coastal waters using multi-source satellite remote sensing data, including ESA OC-CCI, SST-CCI and SSS-GLORYS. We demonstrated that incorporating both optical and environmental variables into one statistical model, like MLR, RF, and XGBoost, improves

DOC estimations in coastal waters from satellite observations. The best-performing RF model used SSS, SST, $R_{rs(560)}$ and a_{tot_412} as regressors, with cross-validation metrics of $R^2 = 0.91$, RMSE = 0.52 mg C L^{-1}, and MAPE = 13.01%, and independent validation on unseen data gave $R^2 = 0.83$, RMSE = 0.64 mg C L^{-1}, and MAPE = 23.15%. This model outperformed many empirical models developed for coastal waters, where MAPD ranged from 7.3% to 33% and R^2 varied between 0.4 and 0.91, showing the importance of considering both optical proxies and key environmental drivers to improve DOC estimations from remote sensing data in coastal waters. While the developed DOC model showed high performance compared to earlier regional satellite-based models, the application of the model to the relatively coarser resolution of OC-CCI remains a challenge, particularly in highly dynamic coastal zones where DOC can vary significantly over short distances, for example, in river plumes or estuaries. Yet, the OC-CCI time series does offer the benefit of a continuous ocean-colour dataset since 1997. There is potential for further development of the DOC model using other ocean-colour or multi-spectral satellite sensors. There are land remote sensing satellites, like Sentinel-2 MSI, that provide sufficient spatial resolution (10 m), 2–5 days revisit time and even the capability to map DOC from surface waters (Toming et al., 2016). However, creating coastal zone maps with Sentinel-2 MSI resolution at the global scale remains computationally challenging.

Funding

This research was funded by the ESA SCOPE project - Satellite-based observations of Carbon in the Ocean: Pools, fluxes and Exchanges (4000142532/23/I-DT) and by the Estonian Research Council (PRG2630, KT and TK). It was also supported by the Simons Foundation 'Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems' (CBIOMES) project (G.K). This work is a contribution to the activities of the National Centre of Earth Observation (NCEO) of the United Kingdom (G.K.).

CRediT authorship contribution statement

K. Toming: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **G. Kulk:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **R. Quast:** Formal analysis, Methodology, Visualization, Writing – review & editing. **R. Shevchuk:** Formal analysis, Methodology, Visualization, Writing – review & editing. **T. Kutser:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

The datasets used in the current study:•

OC-CCI dataset version 6.0, ESA, available online at https://catalogue.ceda.ac.uk/uuid/5011d22aae5a4671b0cbc7d05c56c4f0/•

SST-CCI dataset version 3.0, ESA, available online at https://catalogue.ceda.ac.uk/uuid/4a9654136a7148e39b7feb56f8bb02d2/•

GLORYS12V1 dataset, E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS), available online at https://doi.org/10.48670/moi-00021•

CoastDOM v1, available online at https://doi.pangaea.de/10.1594/PANGAEA.964012

The code for the global DOC satellite retrieval algorithm for coastal waters is openly available in GitHub repository: https://github.com/KaireTo/global-coastal-doc-algorithm-SCOPE

Acknowledgements

We acknowledge the ESA for providing the OC-CCI dataset (version 6.0) and the developers and contributors of CoastDOM v1 for compiling a valuable database of *in situ* DOC measurements in coastal waters. We thank the E.U. Copernicus Marine Service for providing SSS data (GLORYS12V1) and the Copernicus Climate Change Service and ESA for providing SST data (SST-CCI, version 3.0). We also thank the four anonymous reviewers for their valuable and constructive comments, which helped to improve the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2026.115388.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation Hyperparameter optimization framework. Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min. 2623–2631. https://doi.org/10.1145/3292500.3330701.

Aurin, D.A., Mannino, A., Franz, B.A., 2013. Spatially resolving ocean color and sediment dispersion in river plumes. Remote Sens. Environ. 137, 261–272. https://doi.org/10.1016/j.rse.2013.06.018.

Aurin, D., Mannino, A., Lary, D.J., 2018. Remote sensing of CDOM, CDOM spectral slope, and dissolved organic carbon in the Global Ocean. Appl. Sci. 8 (12), 2687. https://doi.org/10.3390/APP8122687.

Aurin, D.A., Dierssen, H.M., 2012. Advantages and limitations of ocean color remote sensing in CDOM-dominated, mineral-rich coastal and estuarine waters. Remote Sens. Environ. 125, 181–197. https://doi.org/10.1016/j.rse.2012.07.001.

Bailey, S.W., Werdell, P.J., 2006. A multi-sensor approach for the on-orbit validation of ocean color satellite data products. Remote Sens. Environ. 102 (1–2), 12–23. https://doi.org/10.1016/j.rse.2006.01.015.

Barrón, C., Duarte, C.M., 2015. Global biogeochemical cycles from the coastal ocean. Glob. Biogeochem. Cycles 29, 1725–1738. https://doi.org/10.1002/2014GB005056. Received.

Bauer, J.E., Cai, W.J., Raymond, P.A., Bianchi, T.S., Hopkinson, C.S., Regnier, P.A.G., 2013. The changing carbon cycle of the coastal ocean. Nature 504 (7478), 61–70. https://doi.org/10.1038/nature12857.

Bowers, D.G., Harker, G.E.L., Smith, P.S.D., Tett, P., 2000. Optical properties of a region of freshwater influence (the Clyde Sea). Estuarine Coastal Shelf Sci. 50 (5), 717–726. https://doi.org/10.1006/ecss.1999.0600.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Brewin, R.J.W., Sathyendranath, S., Platt, T., Bouman, H., Ciavatta, S., Dall'Olmo, G., Dingle, J., Groom, S., Jönsson, B., Kostadinov, T.S., Kulk, G., Laine, M., Martínez-Vicente, V., Psarra, S., Raitsos, D.E., Richardson, K., Rio, M.-H., Rousseaux, C.S., Salisbury, J., Walker, P., 2021. Sensing the ocean biological carbon pump from space: a review of capabilities, concepts, research gaps and future developments. Earth-Sci. Rev. 217 (April), 103604. https://doi.org/10.1016/j.earscirev.2021.103604.

Campbell, A.D., Fatoyinbo, T., Charles, S.P., Bourgeau-Chavez, L.L., Goes, J., Gomes, H., Halabisky, M., Holmquist, J., Lohrenz, S., Mitchell, C., Moskal, L.M., Poulter, B., Qiu, H., Resende De Sousa, C.H., Sayers, M., Simard, M., Stewart, A.J., Singh, D., Trettin, C., Lagomasino, D., 2022. A review of carbon monitoring in wet carbon systems using remote sensing. Environ. Res. Lett. 17 (2), 025009. https://doi.org/10.1088/1748-9326/AC4D4D.

Cao, F., Tzortziou, M., 2021. Capturing dissolved organic carbon dynamics with Landsat-8 and Sentinel-2 in tidally influenced wetland–estuarine systems. Sci. Total Environ. 777, 145910. https://doi.org/10.1016/J.SCITOTENV.2021.145910.

Cao, F., Tzortziou, M., 2024. Impacts of hydrology and extreme events on dissolved organic carbon dynamics in a heavily urbanized estuary and its major tributaries: a view from space. J. Geophys. Res. Biogeosciences 129 (3). https://doi.org/10.1029/2023JG007767.

Cao, F., Tzortziou, M., Hu, C., Mannino, A., Fichot, C.G., Del Vecchio, R., Najjar, R.G., Novak, M., 2018. Remote sensing retrievals of colored dissolved organic matter and dissolved organic carbon dynamics in north American estuaries and their margins. Remote Sens. Environ. 205, 151–165. https://doi.org/10.1016/j.rse.2017.11.014.

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, 13–17-Augu, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Cherukuru, N., Ford, P.W., Matear, R.J., Oubelkheir, K., Clementson, L.A., Suber, K., Steven, A.D.L., 2016. Estimating dissolved organic carbon concentration in turbid coastal waters using optical remote sensing observations. Int. J. Appl. Earth Obs. Geoinf. 52, 149–154. https://doi.org/10.1016/j.jag.2016.06.010.

Cherukuru, N., Martin, P., Sanwlani, N., Mujahid, A., Müller, M., 2021. A semi-analytical optical remote sensing model to estimate suspended sediment and dissolved organic carbon in tropical coastal waters influenced by peatland-draining river discharges off Sarawak, borneo. Remote Sens. 13 (1), 1–31. https://doi.org/10.3390/rs13010099.

Dafner, E.V., Wangersky, P.J., 2002. A brief overview of modern directions in marine DOC studies part II - recent progress in marine DOC studies. J. Environ. Monit. 4 (1), 55–69. https://doi.org/10.1039/b107279j.

Ding, L., Ge, T., Wang, X., 2019. Dissolved organic carbon dynamics in the East China Sea and the Northwest Pacific Ocean. Ocean Sci. 15 (5), 1177–1190. https://doi.org/10.5194/os-15-1177-2019.

Dunne, J.P., Sarmiento, J.L., Gnanadesikan, A., 2007. A synthesis of global particle export from the surface ocean and cycling through the ocean interior and on the seafloor. Global Biogeochem. Cycles. https://doi.org/10.1029/2006GB002907.

Embury, O., Merchant, C.J., Good, S.A., Rayner, N.A., Høyer, J.L., Atkinson, C., Block, T., Alerskans, E., Pearson, K.J., Worsfold, M., McCarroll, N., Donlon, C., 2024. Satellite-based time-series of sea-surface temperature since 1980 for climate applications. Sci. Data 11 (1). https://doi.org/10.1038/s41597-024-03147-w.

Fennel, K., Alin, S., Barbero, L., Evans, W., Bourgeois, T., Cooley, S., Dunne, J., Feely, R. A., Martin Hernandez-Ayon, J., Hu, X., Lohrenz, S., Muller-Karger, F., Najjar, R., Robbins, L., Shadwick, E., Siedlecki, S., Steiner, N., Sutton, A., Turk, D., Aleck Wang, Z., 2019. Carbon cycling in the north American coastal ocean: A synthesis. Biogeosciences 16 (6), 1281–1304. https://doi.org/10.5194/BG-16-1281-2019.

Ferrari, G.M., Dowell, M.D., Grossi, S., Targa, C., 1996. Relationship between the optical properties of chromophoric dissolved organic matter and total concentration of dissolved organic carbon in the southern Baltic Sea region. Mar. Chem. 55 (3–4), 299–316. https://doi.org/10.1016/S0304-4203(96)00061-8.

Fichot, C.G., Benner, R., 2011. A novel method to estimate DOC concentrations from CDOM absorption coefficients in coastal waters. Geophys. Res. Lett. 38 (3), 1–5. https://doi.org/10.1029/2010GL046152.

Fichot, C.G., Benner, R., 2012. The spectral slope coefficient of chromophoric dissolved organic matter (*S*275-295) as a tracer of terrigenous dissolved organic carbon in river-influenced ocean margins. Limnol. Oceanogr. 57 (5), 1453–1466. https://doi.org/10.4319/lo.2012.57.5.1453.

le Fouest, V., Matsuoka, A., Manizza, M., Shernetsky, M., Tremblay, B., Babin, M., 2018. Towards an assessment of riverine dissolved organic carbon in surface waters of the western Arctic Ocean based on remote sensing and biogeochemical modeling. Biogeosciences 15 (5), 1335–1346. https://doi.org/10.5194/bg-15-1335-2018.

Gattuso, J.P., Frankignoulle, M., Wollast, R., 1998. Carbon and carbonate metabolism in coastal aquatic ecosystems. Annu. Rev. Ecol. Syst. 29, 405–434. https://doi.org/10.1146/ANNUREV.ECOLSYS.29.1.405.

Good, S.A., Embury, O., 2024. ESA Sea Surface Temperature Climate Change Initiative (SST_cci): Level 4 Analysis Product, Version 3.0. NERC EDS Centre for Environmental Data Analysis.

Hansell, D.A., Carlson, C.A., 2013. Localized refractory dissolved organic carbon sinks in the deep ocean. Global Biogeochem. Cycles 27 (3), 705–710. https://doi.org/10.1002/GBC.20067.

Harvey, E.T., Kratzer, S., Andersson, A., 2015. Relationships between colored dissolved organic matter and dissolved organic carbon in different coastal gradients of the Baltic Sea. Ambio 44 (3), 392–401. https://doi.org/10.1007/s13280-015-0658-4.

Hoegh-Guldberg, O., Northrop, E., Lubchenco, J., 2019. The ocean is key to achieving climate and societal goals. Science 365 (6460), 1372–1374. https://doi.org/10.1126/SCIENCE.AAZ4390.

Hoyer, S., Hamman, J., 2017. Xarray: N-D labeled arrays and datasets in Python. J. Open Res. Softw. 5 (1), 10. https://doi.org/10.5334/jors.148.

Jackson, T., Sathyendranath, S., Melin, F., 2017. An improved optical classification scheme for the ocean colour essential climate variable and its applications. Remote Sens. Environ. 203, 152–161. https://doi.org/10.1016/j.rse.2017.03.036.

Jean-Michel, L., Eric, G., Romain, B.B., Gilles, G., Angélique, M., Marie, D., Clément, B., Mathieu, H., Olivier, L.G., Charly, R., Tony, C., Charles-Emmanuel, T., Florent, G., Giovanni, R., Mounir, B., Yann, D., Pierre-Yves, L.T., 2021. The Copernicus global 1/12° oceanic and sea ice GLORYS12 reanalysis. Front. Earth Sci. 9. https://doi.org/10.3389/feart.2021.698876.

Kowalczuk, P., Zabłocka, M., Sagan, S., Kuliński, K., 2010. Fluorescence measured in situ as a proxy of CDOM absorption and DOC concentration in the Baltic Sea. Oceanologia 52 (3), 431–471. https://doi.org/10.5697/OC.52-3.431.

Kulk, G., Platt, T., Dingle, J., Jackson, T., Jönsson, B. F., Bouman, H. A., Babin, M., Brewin, R. J. W., Doblin, M., Estrada, M., Figueiras, F. G., Furuya, K., González-Benítez, N., Gudfinnsson, H. G., Gudmundsson, K., Huang, B., Isada, T., Kovač, Ž., Lutz, V. A., … Sathyendranath, S. (2020). Primary production, an index of climate change in the ocean: satellite-based estimates over two decades. Rem. Sens., *Vol. 125, page 826*. doi:https://doi.org/10.3390/RS12050826.

Laine, M., Kulk, G., Jönsson, B.F., Sathyendranath, S., 2024. A machine learning model-based satellite data record of dissolved organic carbon concentration in surface waters of the global open ocean. Front. Mar. Sci. 11 (June), 1–11. https://doi.org/10.3389/fmars.2024.1305050.

Liu, B., D'Sa, E.J., Joshi, I., 2019. Multi-decadal trends and influences on dissolved organic carbon distribution in the Barataria Basin, Louisiana from in-situ and Landsat/MODIS observations. Remote Sens. Environ. 228 (May), 183–202. https://doi.org/10.1016/j.rse.2019.04.023.

Liu, Q., Pan, D., Bai, Y., Wu, K., Chen, C.T.A., Liu, Z., Zhang, L., 2014. Estimating dissolved organic carbon inventories in the East China Sea using remote-sensing data. J. Geophys. Res. Oceans 119 (10), 6557–6574. https://doi.org/10.1002/2014JC009868.

Lønborg, C., Carreira, C., Abril, G., Agustí, S., Amaral, V., Andersson, A., Arístegui, J., Bhadury, P., Bif, M.B., Borges, A.V., Bouillon, S., Calleja, M.L., Cotovicz, L.C., Cozzi, S., Doval, M., Duarte, C.M., Eyre, B., Fichot, C.G., García-Martín, E.E., Álvarez-Salgado, X.A., 2024. A global database of dissolved organic matter (DOM) concentration measurements in coastal waters (CoastDOM v1). Earth Syst. Sci. Data 16 (2), 1107–1119. https://doi.org/10.5194/ESSD-16-1107-2024.

Lønborg, C., Fuentes-Santos, I., Carreira, C., Amaral, V., Arístegui, J., Bhadury, P., Bif, M. B., Calleja, M.L., Chen, Q., Cotovicz, L.C., Cozzi, S., Eyre, B.D., García-Martín, E.E., Giani, M., Gonçalves-Araujo, R., Gruber, R., Hansell, D.A., Holding, J.M., Hunter, W., Álvarez-Salgado, X.A., 2025. Dissolved organic carbon in coastal waters: global patterns, stocks and environmental physical controls. Global Biogeochem. Cycles 39 (5), 1–20. https://doi.org/10.1029/2024GB008407.

Mannino, A., Russ, M.E., Hooker, S.B., 2008. Algorithm development and validation for satellite-derived distributions of DOC and CDOM in the U.S. middle Atlantic bight. J. Geophys. Res. Oceans 113 (C7), 7051. https://doi.org/10.1029/2007JC004493.

Mannino, A., Signorini, S.R., Novak, M.G., Wilkin, J., Friedrichs, M.A.M., Najjar, R.G., 2016. Dissolved organic carbon fluxes in the middle Atlantic bight: an integrated approach based on satellite data and ocean model products. J. Geophys. Res. Biogeosci. 121 (2), 312–336. https://doi.org/10.1002/2015JG003031.

Maritorena, S., Siegel, D.A., Peterson, A.R., 2002. Optimization of a semianalytical ocean color model for global-scale applications. Appl. Opt. 41 (15), 2705–2714. https://doi.org/10.1364/AO.41.002705.

Martin, P., Sanwlani, N., Lee, T.W.Q., Wong, J.M.C., Chang, K.Y.W., Wong, E.W.S., Liew, S.C., 2021. Dissolved organic matter from tropical peatlands reduces shelf sea light availability in the Singapore Strait, Southeast Asia. Mar. Ecol. Prog. Ser. 672, 89–109. https://doi.org/10.3354/meps13776.

Matsuoka, A., Bricaud, A., Benner, R., Para, J., Sempere, R., Prieur, L., Belanger, S., Babin, M., 2012. Tracing the transport of colored dissolved organic matter in water masses of the southern Beaufort Sea: relationship with hydrographic characteristics. Biogeosciences 9 (3), 925–940. https://doi.org/10.5194/bg-9-925-2012.

Muller-Karger, F.E., Varela, R., Thunell, R., Luerssen, R., Hu, C., Walsh, J.J., 2005. The importance of continental margins in the global carbon cycle. Geophys. Res. Lett. 32 (1), 1–4. https://doi.org/10.1029/2004GL021346.

van Oostende, M., Hieronymi, M., Krasemann, H., Baschek, B., Röttgers, R., 2022. Correction of inter-mission inconsistencies in merged ocean colour satellite data. Front. Rem. Sens. 3 (July), 1–17. https://doi.org/10.3389/frsen.2022.882418.

Paavel, B., Arst, H., Metsamaa, L., Toming, K., Reinart, A., 2011. Optical investigations of CDOM-rich coastal waters in Pärnu Bay. Estonian J. Earth Sci. 60 (2), 102–112. https://doi.org/10.3176/EARTH.2011.2.04.

Pan, X., Wong, G.T.F., 2015. An improved algorithm for remotely sensing marine dissolved organic carbon: climatology in the northern South China Sea shelf-sea and adjacent waters. Deep-Sea Res. Part II: Top. Stud. Oceanogr. 117, 131–142. https://doi.org/10.1016/j.dsr2.2015.02.025.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2012. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12 (85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

Qiong, L., Pan, D., Huang, H., Lu, J., Zhu, Q., 2011. Relationship between the colored dissolved organic matter and dissolved organic carbon and the application on remote sensing in East China Sea. Proceedings 8175, 364–371. https://doi.org/10.1117/12.897865.

Quast, R., Shevchuk, R., 2025. Kaleidoscope: Monte Carlo Uncertainty Propagation for ESA SCOPE. https://doi.org/10.5281/ZENODO.17544961.

Regnier, P., Friedlingstein, P., Ciais, P., Mackenzie, F.T., Gruber, N., Janssens, I.A., Laruelle, G.G., Lauerwald, R., Luyssaert, S., Andersson, A.J., Arndt, S., Arnosti, C., Borges, A.V., Dale, A.W., Gallego-Sala, A., Goddéris, Y., Goossens, N., Hartmann, J., Heinze, C., Thullner, M., 2013. Anthropogenic perturbation of the carbon fluxes from land to ocean. Nat. Geosci. 6 (8), 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 (7901), 401–410. https://doi.org/10.1038/S41586-021-04339-9.

Sathyendranath, S., Brewin, R.J.W., Brockmann, C., Brotas, V., Calton, B., Chuprin, A., Cipollini, P., Couto, A.B., Dingle, J., Doerffer, R., Donlon, C., Dowell, M., Farman, A., Grant, M., Groom, S., Horseman, A., Jackson, T., Krasemann, H., Lavender, S., Platt, T., 2019. An ocean-colour time series for use in climate studies: the experience of the ocean-colour climate change initiative (OC-CCI). Sensors 19 (19). https://doi.org/10.3390/s19194285.

Sathyendranath, S., Jackson, T., Brockmann, C., Brotas, V., Calton, B., Chuprin, A., Clements, O., Cipollini, P., Danne, O., Dingle, J., Donlon, C., Grant, M., Groom, S., Krasemann, H., Lavender, S., Mazeran, C., Mélin, F., Müller, D., Steinmetz, F., Valente, A., Zühlke, M., Feldman, G., Franz, B., Frouin, R., Werdell, J., Platt, T., 2023. ESA Ocean Colour Climate Change Initiative (Ocean_Colour_cci): Monthly climatology of global ocean colour data products at 4 km resolution, version 6.0 [dataset]. NERC EDS Centre for Environmental Data Analysis. doi: 10.5285/690fdf8f229c4d04a2aa68de67beb733.

Semiletov, I., Pipko, I., Gustafsson, Ö., Anderson, L.G., Sergienko, V., Pugach, S., Dudarev, O., Charkin, A., Gukov, A., Bröder, L., Andersson, A., Spivak, E., Shakhova, N., 2016. Acidification of East Siberian Arctic Shelf waters through addition of freshwater and terrestrial carbon. Nat. Geosci. 9, 361–365. https://doi.org/10.1038/ngeo2695.

Shinomura, Y., Iwata, T., Suzuki, Y., 2005. Diel changes in dissolved organic carbon in the upper layer of Suruga bay, Japan. Estuar. Coast. Shelf Sci. 62 (4), 699–709. https://doi.org/10.1016/j.ecss.2004.10.006.

Siegel, D.A., Devries, T., Cetinić, I., Bisson, K.M., 2023. Quantifying the ocean's biological pump and its carbon cycle impacts on global scales. Ann. Rev. Mar. Sci. 15, 329–356. https://doi.org/10.1146/ANNUREV-MARINE-040722-115226.

Stramski, D., Boss, E., Bogucki, D., Voss, K.J., 2004. The role of seawater constituents in light backscattering in the ocean. Prog. Oceanogr. 61 (1), 27–56. https://doi.org/10.1016/j.pocean.2004.07.001.

Tehrani, N.C., D'Sa, E.J., Osburn, C.L., Bianchi, T.S., Schaeffer, B.A., 2013. Chromophoric dissolved organic matter and dissolved organic carbon from sea-viewing wide field-of-view sensor (seawifs), moderate resolution imaging spectroradiometer (modis) and meris sensors: case study for the northern gulf of mexico. Remote Sens. 5 (3), 1439–1464. https://doi.org/10.3390/rs5031439.

Toming, K., Tuvikene, L., Vilbaste, S., Agasild, H., Viik, M., Kisand, A., Feldmann, T., Martma, T., Jones, R.I., Nõges, T., 2013. Contributions of autochthonous and allochthonous sources to dissolved organic matter in a large, shallow, eutrophic lake with a highly calcareous catchment. Limnol. Oceanogr. 58 (4), 1259–1270. https://doi.org/10.4319/lo.2013.58.4.1259.

Toming, K., Kutser, T., Laas, A., Sepp, M., Paavel, B., Nõges, T., 2016. First experiences in mapping lakewater quality parameters with sentinel-2 MSI imagery. Remote Sens. 8 (8). https://doi.org/10.3390/RS8080640.

Toming, K., Liu, H., Soomets, T., Uuemaa, E., Nõges, T., Kutser, T., 2024. Estimation of the biogeochemical and physical properties of lakes based on remote sensing and artificial intelligence applications. Remote Sens. 16 (3), 1–28. https://doi.org/10.3390/rs16030464.

Tuzcu Kokal, A., Harringmeyer, J.P., Cronin-Golomb, O., Weiser, M.W., Hong, J., Ghosh, N., Swanson, J., Zhu, X., Musaoglu, N., Fichot, C.G., 2024. Capturing the dynamics of dissolved organic carbon (DOC) in tidal saltmarsh estuaries using remote-sensing-informed models. J. Geophys. Res. Biogeosci. 129 (10). https://doi.org/10.1029/2024JG008059.

Vodacek, A., Blough, N.V., DeGrandpre, M.D., Peltzer, E.T., Nelson, R.K., 1997. Seasonal variation of CDOM and DOC in the middle Atlantic bight: terrestrial inputs and photooxidation. Limnol. Oceanogr. 42 (4), 674–686. https://doi.org/10.4319/lo.1997.42.4.0674.

Zhang, J., Li, H., Miao, Y., Zhou, Z., Lyu, H., Gong, Z., 2024. Remote sensing retrieval method based on few-shot learning: a case study of surface dissolved organic carbon in Jiangsu Coastal Waters, China. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3524257.