Journal Pre-proof

High-resolution satellite retrieval of dissolved organic carbon in coastal waters using machine learning: Application to China's Dongshan Bay

Lingling Li , Xiaolong Yu , Wendian Lai , Nengwang Chen , Weidong Guo , Hongyan Bao , Guizhi Wang , Shuiying Huang , Wenlu Lan , Xiaoyan Peng , Zhongping Lee

PII: S1001-0742(26)00371-2
DOI: https://doi.org/10.1016/j.jes.2026.05.031
Reference: JES 5497

To appear in: *Journal of Environmental Sciences*

Received date: 20 November 2025
Revised date: 8 May 2026
Accepted date: 9 May 2026

Please cite this article as: Lingling Li , Xiaolong Yu , Wendian Lai , Nengwang Chen , Weidong Guo , Hongyan Bao , Guizhi Wang , Shuiying Huang , Wenlu Lan , Xiaoyan Peng , Zhongping Lee , High-resolution satellite retrieval of dissolved organic carbon in coastal waters using machine learning: Application to China's Dongshan Bay, *Journal of Environmental Sciences* (2026), doi: https://doi.org/10.1016/j.jes.2026.05.031

Highlights

- Machine learning enables efficient high-resolution monitoring of nearshore DOC dynamics.
- The AutoGluon-DOC model achieved high DOC retrieval accuracy (error of ~10 %).
- DOC gradients in Dongshan Bay are jointly driven by tides, Chl, and aquaculture.
- Satellite series reveal decreasing DOC trend in Dongshan Bay between 2021 and 2023.

High-resolution satellite retrieval of dissolved organic carbon in coastal waters using machine learning: Application to China's Dongshan Bay

Lingling Li[a], Xiaolong Yu[a,b,*], Wendian Lai[a], Nengwang Chen[c], Weidong Guo[a], Hongyan Bao[a], Guizhi Wang[a], Shuiying Huang[c], Wenlu Lan[d], Xiaoyan Peng[d], Zhongping Lee[a]

a. State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen 361102, China

b. National Observation and Research Station for the Taiwan Strait Marine Ecosystem, Xiamen University, Zhangzhou, Fujian 361102, China

c. College of the Environment and Ecology, Xiamen University, Xiamen 361102, China

d. Marine Environmental Monitoring Center of Guangxi, Beihai 536000, China

Abstract: Monitoring dissolved organic carbon (DOC) concentrations in coastal waters is critical to elucidating carbon dynamics, thereby facilitating the understanding carbon dynamics and quantifying lateral carbon fluxes to the open ocean. Machine learning combined with ocean color remote sensing provides an effective approach to estimate optically inactive substances such as DOC in complex coastal environments. In this study, we compiled a comprehensive training dataset in Google Earth Engine by paring Sentinel-2 radiometric measurements with *in situ* DOC measurements from coastal waters of Southeast China. Based on this dataset, a machine learning-based model, termed AutoGluon-DOC, was developed to retrieve DOC from Rayleigh-corrected top-of-atmosphere reflectance ($\rho_{rc}(\lambda)$) and auxiliary information (*e.g.*, acquisition time). The model achieved a median absolute percentage error of ~10 % on the validation dataset. Application of AutoGluon-DOC to Sentinel-2 images over Dongshan Bay, a mangrove-estuary-aquaculture composite system, revealed a clear spatial gradient of decreasing DOC from the bay head toward the open sea. DOC spatial variability was primarily regulated by tidal fluctuations and Chl in the bay head, while aquaculture activities dominated in the bay mouth. Satellite-derived DOC series also exhibited a distinct seasonal cycle, with higher concentrations in summer and lower in winter, mainly associated with biological production and episodically enhanced by river discharge during extreme rainfall events. These results underscore the effectiveness of AutoGluon-DOC for high-resolution monitoring of DOC dynamics in complex coastal environments, with a framework adaptable to diverse regions as training datasets expand.

Keywords:

Dissolved organic carbon (DOC)

Machine learning

Remote sensing

Coastal waters

Google Earth Engine (GEE)

Driving factors analysis

----

* Corresponding author. E-mail: xlyu@xmu.edu.cn (X. Yu)

1 Introduction

Dissolved organic carbon (DOC) in aquatic environments is a vital component of the global carbon cycle, as it regulates carbon fluxes and CO_2 emissions through its interactions with microbial processes and biogeochemical cycling, thereby strongly influencing ecosystem functioning (Mopper et al., 1991; Wu et al., 2015; Xiao et al., 2024). Furthermore, coastal waters play a key role in the global carbon cycle by facilitating the transport of DOC from terrestrial environments to the ocean (Dai et al., 2012; Pan et al., 2014). Consequently, accurately characterizing the spatiotemporal variations of DOC in coastal regions is crucial for improving assessments of the global carbon cycle, as it aids in understanding and quantifying the contributions of the lateral carbon fluxes into the ocean (Bonelli et al., 2022). Moreover, DOC concentrations in wetlands, estuaries, and coastal ecosystems serve as proxies of carbon export from diverse blue carbon ecosystems (Cao and Tzortziou, 2021; Sullivan et al., 2025). Given its ecological and biogeochemical importance, DOC monitoring remains challenging using conventional field sampling, whereas satellite remote sensing offers an efficient alternative for large-scale observation.

However, as an optically inactive substance, DOC is typically estimated based on its linear or non-linear correlations with the absorption coefficient of colored dissolved organic matter (CDOM, denoted as a_g in m^{-1}) (Cao et al., 2018; Yu et al., 2016). Yet, such relationships are only valid in regions where DOC covaried with a_g (Foltz et al., 2025). Meanwhile, the sources of DOC in coastal waters are highly diverse, encompassing inputs from inland soils, rivers, wetlands, and lakes (Fichot et al., 2023), as well as natural production through marine primary productivity (Lønborg et al., 2025). In addition, atmospheric deposition and human activities also significantly contribute to DOC generation (Li et al., 2025; Qu et al., 2024). As a result, DOC often does not follow a conservative mixing relationship with a_g in most coastal areas (Bonelli et al., 2022; Yu et al., 2016) , posing a challenge for accurately retrieving DOC through remote sensing.

Compounding this issue is the challenge of accurately retrieving a_g from satellite-measured remote sensing reflectance (R_{rs}), which is always subject to substantial uncertainties in coastal waters (Fichot et al., 2023; Wei et al., 2019). These uncertainties are primarily caused by the poor performance of atmospheric correction algorithms due to the complex optical properties of coastal waters and variable atmospheric conditions, including diverse aerosol types (Pahlevan et al., 2021). Such uncertainties could lead to negative R_{rs} values in the blue bands (Zhao et al., 2023a), which further complicate the reliable estimation of a_g. To mitigate the potential uncertainties associated with atmospherically corrected R_{rs}, many studies have utilized the Rayleigh scattering-corrected top-of-atmosphere reflectance ($\rho_{rc}(\lambda)$) for retrieving water-quality parameters (WQPs) (Feng et al., 2018; Hu, 2011). This is because Rayleigh scattering can be precisely modeled, whereas aerosol scattering correction remains challenging in coastal environments (Gordon, 1997). In particular, when coupled with well-trained machine learning (ML) or neural network (NN) models, using $\rho_{rc}(\lambda)$ as model inputs could significantly reduce retrieval uncertainties introduced by aerosol correction in coastal waters, especially when the training datasets are both representative and comprehensive (Lai et al., 2022). Thus, using $\rho_{rc}(\lambda)$ as the primary input features, ML or NN can offer a promising solution for deriving DOC from ocean color satellite measurements.

In addition to atmospheric correction, retrieving DOC in coastal waters is further challenged by the spatial mismatches between satellite and field measurements and the difficulty of efficiently compiling a large matchup dataset. On the one hand, coastal waters typically exhibit high spatial heterogeneity, and traditional ocean color satellite data with resolutions ranging from hundreds of meters to kilometers often fail to capture fine-scale features in these regions (Fichot et al., 2023). Therefore, higher-resolution satellite data is strongly desired to monitor DOC dynamics in coastal waters. In this regard, Sentinel-2 Multispectral Instrument (MSI), with a 10-meter resolution and a revisit of approximately 5 days for both Sentinel-2A/2B satellites, provides an effective data source for high-resolution spatial monitoring of coastal waters. On the other hand, the conventional approaches for pairing satellite-field matchups involve downloading satellite imageries that match *in situ* measurements within a specific temporal window. This process requires substantial storage, time, and manpower, especially when handling large data volumes of high spatial-resolution images. Fortunately, the vast accumulation of satellite data on the Google Earth Engine (GEE) platform now enables efficient matching of *in situ* data with processed satellite data, thereby significantly streamlining the data preprocessing workflow. Consequently, many studies have already utilized the GEE platform for remote sensing retrieval of WQPs (Altuntas et al., 2025; Das et al., 2024).

To address the above-mentioned challenges, this study develops an automated machine-learning framework for estimating DOC concentrations in coastal waters. This DOC retrieval framework consists of three key components: 1) automatically processing Sentinel-2 MSI imagery with Acolite to generate $\rho_{rc}(\lambda)$ data on the GEE platform and compile a dataset pairing $\rho_{rc}(\lambda)$ and *in situ* DOC measurements across different temporal windows; 2) training the AutoGluon-DOC model and assessing its performance by comparing retrieved DOC with *in situ* observations; 3) applying the model to Dongshan Bay to demonstrate its capability for capturing high-resolution DOC dynamics and identifying potential driving factors in a complex coastal environment. By using DSB as a representative study case, we aim to illustrate the practical utility of the AutoGluon-DOC framework for monitoring optically inactive constituents in diverse and dynamic coastal ecosystems.

2 Materials and methods

2.1 Field measurements

Multiple cruises were conducted in Xiamen Bay (XMB), Dongshan Bay (DSB), and Beibu Gulf (BBG) to collect field DOC measurements between 2021 and 2023. XMB and DSB are located in southern Fujian Province, while the BBG is in southern Guangxi Province, China. The collection of DOC data from various regions aimed not only to expand the overall sample size but also to enhance the spatial representativeness of the dataset across diverse nearshore water types. The locations of all sampling sites are shown in **Fig. 1**, while **Appendix A Table S1** summarizes the corresponding sampling dates and the ranges of DOC concentrations (in mg/L) for each cruise. Water samples collected at each sampling site were filtered through pre-combusted GF/F filters (0.7 μm) to obtain the filtrate. DOC concentrations were then determined using the high-temperature catalytic oxidation method with a Shimadzu total organic carbon (TOC) analyzer (model: TOC-CPH). Details on the standard procedure for DOC measurements, including pre-cruise preparation, water sample collection, filtration setup, and analysis of DOC samples, can be found in Wu et al. (2013).

2.2 Satellite data processing

The proposed framework for DOC retrieval is illustrated in **Fig. 2** and comprises three main parts, *i.e.*, 1) compilation of the training dataset from satellite-field matchups, 2) development and evaluation of the DOC retrieval model, and 3) model application and analysis of driving factors.

The Level 1C top-of-atmosphere reflectance ($\rho_t(\lambda)$) from Sentinel-2A/B MSI that matches the *in situ* DOC measurements was extracted via GEE. The matched $\rho_t(\lambda)$ were then processed into $\rho_{rc}(\lambda)$ using the Acolite Python package (v. 20220222.0, https://github.com/acolite/acolite/). Flags were applied to mask out non-water, retaining only pixels exempt from flags such as land, cloud, and sun glint (Vanhellemont, 2019). The matching criteria include a temporal window of ±3.5 days and a spatial window of 3-by-3 pixels. A matchup was considered valid only if it contained more than five valid pixels and the coefficient of variation of $\rho_{rc}(\lambda)$ across the visible (443-560 nm) and 865 nm bands remained below 0.15 (Bailey and Werdell, 2006; Pahlevan et al., 2016). As a result, 1,005 satellite-field matchups were obtained through this initial matching process, forming a preliminary dataset for further analysis.

The matched $\rho_{rc}(\lambda)$ were further screened through a quality control procedure to ensure only reasonable spectra were retained. Specifically, to mitigate adjacency effects in nearshore waters, typical $\rho_{rc}(\lambda)$ spectra from clear to turbid waters, along with surrounding land covers (*e.g.*, vegetation, buildings), were compared to identify and exclude anomalous $\rho_{rc}(\lambda)$ spectra for water pixels (**Appendix A Table S2** and **Fig. S1**). Here, only $\rho_{rc}(\lambda)$ spectra that exhibited typical water spectral features were retained and used as high-quality input to train the AutoGluon-DOC. Based on the spectral features of different objects (**Appendix A Fig. S3**), a straightforward screening approach was applied to filter out anomalous spectra using $\rho_{rc}(\lambda)$ values at 443 nm, 865 nm, and the spectral difference between $\rho_{rc}(704)$ and $\rho_{rc}(740)$. Detailed screening criteria can be found in **Appendix A** Eqs. (S1)-(S3). After this quality control process, 367 matchups were retained for model development, which is hereafter referred to as Dataset-3.5d.

Considering the high spatiotemporal heterogeneity of coastal waters, different temporal windows for matchups were applied to evaluate the robustness of the model performance. A few subsets were extracted from Dataset-3.5d by narrowing the temporal match window between the field measurement and satellite overpass. Three subsets, termed Dataset-2.5d (N = 163), Dataset-1d (N = 56), and Dataset-3h (N = 7), were generated when the temporal windows were set to ±2.5 days, ±1 day, and ±3 hours, respectively.

2.3 Development of AutoGluon-DOC

Selecting an appropriate modeling framework is critical for achieving accurate and efficient retrieval of DOC from satellite data, particularly when dealing with limited training samples and complex hyperparameter tuning. In this context, AutoGluon offers significant advantages as an automated machine learning framework. It simplifies model selection and hyperparameter optimization by automatically adjusting model configurations, thereby enabling rapid, adaptive, and high-performing model development with minimal manual intervention (Erickson et al., 2020). A key strength of AutoGluon lies in its ability to stack and ensemble a diverse set of machine learning and neural network models. Its base models include algorithms such as k-nearest neighbors, neural networks, LightGBM, random forests, CatBoost, and XGBoost, with each layer optimizing the combination of these models. This is achieved through a multi-layer stacking strategy where the predictive outputs from a preceding layer are utilized as augmented features to train the models in the subsequent layer, enabling higher-level models to effectively correct the systematic biases of the base learners.

During training, AutoGluon conducts automated model selection and hyperparameter optimization by exploring model-specific hyperparameter search spaces. Randomized and adaptive search strategies are used to evaluate candidates based on validation performance, with underperforming configurations pruned via early stopping and well-performing models integrated through bagging and stacking ensembles (Erickson et al., 2020). Building upon this automated optimization framework, additional measures were implemented to further enhance model generalization and mitigate overfitting. Specifically, 5-fold bagging and a stacking layer were employed throughout the training process, in conjunction with Bayesian optimization for more efficient hyperparameter exploration. Model performance was evaluated using the root mean squared error (RMSE) as the primary metric. The resulting performance leaderboard for the AutoGluon-DOC model is detailed in **Appendix A Table S3**. Based on this multi-stage optimization, the final optimal model was identified as WeightedEnsemble_L3, a multi-layer stacked ensemble. This model achieved a validation score of -0.175, outperforming all individual base models from the L1 and L2 layers.

The input features for the AutoGluon-DOC consist of $\rho_{rc}(\lambda)$ from Sentinel-2A MSI (at 443, 492, 560, 665, 704, 740, 783, 833, and 865 nm) and Sentinel-2B MSI (at 442, 492, 559, 665, 704, 739, 780, 833, and 864 nm), along with the acquisition time of the imagery. Following established practices in recent water quality studies (Caballero et al., 2020; Cao and Tzortziou, 2021; Pahlevan et al., 2019), we treat Sentinel-2A and Sentinel-2B as a unified "twin satellite mission". Corresponding bands with similar central wavelengths are treated as equivalent inputs during model training. In addition to spectral information, AutoGluon can automatically extract temporal features from the image acquisition time, including year, month, day, and season, which can influence DOC patterns. By incorporating these features, the model is better able to capture the temporal dynamics of DOC concentrations. The output of the AutoGluon-DOC is the logarithmic value of DOC concentration (*i.e.*, ln(DOC)). The logarithmic transformation helps reduce data skewness and improves the robustness and stability of the model. The AutoGluon-DOC model was subsequently trained on 70 % of the randomly selected datasets using $\rho_{rc}(\lambda)$ from Sentinel-2 MSI, with validation on the remaining 30 % confirming its strong performance.

2.4 Statistical analysis

To evaluate the accuracy of the AutoGluon-DOC and its subsequent validation, statistical metrics, including the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute percentage error (MAPE) were employed. These metrics are calculated following,

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(X_i^{estimated} - X_i^{measured})^2}{\sum_{i=1}^{N}(\bar{X} - X_i^{measured})^2}, \tag{1}$$

$$RMSE = \sqrt{\left(\sum_{i=1}^{N}\left(X_i^{estimated} - X_i^{measured}\right)\right)^2 \times \frac{1}{N}}, \tag{2}$$

$$MAPE = \left(\sum_{i=1}^{N}\left|\frac{X_i^{estimated} - X_i^{measured}}{X_i^{measured}}\right|\right) \times \frac{100}{N}, \tag{3}$$

where, $X_i^{estimated}$ and $X_i^{measured}$ are the satellite-estimated and *in situ* measured DOC, respectively, and $\bar{X}$ is the mean value of measured DOC.

2.5 Elasticity factor model

Meteorological, hydrological, and tidal factors, including precipitation (mm), riverine discharge (m^3/month), and tidal height (m), could regulate freshwater input and biogeochemical processes in estuarine ecosystems (Cao and Tzortziou, 2021, 2024). Additionally, as a key indicator of phytoplankton biomass, Chlorophyll-a (Chl, mg/m^3) reflects the contribution of autochthonous biological processes to DOC dynamics (Liang et al., 2025).

Precipitation data from 2021 to 2023 were obtained from the Yunxiao meteorological station. Riverine discharge data at the Zhangjiang Estuary during the same period were estimated using calibrated measurements from surrounding inflow rivers within the estuarine catchment, following the procedure described in detail in Yu et al. (2018). In addition, since DSB is characterized by an irregular semidiurnal tide, with two flooding and ebbing cycles per day, tidal effects significantly influence the biogeochemical cycling within the bay. Tidal height data for the same period were obtained from Dongshan station, as provided by the National Marine Data and Information Service (http://global-tide.nmdis.org.cn/). The locations of the Yunxiao and Dongshan stations are highlighted in **Fig. 5a**. Chl concentrations were derived from Sentinel-2 imagery using the OC3E algorithm (O'Reilly and Werdell, 2019), with Acolite-processed R_{rs} products as inputs. This NASA standard three-band algorithm was selected because its spectral configuration aligns with the center wavelengths of Sentinel-2. Note that the Chl products are used for qualitative analysis only and reducing its uncertainties is out of the scope of this work.

To assess the sensitivity of DOC variations to changes in precipitation, riverine discharge, tidal height, and Chl, we employed an elasticity factor model to quantify the relationships between the relative changes of DOC and these influencing factors (Sankarasubramanian et al., 2001). The elasticity factor model offers a robust, non-parametric approach for assessing the sensitivity of DOC to relative changes in these influencing factors. For example, the climatic elasticity of DOC (e) with respect to precipitation (e_{Pr}) can be expressed as:

$$\frac{\mathrm{d}DOC/DOC}{\mathrm{d}Pr/Pr}=\frac{\Delta DOC_i/\overline{DOC}}{\Delta Pr_i/\overline{Pr}}=\frac{\left(DOC_i-\overline{DOC}\right)/\overline{DOC}}{\left(Pr_i-\overline{Pr}\right)/\overline{Pr}}, \tag{4}$$

where, DOC_i represents an individual DOC concentration estimated by the AutoGluon-DOC model, and Pr_i represents an individual precipitation observation during January 2021 to December 2023. Correspondingly, the $\overline{DOC}$ and $\overline{Pr}$ were calculated as the multi-year means over the same observation period. The numerator and denominator indicate the relative changes in DOC and precipitation compared to their long-term averages. The climatic elasticity value is statistically significant only when a significant correlation ($P < 0.05$) exists between the long-term $\Delta DOC_i/\overline{DOC}$ and $\Delta Pr_i/\overline{Pr}$.

Similarly, the elasticity of DOC with respect to riverine discharge (e_{Rd}), tide height (e_{Th}), and Chl (e_{Chl}) were calculated via Eq. (5) by replacing the precipitation data with the respective factor data. The precipitation, river discharge, and tidal height data used in the elasticity analysis were preprocessed prior to the analysis. Precipitation data were aggregated from daily to monthly totals to examine cumulative effects and account for potential lagged DOC responses. River discharge data were provided at a monthly resolution and were used without further resampling. Tidal height data, available at an hourly resolution, were linearly interpolated to the satellite overpass time (UTC 02:51) to align with the instantaneous satellite-derived AutoGluon-DOC estimates. Notably, since Chl concentrations were derived directly from the same Sentinel-2 imagery as the DOC estimates, they provide perfectly synchronized biological information, allowing for a direct assessment of the coupling between phytoplankton dynamics and DOC variability without further resampling.

3 Results and discussion

3.1 Evaluation of AutoGluon-DOC

The AutoGluon-DOC model was first trained separately using the four compiled datasets (*i.e.*, Dataset-3.5d, Dataset-2.5d, Dataset-1d, and Dataset-3h) to evaluate its robustness across datasets compiled using different temporal windows. The scatter plots in **Fig. 3** illustrate the overall performance of AutoGluon-DOC on the four datasets.

As shown in **Fig. 3**, AutoGluon-DOC demonstrated overall consistent and good performance among the four training datasets. The MAPE of derived DOC remained relatively low (~ 10 %), suggesting the model's robustness in predicting DOC for the data collected in the coastal waters in Southeast China. As the temporal window narrowed, the number of matchups decreased significantly, while the overall performance of AutoGluon-DOC improved, except when the temporal window was set to ±3 hours. However, with only 11 samples in Dataset-3h, the resulting MAPE of derived DOC (= 10.2 %) may lack statistical significance. Thus, balancing the overall sample size and the accuracy of retrieved DOC, the AutoGluon-DOC trained on Dataset-2.5d was ultimately selected for subsequent analyses. In addition, the ±2.5-day temporal window corresponds to the revisit cycle of the MSI sensor, increasing the likelihood of pairing each field DOC observation with at least one valid satellite-measured $\rho_{rc}(\lambda)$, assuming cloud-free conditions.

To identify the most influential spectral and temporal variables for DOC estimation, we conducted a feature sensitivity analysis by calculating importance scores for each predictor within the AutoGluon-DOC model. These scores were quantified using the permutation importance method, which evaluates a feature's contribution by measuring the degradation in model performance when its values are randomly shuffled. A more pronounced decrease in predictive accuracy signifies a higher model dependency on that specific feature (Molnar et al., 2024).

As summarized in **Table 1**, the analysis provides a comprehensive ranking of feature contributions to the DOC retrieval model. The results reveal that imagery acquisition time is the most critical variable, yielding an importance score of 0.035. When the model was executed without temporal features, it experienced a significant degradation in performance, with the MAPE increasing by 2 % compared to the version with temporal features. This finding suggests that temporal dynamics, likely associated with seasonal variations in hydrological runoff or biological processes, play a dominant role in governing DOC fluctuations. Regarding spectral influence, the model demonstrates the highest sensitivity to short-wavelength bands, specifically B1 and B2 with scores of 0.018 and 0.012, respectively. This is physically consistent with the optical properties of CDOM, which is characterized by strong absorption in the blue and ultraviolet bands (Yu et al., 2016). Furthermore, while the near-infrared bands, specifically B8 and B8A, exhibit relatively lower importance scores (0.009 and 0.008), they are fundamental for ensuring the accuracy of the atmospheric correction process (Lai et al., 2025).

3.2 Validation of AutoGluon-DOC in Dongshan Bay

Since most field data was collected from DSB, this region was selected to evaluate the performance of AutoGluon-DOC. To evaluate the reliability of AutoGluon-DOC, time series of DOC derived from all cloud-free MSI images were compared with *in situ* measurements during 2021-2023 (**Fig. 4**).

Due to the lack of continuous long-term monitoring at fixed stations, four regions of interest (ROIs) were utilized to pool synchronous *in situ* measurements within a ±2.5-day window (consistent with the criteria in Section 3.1). This ROI-based approach provides a statistically representative characterization of sub-regional water masses, with the mean and one standard deviation (± 1 std) of *in situ* DOC shown in **Fig. 4**. These ROIs were strategically positioned along a transect from the Zhangjiang River Estuary (ZRE) to the bay mouth **(Fig. 5a**) to capture DOC variability while avoiding aquaculture-induced interference (blue cross-hatched areas). Comparing these aggregated field data with MSI-derived time series (2021-2023) demonstrates the model's reliability in capturing realistic spatial and temporal DOC patterns (**Fig. 4**).

As shown in **Fig. 4**, the time series DOC derived by AutoGluon-DOC showed overall good agreements with the *in situ* measurements across four ROIs in the DSB from January 2021 to December 2023 (**Fig. 4**). However, there are a few exceptions where the derived DOC was notably overestimated, such as retrievals in ROI_1 (2023) and ROI_3-4 (May-June 2022). These overestimations likely stem from tidal-driven hydrodynamic disturbances that trigger rapid DOC fluctuations (Cao and Tzortziou, 2021). Specifically, for these specific outliers, the tidal height differences between satellite overpasses and field samplings ranged from 24 to 64 cm, indicating that measurements were captured during divergent tidal phases. Such phase-related shifts in water mass properties significantly heighten the uncertainty of satellite-based DOC retrievals.

Despite these localized discrepancies, the MSI-derived DOC remained in strong agreement with *in situ* measurements across all four ROIs. Building on this reliability, linear regression analysis of the 2021-2023 time series revealed distinct and significant long-term trends across the ROIs. while ROI_1 and ROI_2 exhibited highly significant declines ($P < 0.01$, slopes of -0.0004 and -0.0003, respectively), a similarly significant, albeit slightly weaker, decrease was observed in ROI_3 ($P < 0.05$). In contrast, the trend in ROI_4 was not statistically significant ($P \geqslant 0.05$), suggesting that DOC levels in this region remained relatively stable compared to the others. The underlying factors driving the declining trend of DOC at the bayhead are discussed in Section 3.5. The progressive decrease in slope magnitude from ROI_1 to ROI_4 reveals a clear spatial gradient in long-term DOC variability within the bay.

3.3 Spatial variations of DOC

AutoGluon-DOC was further applied to MSI images to analyze the spatiotemporal variations in DOC distribution within DSB. Here, we selected two cloud-free MSI images per season from 2021 to 2023, representing low- and high-tide scenarios, to assess the spatiotemporal variations of DOC and the influence of tide on DOC distribution in DSB.

As shown in **Fig. 5**, the spatial distribution of DOC in DSB exhibits a clear gradient, with higher concentrations near the bay head and lower concentrations toward the bay mouth. This spatial pattern is highly consistent with previous local research. Specifically, Guo et al. (2020) reported that DOC concentrations gradually increase from the upper to lower reaches of the Zhangjiang River, peak within the mangrove creek zone, and then decline along the estuary toward the bay. Such a seaward-decreasing trend is a characteristic feature of most nearshore systems. Other studies have similarly documented that DOC levels typically decline as water moves from the estuary into the bay and eventually toward offshore areas (Cao and Tzortziou, 2021; Mannino et al., 2016). This reasonable spatial pattern further confirms the capability of AutoGluon-DOC to capture realistic DOC dynamics in DSB.

Generally, DOC concentrations in DSB are lower during high tide (1.31 - 1.49 mg/L) than during low tide (1.51 - 1.89 mg/L), with low-tide values about 1.2 times higher on average. A notable exception occurred in the summer of 2022, when DOC concentrations in the upper bay exceeded 2 mg/L despite high-tide conditions (**Fig. 5c**). This anomaly contrasts sharply with the typical tidal patterns observed in other seasons (**Fig. 6a, e, g**) and is likely driven by extreme precipitation. According to the discharge data in **Fig. 6a**, riverine flow during summer 2022 was roughly 2.5 times higher than in other years due to rainfall. Such extreme events can trigger massive flushing of terrestrial organic matter from upstream wetlands into the bay. Studies have shown that a single extreme precipitation event can elevate DOC concentrations to more than twice the baseline levels (Cao and Tzortziou, 2024; Letourneau and Medeiros, 2019). This intensified land-to-ocean transport significantly elevates DOC levels and can override the typical diluting effect of the incoming tide (D'Sa et al., 2023).

3.4 Seasonal variations of DOC

Seasonal averages of DOC in four ROIs were calculated using all available cloud-free MSI images acquired between 2021 and 2023, with each season comprising six scenes on average. Based on these seasonal composites, a clear seasonal cycle in DOC concentrations was observed, characterized by a decreasing trend from summer to winter (**Fig. 6**). These seasonal patterns are consistent with previous studies reporting similar variations in DOC concentrations in the Chesapeake Bay estuary and other coastal waters (Cao and Tzortziou, 2024; Lebrasse et al., 2025; Menendez et al., 2022).

A further interpretation of the downward seasonal shift in DOC concentrations is provided in Section 3.5.

Notably, in the summer of 2022, extreme rainfall caused a significant reduction in DOC concentrations in ROI_1, disrupting the characteristic seasonal cycle (**Fig. 6a**). Its spatial distribution is illustrated in **Fig. 5c**. Previous studies have documented that DOC concentrations in the upstream reaches are typically lower than those within the mangrove-dominated zone (Guo et al., 2020). Consequently, as shown in **Fig. 5c**, this decline was driven by strong dilution from upstream riverine runoff, which carried relatively lower DOC than the mangrove zone. In contrast, DOC concentrations in the downstream regions (ROI_2 and ROI_3) were higher than typical summer levels during the same period. This spatial pattern suggests that extreme rainfall events exert a dual effect: they dilute the existing DOC pool at the bay head while simultaneously flushing high-DOC mangrove water further into the bay (Letourneau and Medeiros, 2019; Zhou et al., 2023). Such event-driven anomalies demonstrate that although seasonal composites usually reflect steady-state hydrological transitions, extreme meteorological shocks can momentarily decouple DOC dynamics from their typical seasonal trajectories by overwriting baseline signals with transport-dominated signatures.

3.5 Driving factors of the spatiotemporal distribution of DOC in DSB

To further explore the influence of environmental drivers on DOC dynamics, elasticity models for precipitation, riverine discharge, tide height, and Chl were established based on Eq. (5). The elasticity models focused solely on the proportional response of DOC to a 1 % increase in these driving factors across the four ROIs. R^2 was used to assess the strength of the correlation, thereby evaluating the sensitivity of these factors in influencing DOC dynamics (**Fig. 8**).

The sensitivity analysis highlights a clear zonation of DOC drivers in Dongshan Bay, a phenomenon attributed to the bay’s status as a unique composite system integrating mangroves, estuaries, and aquaculture zones. In the upper bay (ROI_1 and ROI_2), DOC levels are mainly controlled by tides and Chl. Specifically, there is a significant negative correlation between DOC and tidal height in ROI_1 ($R^2 = 0.26$, $P < 0.01$) and ROI_2 ($R^2 = 0.11$, $P < 0.05$). This result is supported by time-series measurements at station S1 (**Fig. 8a**), which show a clear inverse trend between DOC and tidal height. This tidal pattern is consistent with the observations of Jiang et al. (2023) in the Zhangjiang River Estuary, who demonstrated that the mangrove pump drives the outwelling of carbon-rich pore water from mangrove soils into the bay during low tides. This outwelling process represents a major pathway for the transfer of organic matter from wetlands to the coastal ocean. Such mechanisms are common in many subtropical mangrove or salt marsh areas (Knoke et al., 2024; Liang et al., 2023; Xu and Zhang, 2025).

In parallel, Chl emerges as a pivotal biological driver, exhibiting strong positive correlations with DOC ($R^2 = 0.21$, $P < 0.01$ for ROI_1 and $R^2 = 0.16$, $P < 0.01$ for ROI_2). The coupling between Chl and DOC is a characteristic feature of mangrove-dominated ecosystems (English et al., 2025; Sharma et al., 2023), where algal exudation and subsequent microbial processing contribute to the dissolved carbon pool.

The analysis presented above provides a plausible explanation for the seasonal DOC cycle observed in the upper bay areas of ROI_1 and ROI_2. Since DOC is significantly related to Chl, the seasonal variations in DOC likely stem from biological phenology and the corresponding primary production. These observations align with the findings of Cao and Tzortziou (2021) for the Chesapeake Bay marsh estuarine system. Their study described a clear seasonal DOC pattern, with elevated concentrations in summer and autumn followed by a decline in winter. DOC concentrations were consistently higher in summer, coinciding with peak marsh plant biomass, and remained elevated into fall as vegetation began to senesce. In this study, a similar seasonal pattern is evident in the *in situ* DOC measurements from ROI_1, the sub-region in closest proximity to the mangrove fringe of DSB, which show a clear seasonal pattern with higher DOC concentrations in summer (1.52 ± 0.48 mg/L) and lower values in winter (1.11 ± 0.34 mg/L), consistent with the satellite-derived estimates. Mangroves are known to be important sources of dissolved organic matter in tropical and subtropical coastal waters (Dittmar et al., 2006), releasing DOC through root exudation, canopy leaching, and microbial decomposition of organic matter during the warm growing season. In winter, reduced vegetation activity and lower decomposition rates lead to diminished DOC inputs from the mangrove fringe, contributing to the observed seasonal minimum (Qiu et al., 2019; Zhu et al., 2019). A consistent seasonal cycle was also observed in ROI_2, ROI_3, and ROI_4, suggesting that this pattern reflects a bay-wide characteristic driven by the combined effect of biological production and mangrove-associated organic matter inputs.

Furthermore, the seasonal dynamics of organic matter in such systems are heavily influenced by the transport and transformation of carbon within mangrove and marsh environments. As depicted in **Fig. 6a**, the riverine runoff exhibits distinct seasonal fluctuations. However, the driver analysis in **Fig.7 a-b** indicates that the correlation between DOC variability and runoff is relatively low in the upper bay, where the R^2 value is nearly zero. Consequently, seasonal variability in DOC is primarily driven by biological activity cycles, with additional influence from organic matter transport during extreme weather events. On a shorter temporal scale, the daily spatial distribution of DOC is jointly regulated by tidal fluctuations and Chl concentrations (García-Martín et al., 2023; Zhao et al., 2023b).

The underlying factors driving the declining trend of DOC at the bayhead remain unclear. One possibility is that DOC inter-annual variability is linked to Chl, given the significant relationship between DOC and Chl discussed above (**Fig.7 m-n**). However, our analysis did not reveal a significant declining trend in Chl over the 2021-2023 period, which makes it difficult to attribute the DOC decline solely to changes in phytoplankton biomass. This discrepancy may stem from uncertainties in satellite-derived Chl retrieval in optically complex coastal waters, or from the relatively short time series (2021-2023) used here, which may be insufficient to separate long-term trends from natural inter-annual variability. Clarifying the drivers of DOC trends at the bayhead and extending the analysis with longer time series and improved Chl algorithms will be addressed in future work.

On the other hand, the driving mechanisms are different at the bay mouth (ROI_3 and ROI_4). In these regions, the influence of tides and Chl becomes much weaker. Instead, riverine discharge shows a greater statistical influence on DOC ($P < 0.05$ in **Fig. 7g-h**). Time-series data from fixed station S2 (**Fig. 8b**) also show that the expected inverse relationship between DOC and tides is not evident. This shift is closely linked to the intensive aquaculture activities in Dongshan Bay. Cai et al. (2023) reported that sediment total organic carbon concentrations in these specific aquaculture zones were notably high (0.77 %-1.08 %), with carbon-to-nitrogen ratios ranging from 7.0 to 9.0, indicating predominantly endogenous inputs such as excessive feed and fish feces. During periods of high-water flow, strong currents can resuspend organic material from the sediment (Cai et al., 2023; Wang et al., 2025), acting as a fluctuating source that drives irregular variations in DOC levels. Consequently, in such aquaculture zones and under the broader context of seasonal runoff, human activities effectively override both tidal and biological signals.

3.6 Limitations and perspectives

Although AutoGluon-DOC achieved promising performance, several limitations must still be considered. One major limitation concerns the temporal matching window used to construct the dataset for model training. Due to the high-frequency variability of DOC concentrations within DSB, driven primarily by tidal dynamics, expanding the temporal window can lead to mismatches between satellite-derived and *in situ* measurements. This mismatch introduces errors into the retrieval results, particularly when DOC concentrations differ substantially across tidal phases.

Another limitation is associated with the insufficient temporal resolution of the remote sensing observations. Accurately capturing seasonal DOC dynamics requires frequent, high-quality observations. However, frequent cloud cover during the local rainy season, combined with the non-daily revisit cycle of MSI sensors, resulted in a limited number of low-tide measurements in this study. This limitation may lead to seasonal average DOC concentrations that do not fully reflect true conditions and restricts the ability to finely resolve seasonal variability. In particular, DOC concentrations are generally higher during low-tide periods, which may not have been adequately captured. To overcome this issue, future efforts should integrate multi-source remote sensing data to enhance temporal resolution, thereby enabling more accurate assessments of seasonal mean states and long-term DOC dynamics. This framework could be extended to moderate-resolution sensors with higher revisit frequencies, such as VIIRS and OLCI, and further combined with neural network-based approaches (*e.g.*, transfer learning) to learn the mapping between lower- and higher-spatial-resolution observations, thereby deriving higher-spatial-resolution DOC products.

Future efforts will focus on enhancing model performance through multiple improvements. First, although AutoGluon-DOC has demonstrated good capability in handling small-sample datasets, machine learning models are inherently data-dependent. Expanding the training dataset with high-quality, matched samples across different water types (*e.g.*, highly turbid waters) is expected to enhance model's robustness and generalizability. Second, incorporating additional environmental parameters closely related to DOC distribution could further improve retrieval accuracy. In coastal regions, salinity and colored dissolved organic matter (CDOM) typically exhibit conservative mixing behavior (Liu et al., 2018). Experimental results have shown a strong relationship between coastal DOC concentrations and salinity. Integrating salinity as an input feature has been reported to improve DOC retrieval accuracy in several studies (Laine et al., 2024; Lønborg et al., 2025). However, obtaining reliable, high-spatial-resolution salinity data in coastal waters remains a significant challenge (Chen and Hu, 2017; Kim et al., 2020). The development of high-quality salinity datasets is therefore essential before salinity-based DOC retrieval can be fully implemented. In future applications, salinity will be included as an additional input to evaluate its contribution to enhancing model performance in coastal environments.

4 Conclusions

In this study, a machine learning model, termed AutoGluon-DOC, was developed based on the AutoGluon framework to efficiently and accurately retrieve DOC concentrations in coastal waters. The model was trained on a large dataset consisting of field-measured DOC and matched $\rho_{rc}(\lambda)$ measurements from Sentinel-2 MSI, obtained through Google Earth Engine. As a demonstration, AutoGluon-DOC was applied to a series of Sentinel-2 MSI images over Dongshan Bay (DSB) from 2021 to 2023, revealing pronounced spatiotemporal variability in DOC distribution. Spatially, a clear decreasing gradient in DOC concentrations was observed from the bay head toward the bay mouth. Temporally, DOC followed a seasonal cycle, with higher concentrations in summer and lower levels in winter.

Analysis of drivers across the bay showed that, under typical conditions, the spatial distribution of DOC was jointly regulated by tidal fluctuations and Chl in the bay head, and by aquaculture activities overriding these natural drivers in the bay mouth. The seasonal pattern was likely driven mainly by biological production and could be strongly influenced by river discharge during extreme rainfall events. These results demonstrate that AutoGluon-DOC is an effective tool for high-spatial-resolution monitoring of DOC from satellite data and for tracking DOC dynamics in complex coastal systems, with a framework that can be extended to other regions as the training dataset is expanded.

Declaration of competing interest

1) 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.

Acknowledgements

This work was supported by the Guangxi Key R&D Program of China (No. GUIKE AB25069453), the National Key Research and Development Program of China (No. 2022YFC3105402), the Science and Technology Program of Fujian Province, China (No. 2022L3001), the Fundamental Research Funds for the Central Universities (No. 20720240102), the Science and Technology Program of Xiamen, China (No. 3502Z20226021), and the MEL Internal Program (No. MELRI2501). We thank Ms. Ying Su for her assistance during the execution of the above projects. A portion of the data and samples was collected during the open research cruises NORC2023-11, NORC2025-02, NORC2024-04, supported by NSFC Shiptime Sharing Project (project numbers: 42249911, 42449902, 42349904).

Appendix A Supplementary data

Supplementary data associated with this article can be found in the online version at xxxxxx.

Authors' contributions

Lingling Li: data processing and analysis, original draft writing, and manuscript revision. Xiaolong Yu: data collection, research design, manuscript editing, and revision. Wendian Lai: data collection, processing and manuscript revision. Nengwang Chen, Weidong Guo, Hongyan Bao, Guizhi Wang, Shuiying Huang, Wenlu Lan and Xiaoyan Peng: data collection and manuscript revision. Zhongping Lee: research design and manuscript revision.

Data and Code Availability Statements

Satellite measurements from the Sentinel-2 MSI are available from the Copernicus Data Space Ecosystem (https://browser.dataspace.copernicus.eu/). The *in situ* measurements of DOC are under proprietary control but can be shared upon reasonable request. The source code for models developed in this study is publicly available at https://github.com/kilig00/DOC.

References

Altuntas, O.Y., Saygi, M., Colkesen, I., 2025. Long-term monitoring of coastal water quality using Sentinel-2 satellite images and Google Earth Engine: The case study Izmir and Erdek Bays. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 48, 37–43.

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, 12–23.

Bonelli, A.G., Loisel, H., Jorge, D.S.F., 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.

Caballero, I., Fernández, R., Escalante, O.M., Mamán, L., Navarro, G., 2020. New capabilities of Sentinel-2A/B satellites combined with in situ data for monitoring small harmful algal blooms in complex coastal waters. Sci. Rep. 10, 8743.

Cai, Z.Z., Zeng, J., Yang, M.F., Lin, Y.Q., Zheng, H.D., Luo, D.L., et al., 2023. Vertical distribution and pollution assessment of TN, TP, and TOC in the sediment cores of cage farming areas in Dongshan Bay of southeast China. Front. Environ. Sci. 11, 1216868.

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.

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. Biogeosci. 129, e2023JG007767.

Cao, F., Tzortziou, M., Hu, C.M., Mannino, A., Fichot, C.G., Del Vecchio, R., et al., 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.

Chen, S.L., Hu, C.M., 2017. Estimating sea surface salinity in the northern Gulf of Mexico from satellite ocean color measurements. Remote Sens. Environ. 201, 115–132.

D'Sa, E.J., Tzortziou, M., Liu, B.Q., 2023. Extreme events and impacts on organic carbon cycles from ocean color remote sensing: Review with case study, challenges, and future directions. Earth Sci. Rev. 243, 104503.

Dai, M.H., Yin, Z.Q., Meng, F.F., Liu, Q., Cai, W.J., 2012. Spatial distribution of riverine DOC inputs to the ocean: An updated global synthesis. Curr. Opin. Environ. Sustain. 4, 170–178.

Das, S., Nandi, D., Thakur, R.R., Bera, D.K., Behera, D., Durin, B., et al., 2024. A novel approach for ex situ water quality monitoring using the Google Earth Engine and spectral indices in Chilika Lake, Odisha, India. ISPRS Int. J. Geo-Inf. 13, 381.

Dittmar, T., Hertkorn, N., Kattner, G., Lara, R.J., 2006. Mangroves, a major source of dissolved organic carbon to the oceans. Global Biogeochem. Cycles 20, GB1012.

English, C.J., Bell, T.W., Opalk, K., Siegel, D.A., Carlson, C.A., 2025. Senescence-driven solubilization of biomass is the main source of kelp-derived dissolved organic carbon to the coastal ocean. Commun. Biol. 8, 1172.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., et al., 2020. Autogluon-tabular: Robust and accurate automl for structured data. arXiv. https://doi.org/10.48550/arXiv.2003.06505 preprint arXiv:2003.06505.

Feng, L., Hou, X.J., Li, J.S., Zheng, Y., 2018. Exploring the potential of Rayleigh-corrected reflectance in coastal and inland water applications: A simple aerosol correction method and its merits. Int. J. Photogramm. Remote Sens. 146, 52–64.

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.

Foltz, G.R., Eddebbar, Y.A., Sprintall, J., Capotondi, A., Cravatte, S., Brandt, P., et al., 2025. Toward an integrated pantropical ocean observing system. Front. Mar. Sci. 12, 1539183.

García-Martín, E.E., Sanders, R., Evans, C.D., Kitidis, V., Lapworth, D.J., Spears, B.M., et al., 2023. Sources, composition, and export of particulate organic matter across British estuaries. J. Geophys. Res. Biogeosci. 128, e2023JG007420.

Gordon, H.R., 1997. Atmospheric correction of ocean color imagery in the Earth Observing System era. J. Geophys. Res. Atmos. 102, 17081–17106.

Guo, W.D., Wang, C., Li, Y., Qu, L.Y., Lang, M.C., Deng, Y.B., et al., 2020. Characterization of aquatic dissolved organic matter by spectral analysis: From watershed to deep ocean. Adv. Earth Sci. 35, 933.

Hu, C.M., 2011. An empirical approach to derive MODIS ocean color patterns under severe sun glint. Geophys. Res. Lett. 38, L01603.

Jiang, W.Z., Wang, G.Z., Li, Q., Dutta, M.K., Jin, S.L., Dai, G.Y., et al., 2023. The fate of carbon resulting from pore water exchange in a mangrove and Spartina alterniflora ecozone. Acta Ocean. Sin. 42, 61–76.

Kim, D.W., Park, Y.J., Jeong, J.Y., Jo, Y.H., 2020. Estimation of hourly sea surface salinity in the East China Sea using geostationary ocean color imager measurements. Remote Sens. 12, 755.

Knoke, M., Dittmar, T., Zielinski, O., Kida, M., Asp, N.E., de Rezende, C.E., et al., 2024. Outwelling of reduced porewater drives the biogeochemistry of dissolved organic matter and trace metals in a major mangrove-fringed estuary in Amazonia. Limnol. Oceanogr. 69, 262–278.

Lai, W.D., Lee, Z.P., Wang, J.W., Wang, Y.C., Garcia, R., Zhang, H.G., 2022. A portable algorithm to retrieve bottom depth of optically shallow waters from top-of-atmosphere measurements. J. Remote Sens. 2022, 9831947.

Lai, W.D., Yu, X.L., Chen, N.W., Zhang, C.Y., Wu, Y.F., Huang, S.Y., et al., 2025. Consistent nutrient mapping from Sentinel-2 and Sentinel-3 in nearshore waters: A case study in Xiamen Bay. Ecol. Inform. 92, 103470.

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, 1305050.

Lebrasse, M.C., Schaeffer, B.A., Bohnenstiehl, D.R., Osburn, C.L., Coffer, M.M., He, R.Y., et al., 2025. Winter-spring dynamics of dissolved organic carbon fluxes driven by precipitation in a North Carolina tidal marsh. Estuar. Coast. Shelf Sci. 322, 109361.

Letourneau, M.L., Medeiros, P.M., 2019. Dissolved organic matter composition in a marsh-dominated estuary: Response to seasonal forcing and to the passage of a hurricane. J. Geophys. Res. Biogeosci. 124, 1545–1559.

Li, Y., Liu, Z.L., Li, P., Zhou, X.L., Deng, Z.M., Zou, Z.Y., et al., 2025. Dissolved and particulate organic carbon transport among forest, river, and wetland ecosystems: A review of processes, controlling factors, challenges, and prospects. Environ. Rev. 33, 1–22.

Liang, Q.L., Wang, C., Li, Y., Xie, X.F., Guo, X.H., Liu, X., et al., 2025. Kuroshio-derived anticyclonic eddies drive lateral transport and transformation of oceanic dissolved organic matter along the continental slope of South China Sea. J. Geophys. Res. Oceans 130, e2024JC021718.

Liang, W.Z., Chen, X.G., Zhao, C., Li, L., He, D., 2023. Seasonal changes of dissolved organic matter chemistry and its linkage with greenhouse gas emissions in saltmarsh surface water and porewater interactions. Water Res. 245, 120582.

Liu, D., Bai, Y., He, X.Q., Pan, D.L., Wang, D., Wei, J.A., et al., 2018. The dynamic observation of dissolved organic matter in the Zhujiang (Pearl River) Estuary in China from space. Acta Ocean. Sin. 37, 105–117.

Lønborg, C., Fuentes-Santos, I., Carreira, C., Amaral, V., Arístegui, J., Bhadury, P., et al., 2025. Dissolved organic carbon in coastal waters: Global patterns, stocks and environmental physical controls. Global Biogeochem. Cycles 39, e2024GB008407.

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, 312–336.

Menendez, A., Tzortziou, M., Neale, P., Megonigal, P., Powers, L., Schmitt-Kopplin, P., et al., 2022. Strong dynamics in tidal marsh DOC export in response to natural cycles and episodic events from continuous monitoring. J. Geophys. Res. Biogeosci. 127, e2022JG006863.

Molnar, C., König, G., Bischl, B., Casalicchio, G., 2024. Model-agnostic feature importance and effects with dependent features: A conditional subgroup approach. Data Min. Knowl. Discov. 38, 2903–2941.

Mopper, K., Zhou, X.L., Kieber, R.J., Kieber, D.J., Sikorski, R.J., Jones, R.D., 1991. Photochemical degradation of dissolved organic carbon and its impact on the oceanic carbon cycle. Nature 353, 60–62.

O'Reilly, J.E., Werdell, P.J., 2019. Chlorophyll algorithms for ocean color sensors-OC4, OC5 & OC6. Remote Sens. Environ. 229, 32–47.

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

Pahlevan, N., Mangin, A., Balasubramanian, S.V., Smith, B., Alikas, K., Arai, K., et al., 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.

Pahlevan, N., Sarkar, S., Franz, B.A., 2016. Uncertainties in coastal ocean color products: Impacts of spatial sampling. Remote Sens. Environ. 181, 14–26.

Pan, D.L., Liu, Q., Bai, Y., 2014. Review and suggestions for estimating particulate organic carbon and dissolved organic carbon inventories in the ocean using remote sensing data. Acta Ocean. Sin. 33, 1–10.

Qiu, Y., Wang, Y.Y., Xie, Y.Y., Huang, B.Q., 2019. Effects of environmental factors on primary production of microphytobenthos in the mangrove system of Zhangjiang Estuary, Fujian Province. J. Appl. Oceanogr. 38, 53–61.

Qu, L.Y., Dahlgren, R.A., Gan, S.C., Ren, M.X., Chen, N.W., Guo, W.D., 2024. Spatial variation of anthropogenic disturbances within watersheds determines dissolved organic matter composition exported to oceans. Water Res. 262, 122084.

Sankarasubramanian, A., Vogel, R.M., Limbrunner, J.F., 2001. Climate elasticity of streamflow in the United States. Water Resour. Res. 37, 1771–1781.

Sharma, S., Ray, R., Martius, C., Murdiyarso, D., 2023. Carbon stocks and fluxes in Asia-Pacific mangroves: Current knowledge and gaps. Environ. Res. Lett. 18, 044002.

Sullivan, C.W., Hartstein, N.D., Müller, M., 2025. Quantifying mangrove export of dissolved organic carbon on large scales and at fine resolution: A review of current technologies and the path forward. Discov. Oceans 2, 18.

Vanhellemont, Q., 2019. Adaptation of the dark spectrum fitting atmospheric correction for aquatic applications of the Landsat and Sentinel-2 archives. Remote Sens. Environ. 225, 175–192.

Wang, T.X., Ren, M.X., Qu, L.Y., Sun, S.Y., Jiang, Y.W., Li, Y., et al., 2025. Dynamic variation of dissolved organic matter along the Zhangjiang River-Dongshan Bay continuum. Acta Sci. Circumstantiae 45, 277–287.

Wei, J.W., Lee, Z.P., Shang, S.L., Yu, X.L., 2019. Semianalytical derivation of phytoplankton, CDOM, and detritus absorption coefficients from the Landsat 8/OLI reflectance in coastal waters. J. Geophys. Res. Oceans 124, 3682–3699.

Wu, K., Dai, M.H., Chen, J.H., Meng, F.F., Li, X.L., Liu, Z.Y., et al., 2015. Dissolved organic carbon in the South China Sea and its exchange with the Western Pacific Ocean. Deep Sea Res. Part II 122, 41–51.

Wu, Y., Bao, H.Y., Unger, D., Herbeck, L.S., Zhu, Z.Y., Zhang, J., et al., 2013. Biogeochemical behavior of organic carbon in a small tropical river and estuary, Hainan, China. Cont. Shelf Res. 57, 32–43.

Xiao, Q.T., Xu, X.F., Qi, T.C., Luo, J.H., Lee, X.H., Duan, H.T., 2024. Lakes shifted from a carbon dioxide source to a sink over past two decades in China. Sci. Bull. 69, 1857–1861.

Xu, Y.F., Zhang, Z.H., 2025. Tidally driven outwelling of dissolved carbon and nitrate from the largest mangroves of China. Front. Mar. Sci. 12, 1590259.

Yu, Q.B., Wang, F., Li, X.Y., Yan, W.J., Li, Y.Q., Lv, S.C., 2018. Tracking nitrate sources in the Chaohu Lake, China, using the nitrogen and oxygen isotopic approach. Environ. Sci. Pollut. Res. Int. 25, 19518–19529.

Yu, X.L., Shen, F., Liu, Y.Y., 2016. Light absorption properties of CDOM in the Changjiang (Yangtze) estuarine and coastal waters: An alternative approach for DOC estimation. Estuar. Coast. Shelf Sci. 181, 302–311.

Zhao, D., Feng, L., He, X.Q., 2023a. Global gridded aerosol models established for atmospheric correction over inland and nearshore coastal waters. J. Geophys. Res. Atmos. 128, e2023JD038815.

Zhao, W.Q., Bao, H.Y., Huang, D.K., Niggemann, J., Dittmar, T., Kao, S.J., 2023b. Evidence from molecular marker and FT-ICR-MS analyses for the source and transport of dissolved black carbon under variable water discharge of a subtropical Estuary. Biogeochemistry 162, 43–55.

Zhou, Y.Q., Yu, X.Q., Zhou, L., Zhang, Y.L., Xu, H., Zhu, M.Y., et al., 2023. Rainstorms drive export of aromatic and concurrent bio-labile organic matter to a large eutrophic lake and its major tributaries. Water Res. 229, 119448.

Zhu, X.D., Song, L.L., Weng, Q.H., Huang, G.M., 2019. Linking in situ photochemical reflectance index measurements with mangrove carbon dynamics in a subtropical coastal wetland. J. Geophys. Res. Biogeosci. 124, 1714–1730.

List of tables

Table 1 The feature importance scores of the AutoGluon-DOC model. Input features include the imagery acquisition time and the Rayleigh-corrected reflectance $\rho_{rc}(\lambda)$ at nine MSI bands (B1-B8A). Values in parentheses denote the central wavelengths of Sentinel-2A MSI/Sentinel-2B MSI (in nm), respectively.

Feature	Importance scores	Feature	Importance scores
Imagery acquisition time	0.035	B8A (865/864 nm)	0.008
B1 (443/442 nm)	0.018	B7 (783/780 nm)	0.007
B2 (492 nm)	0.012	B5 (704 nm)	0.007
B6 (740/739 nm)	0.011	B3 (560/559 nm)	0.006
B8 (833 nm)	0.009	B4 (665 nm)	0.002

List of figures

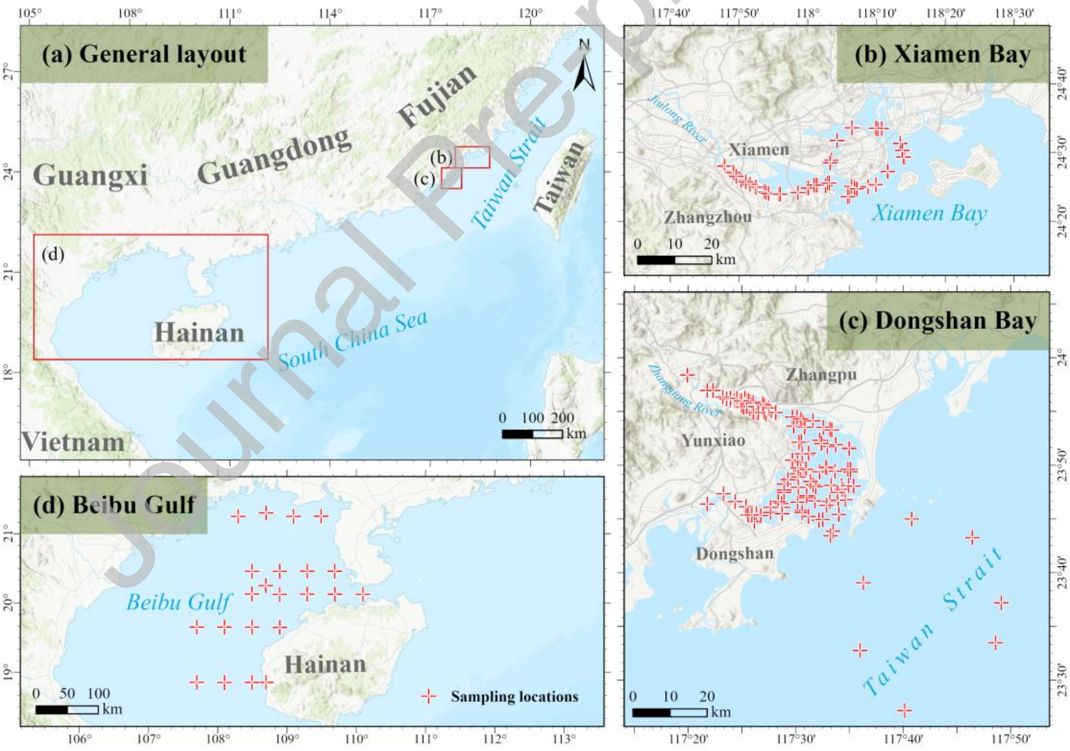


Fig. 1 Sampling locations (red crosses) of all field dissolved organic carbon (DOC) measurements used in this study, covering several coastal regions in Southeast China. General layout of all sampling locations (a) and detailed sampling locations in (b) Xiamen Bay, (c) Dongshan Bay, and (d) Beibu Gulf.

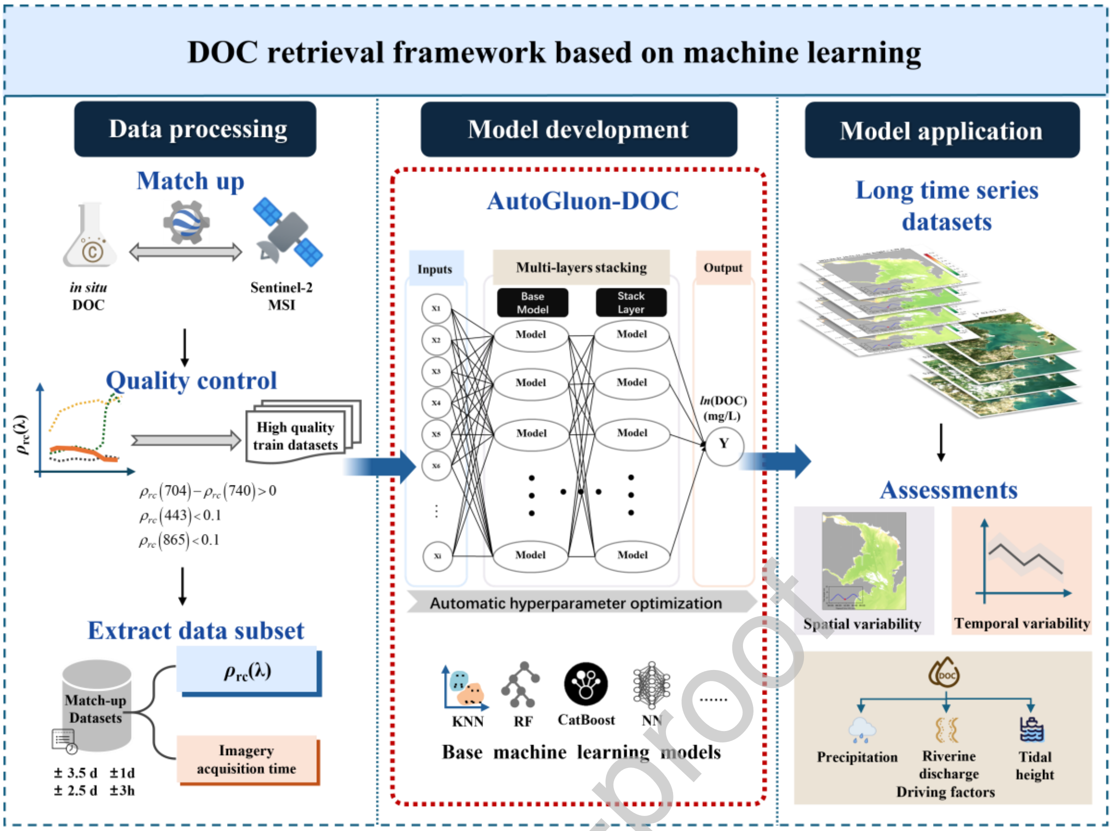


Fig. 2 Framework for DOC retrieval from Sentinel-2 MSI Rayleigh-corrected reflectance ($\rho_{rc}(\lambda)$) using Google Earth Engine and the automated machine learning model AutoGluon-DOC, including data processing, model development, and model application.

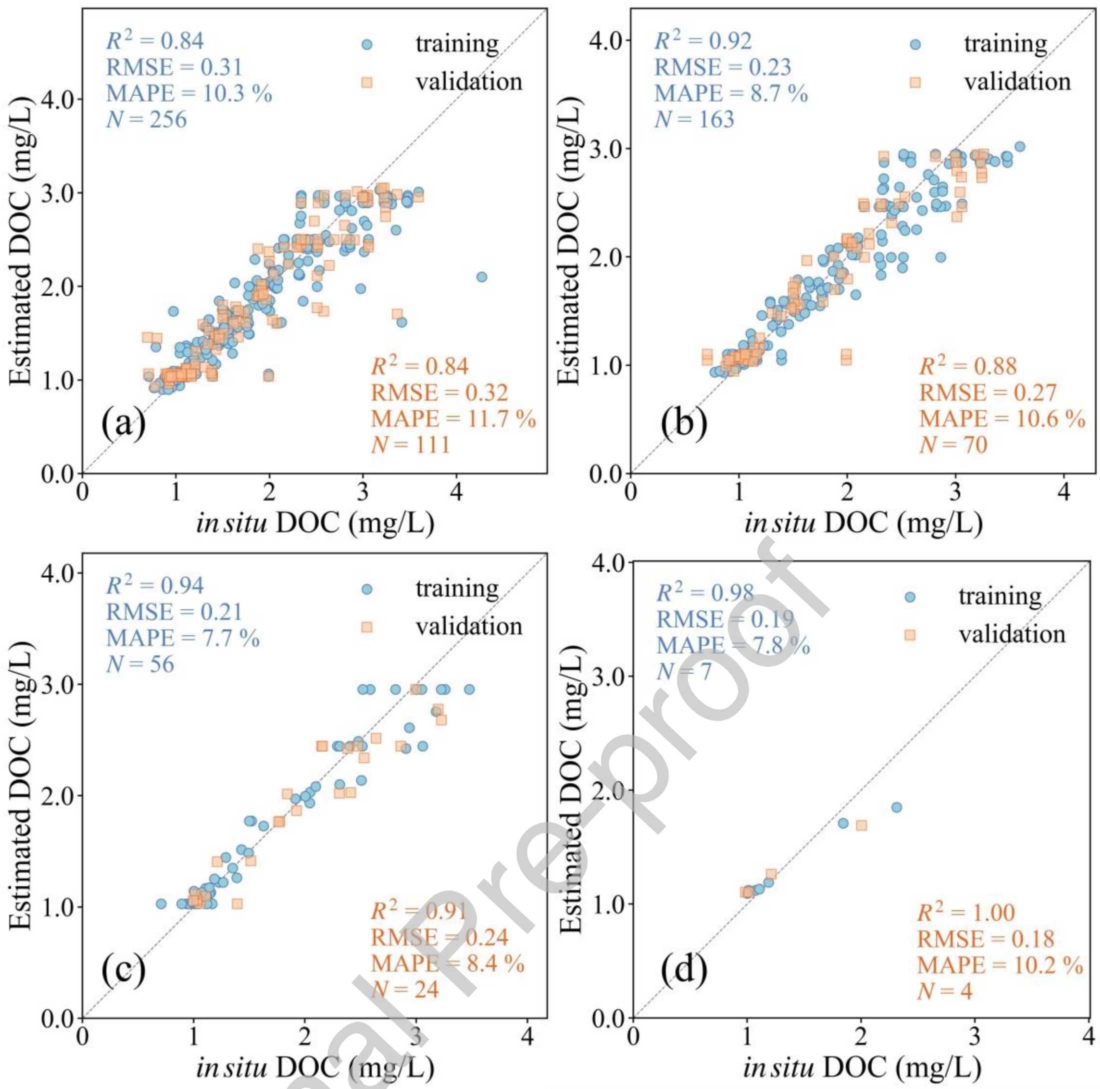


Fig. 3 Training and validation performance of the AutoGluon-DOC based on four training data subsets constructed across different temporal windows. (a) Dataset-3.5d, (b) Dataset-2.5d, (c) Dataset-1d, (d) Dataset-3h. Blue circles indicate training samples and orange squares indicate validation samples. The solid diagonal denotes the 1:1 line. Statistics shown in the upper-left and lower-right corners of each panel correspond to the training and validation sets, respectively, including the coefficient of determination (R^2), root mean square error (RMSE, mg/L), mean absolute percentage error (MAPE, %), and the number of samples (N).

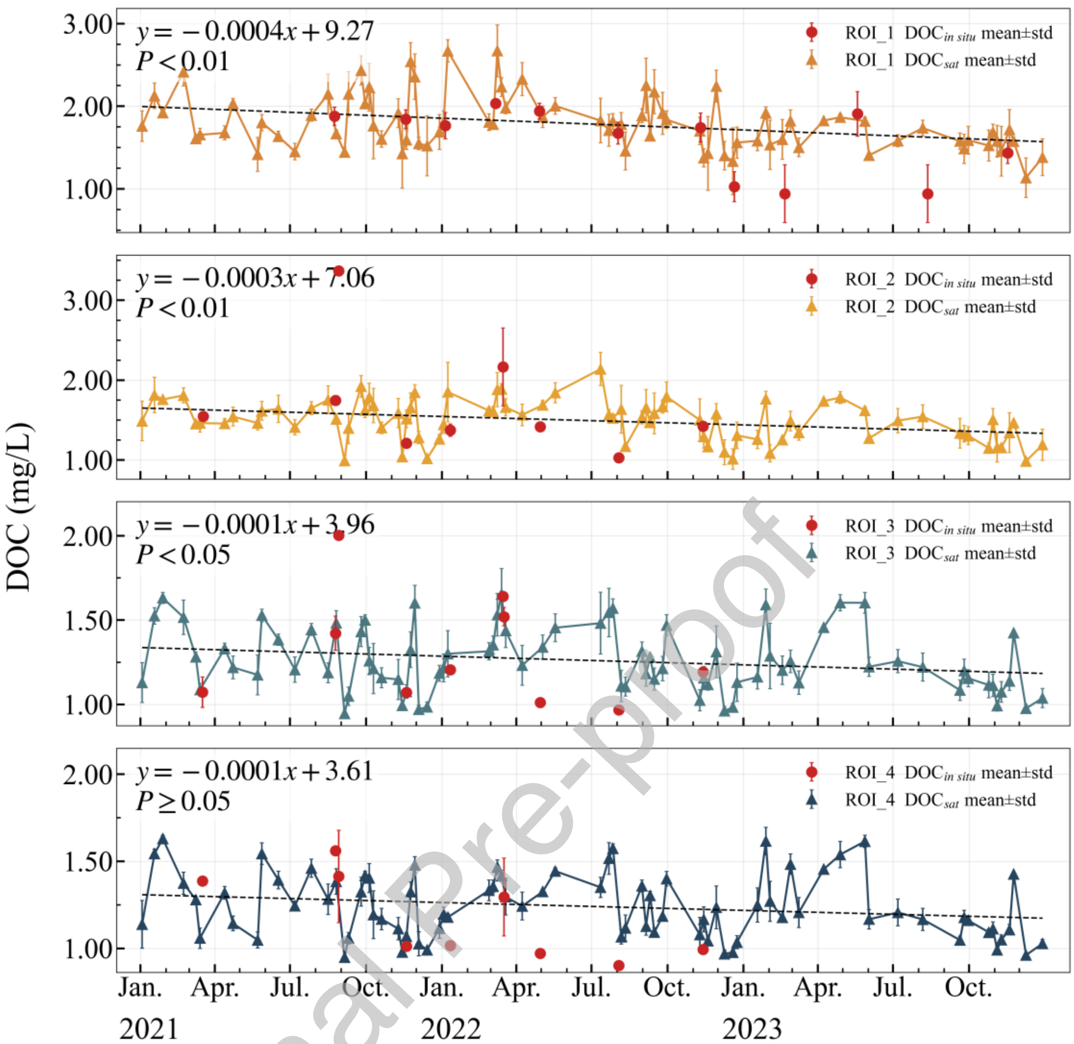


Fig. 4 Comparison between the time series of derived Dissolved Organic Carbon (DOC) retrieved by AutoGluon-DOC and *in situ* measurements across four Regions of Interest (ROIs) in Dongshan Bay (DSB) from January 2021 to December 2023. The red points and error bars represent *in situ*-measured DOC and the standard deviation within each ROI, respectively. Triangles and error bars represent mean DOC concentrations and the standard deviation in each ROI retrieved by AutoGluon-DOC. The black dashed lines indicate the linear trend of the derived DOC. The linear regression equations and statistical significance are displayed in the upper-left corner.

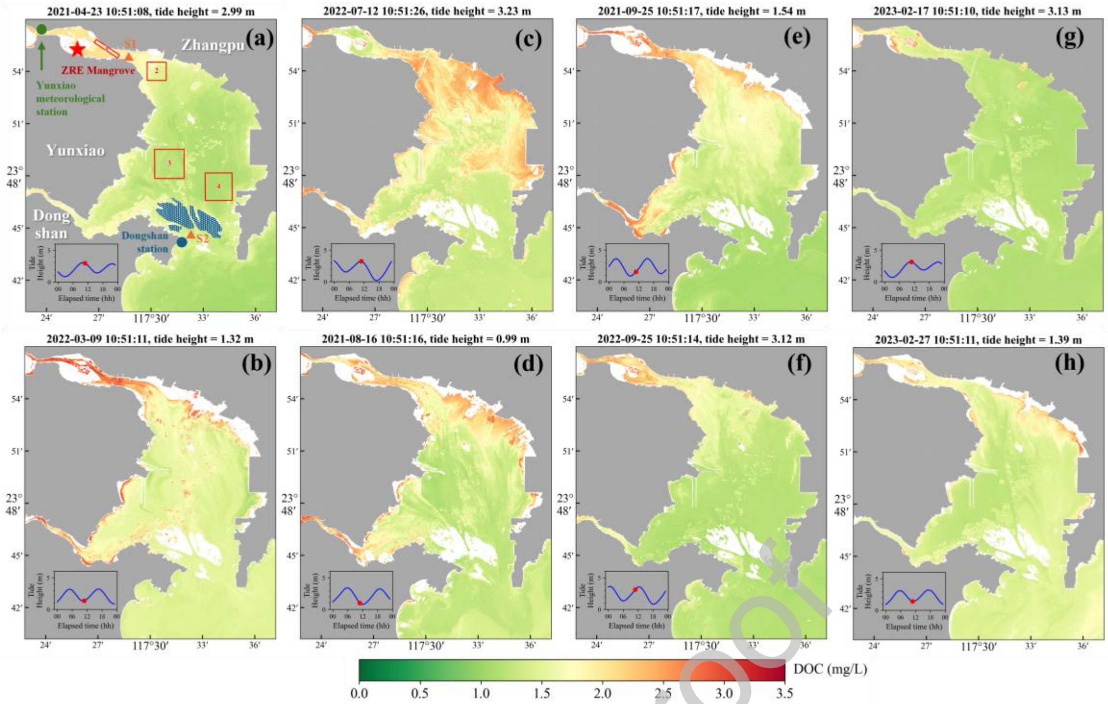


Fig. 5 Spatial distribution of Dissolved Organic Carbon (DOC) retrieved from Sentinel-2 MSI by AutoGluon-DOC in Dongshan Bay (DSB) during relatively high (the upper panel) and low tides (the lower panel) in (a-b) spring from March to May, (c-d) summer from June to August, (e-f) autumn from September to November, and (g-h) winter from December to the following February, respectively. In panel (a) (omitted in the others for clarity): red boxes mark the selected Regions of Interest (ROIs); the red star indicates the mangrove area; the green and blue dots denote the Yunxiao meteorological station and the Dongshan station, respectively; the orange triangles denote the two fixed stations used for time-series DOC analysis; and the blue cross-hatched areas indicate aquaculture zones. The inset in the lower-left corner of each panel shows the daily tidal cycle, with the red dot marking the tidal height at the time of acquisition for each MSI image.

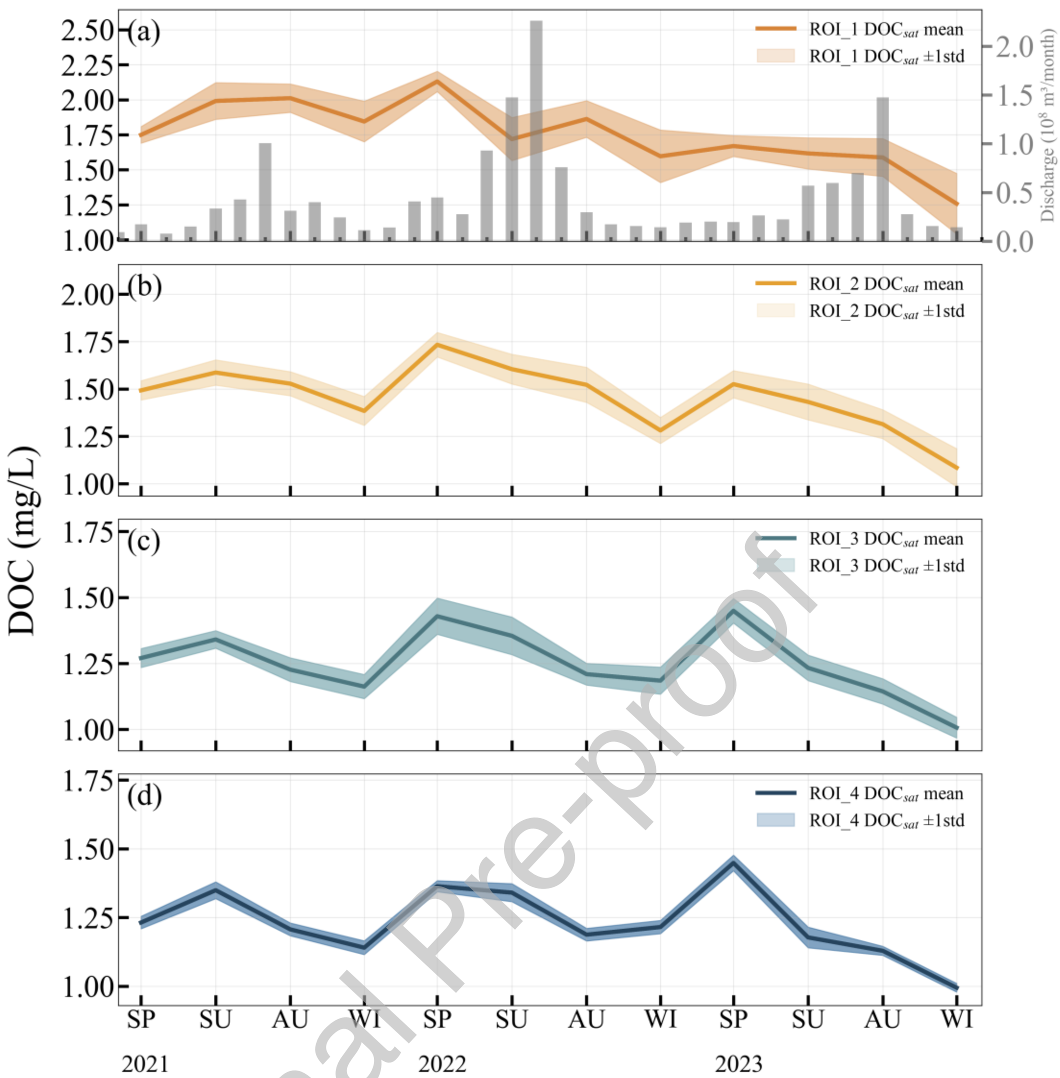


Fig. 6 Seasonal time series of AutoGluon-DOC-retrieved DOC concentrations averaged over four Regions of Interest (ROIs) in Dongshan Bay (DSB) from January 2021 to December 2023: (a) ROI_1, (b) ROI_2, (c) ROI_3, and (d) ROI_4. In each panel, the solid line represents the seasonal mean DOC and the shaded band denotes ± 1 standard deviation. Seasonal abbreviations on the x-axis: SP — spring (March-May); SU — summer (June-August); AU — autumn (September-November); WI — winter (December-February). The gray bars in panel (a), referenced to the right y-axis, show the corresponding monthly riverine discharge (10^8 m^3/month) at the Zhangjiang Estuary, the main freshwater source of DSB.

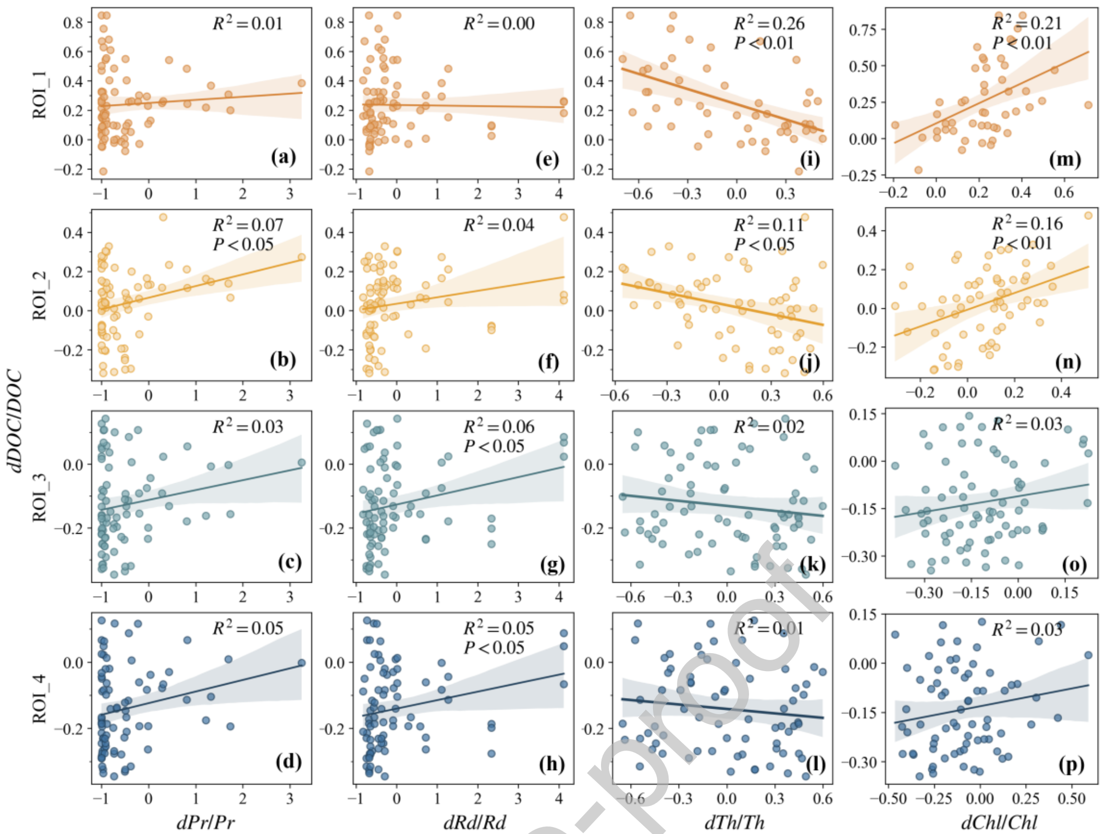


Fig. 7 Sensitivity analysis and correlation analysis for the four regions of interest (ROIs) in DSB. Sensitivity analysis of the relative changes in DOC concentration with respect to (a-d) precipitation (*Pr*), (e-h) riverine discharge (*Rd*), (i-l) tidal height (*Th*) and (m-p) chlorophyll-a concentration (*Chl*). In each panel, the solid line shows the linear best fit and the shaded areas represent the 95 % confidence interval of the best-fit line. R^2 quantifies the goodness of the linear fit between *dDOC*/*DOC* and *dPr*/*Pr*, *dRd*/*Rd*, *dTh*/*Th* or *dChl*/*Chl* as calculated using the elasticity model, and *P*-values are shown only when $P < 0.05$.

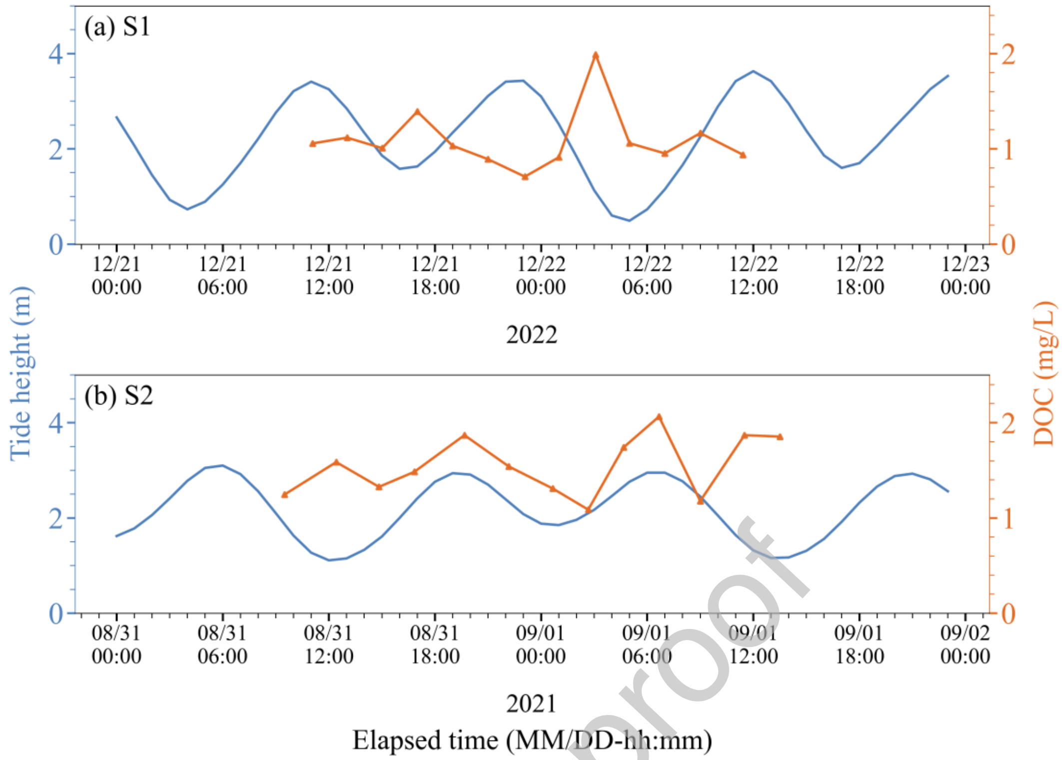
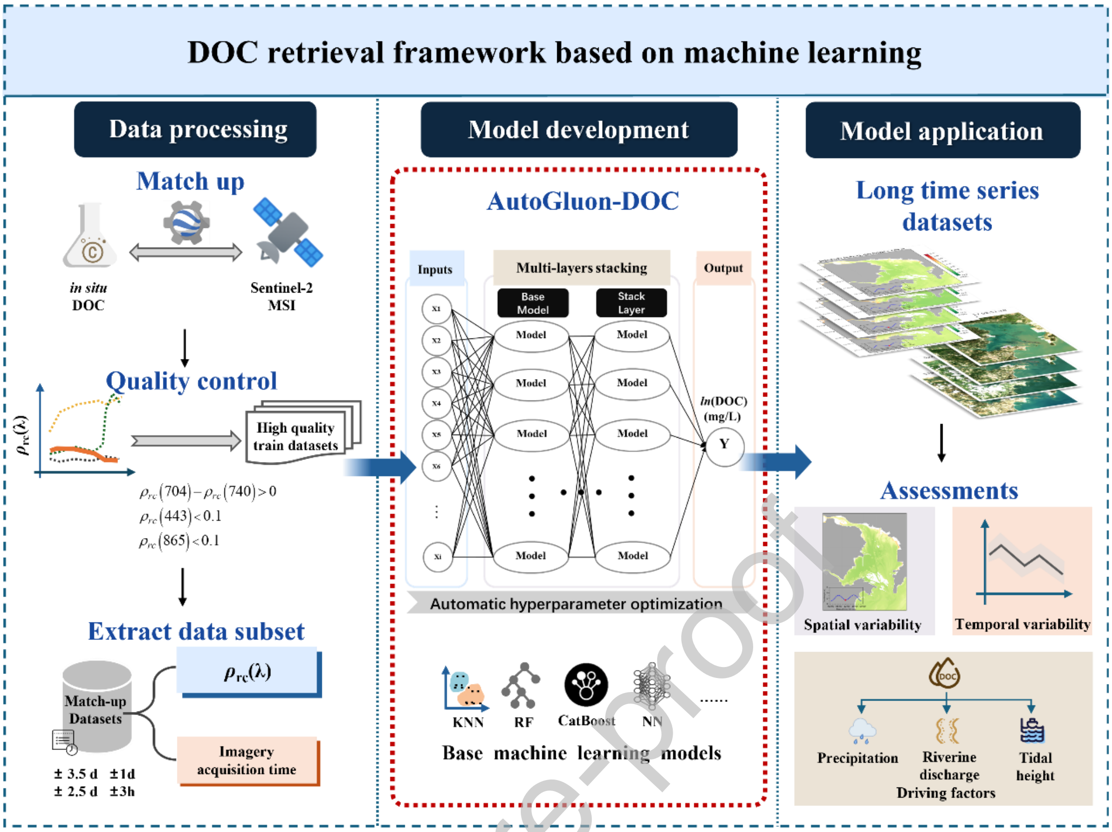


Fig. 8 Time series of tidal height (blue curves, left y-axis) and concurrent *in situ* DOC measurements (orange markers with connecting lines, right y-axis) at two fixed stations in Dongshan Bay (DSB): (a) station S1, from 21 to 23 December 2022; (b) station S2, from 31 August to 2 September 2021. Orange markers represent individual DOC samples. Tidal heights are from the Dongshan station. Station locations are shown in Fig. 5a.

Graphical abstract



Declaration of Interest Statement

☒ 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.