Ecological Informatics 87 (2025) 103094

Contents lists available at ScienceDirect

Ecological Informatics

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

Research on Atlantic surface pCO_2 reconstruction based on machine learning

Jiaming Liu, Jie Wang [*], Xun Wang, Yixuan Zhou, Runbin Hu, Haiyang Zhang

College of oceanography and ecological science, Shanghai Ocean University, Shanghai 201306, China

ARTICLE INFO

Keywords:
Atlantic Ocean
Ocean acidification
Sea surface carbon dioxide partial pressure
Machine learning

ABSTRACT

Ocean acidification is transforming marine ecosystems at an unprecedented rate, which in turn requires the estimation of sea surface carbon dioxide partial pressure (pCO_2) as a crucial metric to gauge acidification. This has substantial implications for marine resource assessment and management, marine ecosystems, and global climate change research. This study utilizes SOCAT cruise survey data to assess the accuracy of global sea surface pCO_2 products offered by Copernicus Marine Service and the Chinese Academy of Sciences Ocean Science Research Center. Through the application of a geographic information analysis method—geographical detector—the study quantitatively reveals the significance of environmental influencing factors, such as longitude, latitude, sea surface 10 m wind speed (U_{10}), total precipitation (TP), evaporation (E), and significant height of combined wind waves and swell (SHWW), in the reconstruction of sea surface pCO_2. Subsequently, various machine learning models, which include convolutional neural network (CNN), back propagation neural network (BP), long short-term memory network (LSTM), extreme learning machine (ELM), support vector regression (SVR), and extreme gradient boosting tree (XGBoost), are used to reconstruct the monthly sea surface pCO_2 data for the Atlantic Ocean from 2001 to 2020 to investigate the potential and suitability of high-precision reconstruction of the sea surface pCO_2 dataset for this sea area. The findings indicate that: (1) The geographical detector effectively quantifies the contribution of various environmental factors used in sea surface pCO_2 reconstruction. Notably, the Copernicus pCO_2 and CODC-GOSD pCO_2 contribute the most, with both contributing ~0.72. These are followed by TP, latitude, longitude, SHWW, U_{10}, and E. (2) After comprehensive data testing, the six machine learning models select the optimal hyperparameters for reconstruction. Among these, the XGBoost model notably improved the quality of the original dataset when using Copernicus pCO_2 and CODC-GOSD pCO_2 products in conjunction with SHWW, U_{10}, and TP environmental variable data. Compared with SOCAT data, the overall reconstruction accuracy in the Atlantic Ocean reached an impressive 94 %, outperforming the standalone use of either Copernicus pCO_2 or CODC-GOSD pCO_2 products. Furthermore, the XGBoost model demonstrated strong applicability in regions with numerous outliers, maintaining a reconstruction accuracy of $\geq$95 %. (3) Stability test results reveal that the XGBoost model exhibits low sensitivity to uncertainties in all input variables. This indicates that the model can accommodate environmental data errors induced by abrupt changes in marine environments. Such robustness enhances its reliability in sea surface pCO_2 reconstruction. The reconstruction of the Atlantic sea surface pCO_2 is conducive to the assessment of global ocean acidification and provides a theoretical basis for the sustainable development of the marine environment.

1. Introduction

Since the Industrial Revolution, the extensive utilization of fossil fuels has precipitated a sharp surge in global carbon emissions, notably elevating the concentration of atmospheric carbon dioxide (CO_2). This heightened presence of CO_2 has investigated a pronounced global greenhouse effect, jeopardizing both human civilization and the Earth's ecological equilibrium (Caldeira and Michael, 2003; James et al., 2005; Kevin, 2001; Richard et al., 2009). The sea-air interface, a significant carbon sink, facilitates the largest natural carbon exchange. Over the past decade, the ocean has sequestered an annual average of 2.78Gt of carbon, comprising 26 % of total anthropogenic carbon emissions

[*] Corresponding author.
E-mail addresses: m220200711@st.shou.edu.cn (J. Liu), wangjie@shou.edu.cn (J. Wang), m220200705@st.shou.edu.cn (X. Wang), m230200722@st.shou.edu.cn (Y. Zhou), m230200727@st.shou.edu.cn (R. Hu), m230200734@st.shou.edu.cn (H. Zhang).

https://doi.org/10.1016/j.ecoinf.2025.103094
Received 7 September 2024; Received in revised form 25 February 2025; Accepted 25 February 2025
Available online 6 March 2025

(Friedlingstein et al., 2022). Consequently, the ocean is instrumental in maintaining the Earth's ecological balance and modulating climate change. Augmenting real-time surveillance and quantitative evaluation of sea-air CO2 flux is imperative for comprehending the dynamics of the global carbon cycle and for informing strategies aimed at achieving carbon neutrality (Liu et al., 2018; Song et al., 2023; Yu et al., 2023).

The sea-air carbon dioxide flux is typically quantified using the sea surface carbon dioxide partial pressure (pCO_2), a measure of the CO_2 content at equilibrium between the surface ocean water and the atmosphere (Chen et al., 2019). However, traditional ship measurements of sea surface pCO_2, which are sparse and have limited spatial resolution, introduce considerable uncertainties in estimating the marine carbon sink. Given the intricate and variable nature of the marine environment, there exists an intrinsic linkage between various environmental factors and sea surface pCO_2. This linkage is instrumental for modeling and forecasting the distribution and temporal variations of sea surface pCO_2. Thus, in the face of global climate change, leveraging multi-source data and sophisticated estimation models becomes crucial for generating long-term, high-quality reconstructions of sea surface pCO_2 across expansive marine regions (Bai et al., 2015; Chau et al., 2022; Krishna et al., 2020).

A large number of studies have confirmed that thermodynamic effects, biochemical effects, ocean circulation and air-sea exchange, human disturbance and continental margin input are important factors restricting the development of pCO_2 (Dixit et al., 2019; Zhong et al., 2021). Specifically, in terms of thermodynamic effects, total alkalinity (TA), dissolved inorganic carbon (DIC), sea surface temperature (SST), and sea surface salinity (SSS), as important indicators of seawater carbonate system, control the change of pCO_2 in surface seawater (Lee et al., 2006; Weiss, 1974; Yang et al., 2015), for example, extreme changes in SST in winter and summer can lead to a significant increase or decrease in sea surface pCO_2. In terms of biochemical effects, the biocalcification process of zooplankton and phytoplankton consumes a large amount of carbon in seawater (Fay and Mckinley, 2017; Reynaud et al., 2003; Salisbury et al., 2008), carbon in the surface layer of seawater is transferred and deposited to deep seawater, resulting in a decrease in surface pCO_2, which in turn accelerates the circulation of air-sea carbon flux. In addition, solar radiation and limiting nutrients, such as nitrogen, phosphorus, iron, and manganese, are directly related to the physiological process of phytoplankton photosynthesis, which is also important for the transport of carbon in seawater (Zhong et al., 2021). In terms of ocean circulation and air-sea exchange, sea breeze affects the absorption capacity of surface seawater for CO_2. The cold-water mass in the high-latitude sea area absorbs CO_2 in the atmosphere and sinks, with the change of upwelling and turbulence, it surges to the low-latitude sea area and rises to the sea surface to release CO_2 to enhance sea surface pCO_2 (Bates et al., 1998; Bates and Merlivat, 2001; Turk et al., 2013). In terms of human disturbance and continental margin input, coastal rivers carry a large amount of nutrients and inorganic carbon into the ocean. Coupled with human disturbance, the mechanism of pCO_2 increase or decrease on the coastal surface is complex, and it is difficult to carry out quantitative description. In practical research, the dominant pCO_2 changes corresponding to different ocean systems are often different, which is also the difficulty of large-scale pCO_2 reconstruction.

Within the realm of big data, machine learning serves as a potential tool for uncovering the intrinsic laws of the data, functioning as an extension and expansion of traditional statistical methods. Over recent years, its utilization in the environmental sector has garnered considerable attention from both academic and industrial research (Laith et al., 2024; Liu and Robert, 2005; Rana et al., 2021; Reusch et al., 2007; Richardson et al., 2003; Salim et al., 2023; Zafar et al., 2021). The capacity of machine learning algorithms to efficiently manage intricate interrelationships and deliver precise outcomes after simulating substantial volumes of data makes them particularly valuable in the estimation and reconstruction studies of sea surface pCO_2. Telszewski et al. (2009) employed the Self-Organizing Map neural network (SOM) in conjunction with measured data from the North Atlantic to reconstruct sea surface pCO_2 from 2004 to 2006. The spatial resolution was 1° × 1° in the North Atlantic, and the Root Mean Square Deviation (RMSD) was 11.6μatm. Moussa et al. (2015) utilized the feedforward neural network (FNN) using remotely sensed chlorophyll concentration (Chl), SST, and SSS data. This allowed them to obtain sea surface pCO_2 from 2001 to 2009 in the North Atlantic with a spatial resolution of 4 km × 4 km and an RMSD of 8.7μatm. Landschützer et al. (2016) used an enhanced SOM-FFNN based on SOM, incorporating SST, SSS, Chl, and ocean mixed layer depth (MLD) reanalysis and model data. They constructed global sea surface pCO_2 in various oceanic regions worldwide with a spatial resolution of 1° × 1° and an RMSD of 20μatm. Chen et al. (2019) integrated the downhill irradiance diffuse attenuation index (Kd), SST, SSS, Chl parameters, and measured data. They employed the regression tree ensemble (RFRE) algorithm based on random forests to obtain sea surface pCO_2 from 2002 to 2017 in the Gulf of Mexico with a spatial resolution of 1° × 1° and an RMSD of 9.1μatm. Dixit et al. (2019) used support vector regression (SVR) incorporating SST and SSS data into the model, reconstructing sea surface pCO_2 from 2011 to 2018 in the Bay of Bengal with a spatial resolution of 1° × 1° and an RMSD of 7.68μatm. Yu et al. (2023) developed an XGBoost algorithm based on a semi-analytical remote sensing model framework (MeSAA). They added an upwelling index related to SST (UI_{SST}) to estimate and obtain sea surface pCO_2 from 2003 to 2019 in the Bohai Sea, Yellow Sea, and East China Sea with a spatial resolution of 1° × 1° and an RMSD of 20μatm. Similar studies have been extensively carried out in various sea areas around the world, and diverse machine learning algorithms have provided new insights for the reconstruction of sea surface pCO_2 (Friedrich and Oschlies, 2009; Hales et al., 2012; Jo et al., 2012; Signorini et al., 2013; Nakaoka et al., 2013; Marrec et al., 2015; Rödenbeck et al., 2015; Lohrenz et al., 2018; Zhong et al., 2021). The above research indicates that it is entirely feasible to spatially reconstruct regional sea surface pCO_2 by synergizing multiple environmental variables. Notably, a large number of studies start from the environmental factors themselves that affect sea surface pCO_2, mostly choosing variables like SST and SSS that are strongly related. Although good results have been obtained in different regions, the sources of data acquisition are diverse and difficult to unify, limiting their widespread application.

Despite extensive efforts in models and algorithms, there are still some issues with the current estimation of sea surface pCO_2 over large areas, mainly reflected in the significant differences in the estimation accuracy of sea surface pCO_2 in different sea areas, with RMSD ranging from 10μatm to 90μatm. In fact, each sea area has unique marine processes dominating it, and most models tend to focus on localized parameters, which leads to their lack of adaptability to different water environments and makes it difficult to promote them on a large scale. The Atlantic Ocean, as the second largest ocean in the world, extends from south to north and is surrounded by the equator. Its distinct characteristics include symmetrical climates in the north and south and complete climate zones. Additionally, the interaction of factors such as ocean currents, atmospheric circulation, and sea-land contours makes the climate of each sea area vastly different. If we could quantify the role of the Atlantic Ocean in regulating CO_2 flux and ocean acidification by estimating sea surface pCO_2, it would have significant implications for the regulation of the marine ecological environment. Therefore, the focus of this study is to fully consider the generalized characteristics of the ocean, find appropriate parameters to simulate changes in the marine environment, and develop an empirical method with universal applicability and strong stability for estimating sea surface pCO_2 in the Atlantic Ocean, which after comprehensive accuracy verification will be extended to global sea areas.

The novel contributions of this study encompass: (1) the employment of geographical information analysis models to quantitatively identify and select significant environmental variables as parameters for machine learning models; (2) a reconstruction process based on established global sea surface pCO_2 data products, wherein the accuracy and stability of six distinct machine learning techniques are compared; and (3) an examination of the relationship between the reconstructed Atlantic sea surface pCO_2 and phenomena such as greenhouse effects and ocean acidification.

2. Data and methods

2.1. Cruise pCO_2 data

This study examines the Atlantic region, delineated by coordinates 66.5°S-66.5°N and 100°W-40°E. The cruise survey data on pCO_2 for the Atlantic is sourced from the Surface Ocean CO_2 Atlas (SOCAT). SOCAT offers an extensive dataset of carbon dioxide fugacity (fCO_2) variations in global surface oceans and coastal regions, meticulously quality-controlled by the international marine carbon research community (Bakker et al., 2016; Pfeil, 2013). This dataset, accessible globally, facilitates the quantification of marine carbon sinks and ocean acidification. The most recent SOCAT version encompasses observations spanning 1957 to 2022, comprising 35.6 million records from global oceans and coastal zones, alongside 7.2 million calibrated sensor readings. In comparison to the pCO_2 data from the ESTOC (European Station for Time series in the Ocean at the Canary Islands) and BATS (Bermuda Atlantic Time Series Research Station) stations in the North Atlantic, SOCAT offers robust support for long-term, large-scale investigations. For this study, we utilized the SOCAT version 2020 dataset, with the spatiotemporal distribution of the study area and cruise survey routes depicted in Fig. 1. (sourced from https://socat.info/). Given that the SOCAT solely provides fCO_2, it is necessary to adjust these values to pCO_2 in accordance with (Dickson et al., 2007).

$$pCO_2 = fCO_2 \bullet exp\left[-\frac{P_{atm}(B+2\delta)}{RT}\right] \tag{1}$$

$$\delta = (57.7 - 0.118T) \bullet 10^{-6} \tag{2}$$

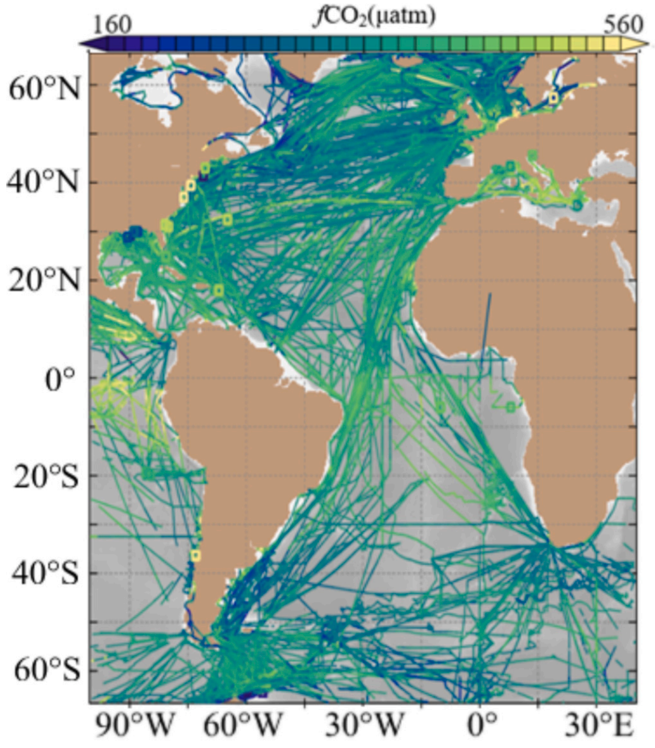

Fig. 1. The spatial distribution of the routes in the Atlantic Ocean.

$$B = \left(-1636.75 + 12.0408T - 3.27957 \bullet 10^{-2}T^2 + 3.16528 \bullet 10^{-5}T^3\right) \bullet 10^{-6} \tag{3}$$

In Eqs. (1), (2), and (3), pCO_2 signifies the partial pressure of carbon dioxide in the seawater surface layer, fCO_2 denotes the fugacity of carbon dioxide in the same layer, P_{atm} stands for atmospheric pressure, measured in Pascals, R is the ideal gas constant with a value of 8.314 J/(mol·K), B and δ represent correction coefficients associated with temperature T(K), measured in cubic meters per mole.

2.2. Reanalysis pCO_2 data

2.2.1. Copernicus data

The research employs surface pCO_2 data (Copernicus pCO_2) spanning 2001–2020, obtained from the Copernicus Marine Service (https://marine.copernicus.eu/). This data has been preprocessed and reanalyzed to yield monthly averaged results, with a water depth not exceeding 30 m for surface data and a spatial resolution of 0.25° × 0.25° (Table 1). The data utilized in this study is derived from the Global Ocean Biogeochemistry Hindcast dataset, available at the following website. This dataset employs a moderately complex biogeochemical model, PISCES, for its simulations. The simulation values generated from this model exhibit a high degree of consistency with Argo data and have been extensively employed in research focusing on the distribution and long-term alterations of the marine carbonate system (Monaco et al., 2021; Sridevi and Sarma, 2021). (See Table 2.)

2.2.2. CODC Global Ocean science data

The CODC-GOSD Marine Science Data, also known as the CODC Global Ocean Science Data, is a comprehensive global marine field observation dataset assembled by the Chinese Academy of Sciences Ocean Science Research Center (https://www.casodc.com/data/) (Zhong et al., 2021). The CODC has pioneered a unique marine observation data quality control system, termed CODC-QC, which systematically manages the quality of raw observations, thereby facilitating real-time precise monitoring of marine environmental conditions. Since 1900, the CODC has amassed a substantial volume of global marine observation data, encompassing 13 physical or biogeochemical elements such as SST, SSS, pCO_2, and others. The research center employs SOM to segment the global ocean into 11 distinct regions, and identifies prediction parameters that are intimately associated with sea surface pCO_2 in these various regions. Based on this, the center identifies the parameter combination that yields the lowest average error in predicting sea surface pCO_2, and subsequently uses FFNN to construct a global ocean surface pCO_2 grid data with a spatial resolution of 1° × 1° from January 1992 to the present (CODC-GOSD pCO_2) (Table 1).

2.2.3. Environmental data

The marine environment is inherently complex and dynamic. For a more accurate reconstruction of sea surface pCO_2, it is imperative to

Table 1
Summary of the input data used to produce high-quality pCO_2 data.

Data	Source	Variable	Spatial Resolution
SOCAT fCO_2	SOCAT	Surface Ocean fCO_2(μatm)	1° × 1°
Copernicus pCO_2	Copernicus	Surface Ocean pCO_2(μatm)	0.25° × 0.25°
CODC-GOSD pCO_2	CODC-GOSD	Surface Ocean pCO_2(μatm)	1° × 1°
E	ECMWF	Evaporation(mm)	0.25° × 0.25°
U_{10}	ECMWF	10 m wind speed(m/s)	0.25° × 0.25°
SHWW	ECMWF	Significant height of combined wind waves and swell(m)	0.5° × 0.5°
TP	ECMWF	Total precipitation(m)	0.25° × 0.25°

Table 2
Control parameters of the machine learning model.

Model	Parameter	Specification and range
BP	Learning rate	0.01
	Iterations number	1000
	Hidden layers and neurons	[12]
	Error threshold	1e-6
CNN	Learning rate	0.001
	Optimizer	SGDM
	Batch size	100
	Max epochs	30
	Learn rate drop factor	0.1
	Dropout layer	0.2
LSTM	Learning rate	0.01
	Optimizer	Adam
	Batch size	100
	Max epochs	60
	Learn rate drop factor	0.1
	Dropout layer	0.2
ELM	Activate model	Sigmoid
	Number of hidden neurons	50
SVR	c	4
	Gamma	0.8
XGBoost	Learning rate	0.1
	max_depth	5
	min_child_weight	1
	subsample	0.9
	colsample_bytree	1

consider various environmental factors that influence it. During periods of rainfall, the SSS in different marine regions decreases, which can indirectly impact the value of sea surface pCO_2. This, in turn, affects the accuracy of the reconstructed sea surface pCO_2 product. Similarly, seawater evaporation can lead to analogous outcomes (Jacob et al., 2019). Consequently, both precipitation and evaporation data should be incorporated into the reconstruction model to ensure its precision. This study utilizes the TP and E from the fifth generation of atmospheric reanalysis datasets (ERA5, https://cds.climate.copernicus.eu/) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) as crucial auxiliary data for reconstruction. Both datasets are single-layer monthly averages with a spatial resolution of 0.5° × 0.5° (Wang and Wang, 2022) (Table 1).

Research suggests that sea breezes can influence the capacity of surface seawater to absorb CO_2 at any given moment. Cold water masses are known to absorb atmospheric CO_2 before sinking, and subsequently rise due to upwelling and turbulence changes, transferring to other marine areas. This process releases CO_2 at the sea surface, thereby altering the pCO_2 of the marine area (Bates et al., 1998a; Bates and Merlivat, 2001; Turk et al., 2013). This study utilizes the ERA5-provided U_{10} data, input into the model to characterise this effect. The data used is single-layer monthly averaged with a spatial resolution of 0.5° × 0.5° (Table 1).

Jang et al. (2022) not only considered wind speed in their study of reconstructing global SSS but also used the significant wave height (SWH) as auxiliary data input into the model, given the close relationship between SSS and pCO_2. The impact of ocean waves is evidently indispensable. To further verify this conjecture, this paper selects the more intuitive SHWW as one of the variables input into the model to validate its role in reconstruction. The data, again derived from ERA5, is a single-layer monthly average with a resolution of 0.5° × 0.5°(Table 1).

2.3. Methods

2.3.1. Preprocessing of variables and data

The SOCAT fCO_2 is adjusted to align with the pCO_2 value as detailed in Eq. 2.1. This ensures that the spatial resolution, at 1° × 1°, aligns with that of both the Copernicus pCO_2 and CODC-GOSD pCO_2 data products. Subsequent variables undergo interpolation and resampling to achieve a uniform 1° × 1° resolution. Once all variables have this consistent spatial resolution, time is employed as a reference for further alignment, ensuring that both input and output datasets share the same spatio-temporal resolution. This study utilizes a total of 9247 cruise data entries from the SOCAT version 2020 dataset. Approximately 80 % of these data are designated for model training and optimization via hyperparameter cross-validation, while the remaining 20 % are set aside for independent model performance verification.

2.3.2. Geographical detector

In spatial analysis, a significant impact of an independent variable on a dependent variable implies a certain similarity in their spatial distribution (Wang et al., 2010). The geographical detector is a statistical method developed based on this premise to identify spatial differentiation characteristics and uncover potential driving forces (Ren et al., 2014; Todorova et al., 2016). To ensure that all selected environmental factors effectively reflect changes in sea surface pCO_2 and yield high-precision reconstruction results, it is essential to preliminarily screen the chosen variables. This paper employs the factor detector in the geographical detector to quantitatively examine the influence of each input parameter on sea surface pCO_2 after data discretization (Cao et al., 2013), measured using the Q value, which ranges from [0,1]. The closer the Q value is to 1, the stronger the explanatory power of the factor for the variable. Its calculation method is as follows (Luo et al., 2016):

$$Q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2} \tag{4}$$

In the formula, $h = 1, \ldots L$ denotes the partition of the dependent variable or factor. N_h and N represent the number of units in partition h and the total area respectively. The variances for partition h and the entire area are represented by σ_h^2 and σ^2, respectively.

2.3.3. Machine learning models

BP is a multi-layer feedforward neural network trained according to the error backpropagation algorithm, and is one of the most widely used neural network models. It consists of an input layer, hidden layers, and an output layer, and through repeated learning of training samples, it can continuously adjust the connection weights and thresholds between layers to ensure optimal output results (Ma and Liu, 2016). The excellent multidimensional function mapping capability of BP makes it fast and efficient in dealing with complex pattern problems, and various improved models have been widely applied in oceanographic research (Wang et al., 2021; Wang et al., 2023; Zhao et al., 2021).

CNN is a type of feedforward neural network with convolutional computation and a deep structure, which is an extended variation of the Multilayer Perceptron (MLP). It consists of an input layer, convolutional layer, pooling layer, and fully connected layer. The use of local connections and weight sharing reduces the number of weights, making the network easier to optimize, while also reducing the complexity of the model and thus the risk of overfitting (Zhao et al., 2024). The powerful feature extraction capability of CNNs not only demonstrates outstanding performance in tasks such as image classification and object detection, but also plays a crucial role in time series prediction and data regression (Krivoguz et al., 2024; Long et al., 2024).

ELM is a machine learning algorithm based on feedforward neural networks, which has significant advantages in terms of learning speed and generalization ability compared to other shallow learning systems. The innovation of ELM lies in the input weights and biases of its hidden layer nodes, which are randomly or manually set and remain unchanged throughout the learning process. This is a significant difference from conventional neural network algorithms that require iterative optimization of weights. ELM can be applied to both supervised learning tasks, such as classification and regression, as well as some unsupervised learning scenarios. It has application examples in fields such as computer vision, bioinformatics, and environmental science (Krishna et al., 2018; Sujatha et al., 2023).

LSTM is a type of temporal recurrent neural network, specifically designed to address the long-term dependency problem that exists in general Recurrent Neural Networks (RNN). By introducing cell state, it continuously enhances the network's ability to capture long-term dependencies, thereby solving the gradient vanishing problem of RNN when processing long sequence data (Zhao et al., 2024). The unique capabilities of LSTM have made it the mainstream model for processing sequence data, with widespread applications in natural language analysis, time series prediction, and speech recognition (Aliakbar et al., 2023; Hu et al., 2023; Zhang et al., 2023).

SVR is a regression analysis method based on Support Vector Machine (SVM). For different data distribution types, SVR can use various kernel functions such as linear, polynomial, and radial basis functions to find a hyperplane in the feature space to achieve regression prediction of the data, minimizing the error between the predicted values and the true values of the training samples (Chen et al., 2021; Cho et al., 2020; Jang et al., 2017; Jang et al., 2022). The sensitive recognition ability of SVR for outliers helps it effectively handle various high-dimensional data and non-linear problems, and in recent years, it has been successfully applied to diverse environmental numerical simulation work (Chen et al., 2019; Jang et al., 2022; Rana et al., 2024).

XGBoost is an optimized implementation based on the Gradient Boosting algorithm, and it is an efficient ensemble learning algorithm (Chen and Guestrin, 2016). It controls model complexity by adding weighted regularization terms according to the loss function, and has stronger recognition of overfitting situations. In addition, due to the adoption of parallel learning methods, its learning speed is faster than most gradient trees. The advantages of XGBoost, such as high accuracy, scalability, interpretability, and robustness, make it widely used in classification, regression, and ranking problems (Jang et al., 2022; Yu et al., 2023).

2.3.4. *Evaluation and interpretation of machine learning model performance*

This study uses statistical indicators such as R-squared (R^2), root mean square deviation (RMSD), mean absolute error (MAE), mean absolute percentage error (MAPE), Nash-Sutcliffe efficiency (NSE), and combined accuracy (CA) to evaluate the performance and accuracy of machine learning methods (Eray et al., 2018; Rana et al., 2019&2024). NSE is an indicator used to assess the prediction accuracy of hydrological models, with a range of [−∞-1], the closer the NSE is to 1, the more credible the model is. CA combines RMSD, MAE, and R^2, providing a general evaluation method similar to ideal point error for models, the lower the CA value indicates the better prediction and fitting effect of the model. The calculation formulas are as follows:

$$R^2 = \left(\frac{\sum_{i=1}^{n}\left(Q_i^m - \overline{Q}\right) \bullet \left(Q_i^0 - \overline{Q}\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_i^m - \overline{Q}\right)^2 \sum_{i=1}^{n}\left(Q_i^0 - \overline{Q}\right)^2}}\right)^2 \tag{5}$$

$$RMSD = \frac{\sqrt{\sum_{i=1}^{n}\left(Q_i^0 - Q_i^m\right)^2}}{n} \tag{6}$$

$$MAE = \frac{\sum_{i=1}^{n}\left|Q_i^0 - Q_i^m\right|}{n} \tag{7}$$

$$MAPE = \left(\frac{100}{n}\right)\sum_{i=1}^{n}\left|\frac{Q_i^0 - Q_i^m}{Q_i^0}\right| \tag{8}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(Q_i^0 - Q_i^m\right)^2}{\sum_{i=1}^{n}\left(Q_i^0 - \overline{Q}\right)^2} \tag{9}$$

$$CA = 0.33\left(RMSD + MAE + \left(1 - R^2\right)\right) \tag{10}$$

In the above formulas, n represents the amount of data, and Q_i^0、Q_i^m、$\overline{Q}$ represents the actual value, model estimation value, and average value of sea surface pCO$_2$, respectively.

3. Results

3.1. *Comparison of Copernicus and CODC-GOSD products with cruise survey data*

This study presents a comprehensive comparative analysis of Copernicus pCO$_2$ and CODC-GOSD pCO$_2$ against SOCAT pCO$_2$ data for the period of January 2001 to December 2020 (Fig. 2). Compared with SOCAT pCO$_2$, the R^2 value of both Copernicus pCO$_2$ and CODC-GOSD pCO$_2$ is 0.82. The RMSD values are 8.58 μatm and 8.53 μatm, respectively, which indicates a relatively close alignment. The MB values are −0.27 μatm and 0.50 μatm, respectively, which indicates a negative deviation between Copernicus pCO$_2$ and SOCAT pCO$_2$ and a positive deviation between CODC-GOSD pCO$_2$ and SOCAT pCO$_2$. However, the absolute difference between the two remains within an acceptable range (Fig. 2A and B). Further comparison reveals that, despite some information being obscured, Copernicus pCO$_2$ and CODC-GOSD pCO$_2$ maintain relatively consistent data characteristics. However, Copernicus pCO2 provides coverage across almost the entire Atlantic Ocean, whereas CODC-GOSD pCO2 has gaps in coastal areas. This finding provides a basis for the subsequent integration of the two products (Fig. 2C and D). A comparison of the average differences between Copernicus pCO$_2$, CODC-GOSD pCO$_2$, and SOCAT pCO$_2$ (Fig. 2E and F) reveals that for Copernicus pCO$_2$ and SOCAT pCO$_{2,}$ difference points with high negative values are dominant in the mid-to-high latitude sea areas of the North Atlantic. The sea surface pCO$_2$ difference points with high positive and negative values in the equatorial region exhibit complex interactions without a distinguishable pattern. Most of these points are concentrated along the 0° latitude line and extend into the southern sea areas, with a few negative high-value difference points located in the northern hemisphere. The same pattern is evident in the South Atlantic, specifically in the northeast Weddell Sea, where the number of difference points with high positive and negative values is relatively large and the points exhibit complex interactions. For CODC-GOSD pCO$_2$ and SOCAT pCO$_2$, many high-value difference points are also observed in the mid-to-high latitude sea areas of the North Atlantic. Unlike Copernicus pCO$_2$, CODC-GOSD pCO$_2$ predominantly exhibits positive high-value difference points and their frequency is considerable. The sea surface pCO$_2$ difference points with high positive and negative values in the equatorial sea area exhibit complex interactions and are predominantly located along the 0° latitude line and in the southern sea areas. A similar pattern to that observed in Copernicus pCO$_2$ also appears in the South Atlantic, where the difference points with high positive and negative values exhibit complex interactions. However, the density of these points is lower than that of Copernicus pCO$_2$, which ultimately indicates that CODC-GOSD pCO$_2$ has been partially optimized.

These results demonstrate the presence of varying degrees of error between Copernicus pCO$_2$ and CODC-GOSD pCO$_2$ compared with SOCAT pCO$_2$. However, the sea areas with large errors are consistently the same. This finding highlights the respective advantages of the two products. Notably, the differences between Copernicus pCO$_2$, CODC-GOSD pCO$_2$, and SOCAT pCO$_2$ are influenced by location and environmental factors. Therefore, high-quality surface pCO$_2$ data for the Atlantic Ocean can be generated by synergistically combining these two products.

To further ensure the accuracy of the machine learning model, two sea surface pCO2 products with large errors compared with SOCAT are discussed separately after being divided into different sea areas (Fig. 3I, II, and III). Consequently, three specific sea areas with anomalous values are identified as follows: the northeast sea of Canada, the eastern sea of Brazil, and the northeast of Weddell Sea.

The data sample size in the northeast sea of Canada is 386 (Fig. 3A

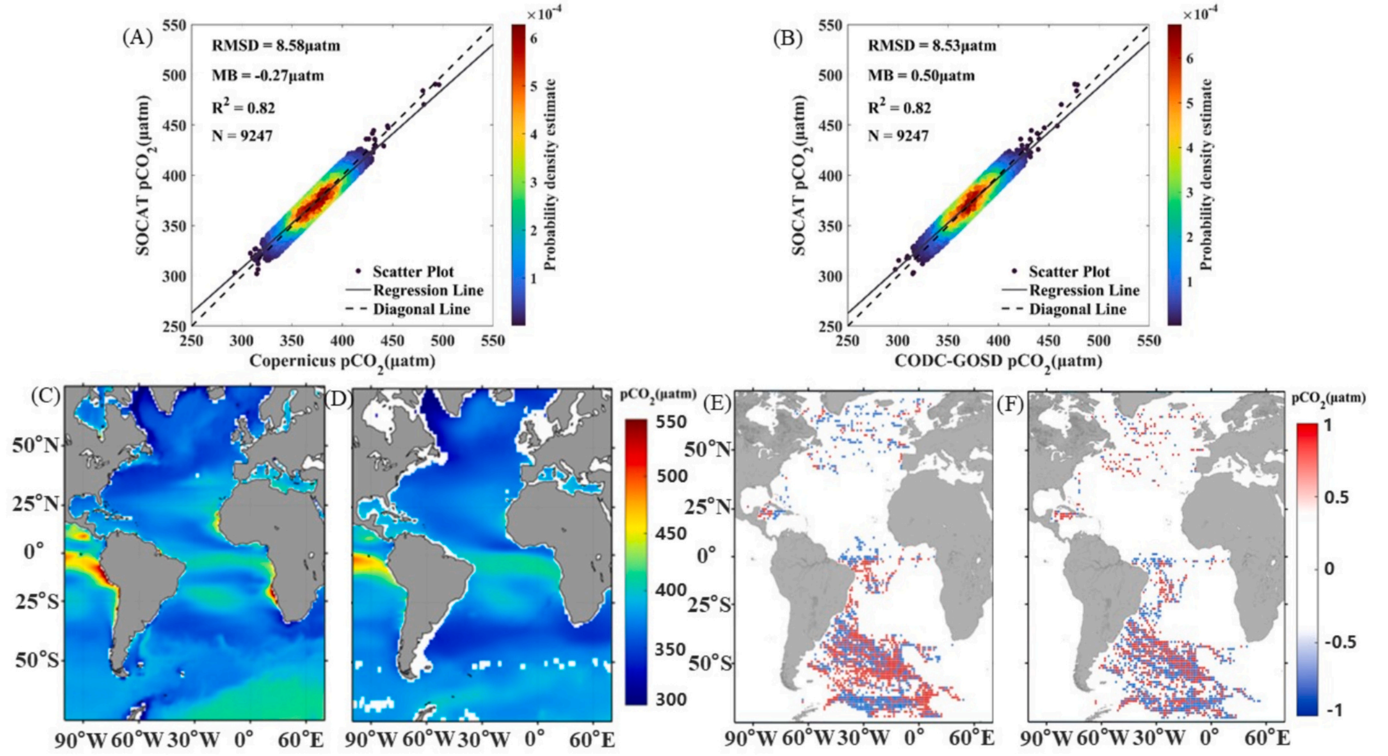

Fig. 2. Comparative analysis of Copernicus pCO_2 and CODC-GOSD pCO_2 data products against SOCAT cruise survey dataset. (A) and (B) Correlation analysis results of Copernicus pCO_2 and CODC-GOSD pCO_2 data against SOCAT data, respectively. As the colour of the scatter plot transitions from blue to red, data density increases. (C) and (D) Annual mean distribution of Copernicus pCO_2 and CODC-GOSD pCO_2 from 2001 to 2020, respectively. (E) and (F) Spatial distribution differences between Copernicus pCO_2 and CODC-GOSD pCO_2 data and SOCAT pCO_2 (pCO_2 product - cruise survey pCO_2), respectively. Data, ranging from [−1–1], are standardised to visualise the differences. The northeast sea of Canada, the eastern sea of Brazil, and the northeast of the Weddell Sea exhibit high uncertainties. Dataset quality is characterised using R^2, RMSD, and MB (mean bias, which measures the deviation between modelled and actual values). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and B) and the Copernicus pCO_2 and CODC-GOSD pCO_2 datasets are matched for correlation analysis, respectively. The R^2 values are both 0.81 and the RMSD values are 8.65 μatm and 8.57 μatm, respectively. The error in CODC-GOSD pCO_2 is lower than that in Copernicus pCO_2, although both are higher than the errors observed for the entire Atlantic Ocean. The MB values are 1.35 μatm and − 0.17 μatm, respectively. Compared with the entire Atlantic Ocean, the absolute deviation of Copernicus pCO_2 is larger but in the opposite direction, while the absolute deviation of CODC-GOSD pCO_2 is smaller, also in the opposite direction. The data sample size in the eastern sea of Brazil is 269 (Fig. 3C and D) and the R^2 values after the comparative analysis are 0.73 and 0.74, respectively. Although a significant decrease in R^2 value is observed compared with that of the entire Atlantic Ocean (0.82), the value is still very close to and higher than 0.7, which ultimately indicates that the fitting effect meets the model requirements. These anomalous results are attributed to data loss in the South Atlantic Ocean, which poses challenges in establishing high-precision data products based on existing data sources. The RMSD values are 8.68 μatm and 8.97 μatm, with the error in Copernicus pCO_2 being lower than that in CODC-GOSD pCO_2, although both are higher than that of the entire Atlantic Ocean. The MB values are 0.04 μatm and − 0.66 μatm, respectively. Compared with the entire Atlantic Ocean, the absolute deviation of Copernicus pCO_2 in the eastern sea of Brazil is smaller but in the opposite direction, while the absolute deviation of CODC-GOSD pCO_2 is larger, also in the opposite direction. The data sample size in the northeast of Weddell Sea is 3262 (Fig. 3E and F) and the R^2 values after comparative analysis are both 0.74, which is almost the same as those of the eastern sea of Brazil. The RMSD values are 8.53 μatm and 8.63 μatm, with the error in Copernicus pCO_2 being lower than that in CODC-GOSD pCO_2 and the entire Atlantic Ocean. The MB values are −0.74 μatm and 0.94 μatm, respectively. Compared with the entire Atlantic Ocean, the absolute deviations for both products are larger, but in the same direction, which is significantly different from the situation in the northeast sea of Canada and the eastern sea of Brazil.

The analysis of the three selected sea areas with relatively notable anomalies reveals that although R^2 is lower than that of the entire Atlantic Ocean, particularly in the South Atlantic, the R^2 value is still high enough. This validation demonstrates that, whether for the entire Atlantic Ocean or specific sea areas, the quality of both Copernicus pCO_2 and CODC-GOSD pCO_2 products is high and their levels tend to be consistent, without significant differences. Therefore, these products can be used as input variables for machine learning reconstruction of sea surface pCO_2.

3.2. Model parameter selection based on geographical detector

The Copernicus pCO_2, CODC-GOSD pCO_2, SHWW, U_{10}, longitude, latitude, TP, and E are input into the geographical detector to examine their respective contributions to sea surface pCO_2. The explanatory power of both the Copernicus pCO_2 and CODC-GOSD pCO_2 data products for sea surface pCO_2 is ~0.72, which is relatively high and indicates that they will play a crucial role in the machine learning model (Fig. 4). In addition, the order of explanatory power from high to low is as follows: TP, latitude, longitude, SHWW, U_{10}, and E. Notably, E performs the worst, with a Q value of only 0.003, which indicates that its association with sea surface pCO_2 is very low. To ensure the accuracy of the model while maintaining high efficiency, parameters with extremely low contributions are minimised. Therefore, E is excluded from the machine learning model, which in turn helps mitigate the problem of overfitting.

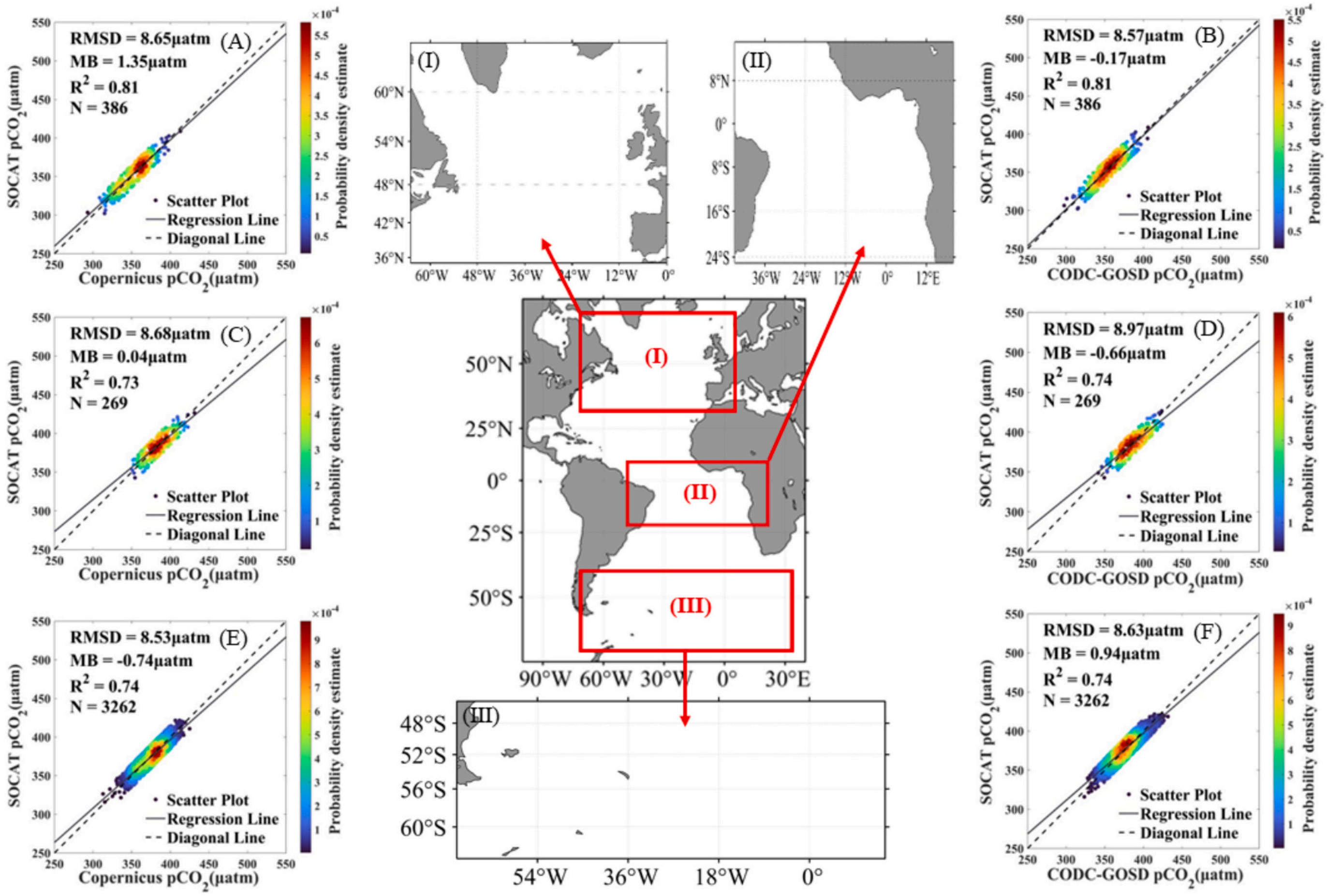

Fig. 3. Copernicus pCO$_2$ and CODC-GOSD pCO$_2$ exhibit notable errors compared with SOCAT pCO$_2$. (I), (II), and (III) Three selected anomalous sea areas: (I) northeastern sea area of Canada (latitude: 35°N-65°N, longitude: 65°W-0°); (II) eastern sea area of Brazil (latitude: 25°S-15°N, longitude: 45°W-20°E); and (III) northeastern part of Weddell Sea (latitude: 63°S-45°S, longitude: 70°W-15°E). (A), (C), and (E) Correlation analysis results of Copernicus pCO$_2$ and SOCAT pCO$_2$ for northeastern Canada, eastern Brazil, and northeastern Weddell sea areas, respectively. (B), (D), and (F) Correlation analysis results of CODC-GOSD pCO$_2$ and SOCAT pCO$_2$ for the same sea areas. Dataset quality is characterised using R^2, RMSD, and MB.

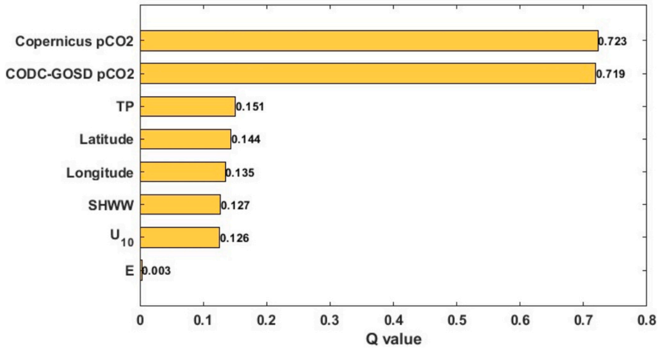

Fig. 4. Importance of each input variable is quantitatively described by the geographical detector. Q value for each variable is obtained using the factor detector to screen the variables that generate high-quality Atlantic surface pCO$_2$. Q values range from 0 to 1, with higher values indicating greater importance for pCO$_2$. Variables in the figure are arranged from top to bottom in order of importance, from high to low.

3.3. *Construction of high-quality Atlantic surface pCO$_2$ model*

All machine learning models developed in this study utilise a singular Copernicus pCO$_2$, CODC-GOSD pCO$_2$, and various environmental variables for preliminary testing. This guarantees model usability before integrating the two datasets to further enhance accuracy. Tables 3-1 and 3-2 present the performance outcomes of each model under different conditions.

From the perspective of various models, the R^2 and NSE values of the CNN are the lowest across all models, while its RMSD, MAE, MAPE, and CA values are the highest. This indicates that CNN performance in reconstructing sea surface pCO$_2$ is relatively suboptimal. In contrast, under the same conditions, the XGBoost model achieves the highest R^2 and NSE values among all models, with both its training set and validation set exceeding 0.85 and 0.87, respectively. Meanwhile, its RMSD, MAE, MAPE, and CA metrics are the lowest. This result indicates a substantially superior performance by the XGBoost model in reconstructing sea surface pCO$_2$ compared with its counterparts. Evaluating different variable combinations reveals that the differences in accuracy and error among M1, M2, and M3 within each model are subtle and lack a distinguishable progression. This finding indicates that the reconstruction quality of sea surface pCO$_2$ remains largely consistent when any two variables are paired together. Notably, M4 demonstrates the most robust performance among all models. M4 significantly enhances accuracy and reduces error compared with M1, M2, and M3. This finding indicates that the collaborative effect of three variables substantially improves sea surface pCO$_2$ reconstruction outcomes. In every instance, the combination of XGBoost with Copernicus pCO$_2$ and CODC-GOSD pCO$_2$ under the M4 framework yields R^2 and NSE values >0.90 and 0.94 for the training and validation sets, respectively. Meanwhile, the RMSD, MAE, MAPE, and CA values are 4.28 μatm/6.02 μatm (T/V),

Table 3-1
The construction results of the high-quality Atlantic surface pCO_2 model based on machine learning. The model includes CNN, LSTM, ELM, BP, SVR, and XGBoost, T is the training set, V is the validation set. Numerical thickening represents the optimal performance of each model and the corresponding parameter combination. The optimal results based on the XGBoost model are highlighted in italics. All models use the same data set (Table 1) and are optimized in the test.

Approach	Variable	Copernicus			CODC-GOSD			Copernicus and CODC-GOSD		
		R^2	RMSD(μatm)	MAE(μatm)	R^2	RMSD(μatm)	MAE(μatm)	R^2	RMSD(μatm)	MAE(μatm)
		T/V	T/V	T/V	T/V	T/V	T/V	T/V	T/V	T/V
CNN	M1	0.73/0.74	10.06/10.27	7.77/7.73	0.64/0.62	11.83/11.68	8.40/8.39	0.75/0.72	9.90/9.99	6.67/6.53
	M2	0.70/0.68	10.77/10.83	7.93/7.79	0.67/0.67	11.42/11.02	8.06/8.18	0.78/0.77	9.23/9.35	6.25/6.32
	M3	0.75/0.74	9.69/9.61	8.36/8.58	0.63/0.62	11.99/11.97	8.94/8.80	0.66/0.69	11.41/11.31	6.57/6.54
	M4	0.76/0.76	9.55/9.60	7.65/7.72	0.75/0.75	9.81/9.98	7.91/8.12	0.82/0.83	8.10/8.20	6.05/6.12
LSTM	M1	0.82/0.81	8.37/8.43	7.33/7.25	0.81/0.80	8.56/8.61	7.19/7.27	0.85/0.86	7.21/7.01	5.26/5.31
	M2	0.81/0.80	8.48/8.45	7.28/7.28	0.81/0.80	8.27/8.38	7.15/7.26	0.88/0.88	6.54/6.50	5.59/5.48
	M3	0.81/0.80	8.46/8.62	7.06/7.20	0.81/0.81	8.37/8.41	7.34/7.36	0.88/0.87	6.65/6.81	5.26/5.36
	M4	0.82/0.82	8.35/8.40	7.11/7.13	0.82/0.81	8.20/8.19	7.14/7.26	0.89/0.89	6.42/6.37	5.23/5.28
ELM	M1	0.83/0.82	7.95/8.19	6.76/6.96	0.83/0.82	7.92/7.93	6.69/6.83	0.88/0.88	6.25/6.34	5.12/5.23
	M2	0.83/0.83	7.98/8.09	6.72/6.82	0.83/0.83	7.94/7.96	6.70/6.76	0.88/0.89	6.20/6.39	5.12/5.19
	M3	0.83/0.82	7.95/8.05	6.74/6.82	0.83/0.83	7.88/8.03	6.67/6.82	0.89/0.89	6.23/6.33	5.11/5.21
	M4	0.83/0.84	7.82/7.98	6.72/6.81	0.84/0.83	7.70/7.91	6.71/6.70	0.89/0.89	6.12/6.27	5.11/5.16
BP	M1	0.84/0.84	7.74/7.89	6.55/6.69	0.84/0.84	7.81/7.88	6.61/6.75	0.89/0.89	6.16/6.23	5.08/5.24
	M2	0.84/0.84	7.82/7.92	6.55/6.75	0.84/0.83	7.83/7.93	6.60/6.79	0.88/0.89	6.25/6.19	5.01/5.29
	M3	0.84/0.83	7.74/7.79	6.76/6.82	0.83/0.83	7.87/7.88	6.45/6.73	0.88/0.87	6.21/6.39	5.09/5.20
	M4	0.85/0.84	7.63/7.69	6.63/6.62	0.84/0.84	7.78/7.81	6.46/6.57	0.89/0.89	6.12/6.18	5.02/5.16
SVR	M1	0.83/0.84	7.91/8.04	6.55/6.74	0.84/0.83	7.85/7.95	6.57/6.63	0.87/0.89	6.16/6.24	5.02/5.13
	M2	0.83/0.84	7.91/7.96	6.53/6.76	0.84/0.82	7.85/8.03	6.55/6.69	0.88/0.89	6.11/6.41	5.03/5.12
	M3	0.84/0.83	7.91/7.91	6.57/6.61	0.84/0.83	7.87/7.96	6.56/6.61	0.89/0.89	6.16/6.18	5.01/5.14
	M4	0.84/0.84	7.90/7.87	6.54/6.54	0.84/0.85	7.83/7.68	6.51/6.53	0.90/0.89	6.10/6.13	5.00/5.08
XGBoost	M1	0.90/0.84	5.58/7.42	4.40/6.29	0.90/0.85	5.62/7.61	4.49/6.22	0.90/0.89	4.39/6.11	3.46/4.99
	M2	0.90/0.84	5.53/7.60	4.36/6.28	0.90/0.84	5.57/7.71	4.45/6.20	0.89/0.91	4.40/6.13	3.40/4.95
	M3	0.90/0.83	5.56/7.56	4.48/6.47	0.90/0.84	5.53/7.57	4.50/6.22	0.90/0.88	4.44/6.06	3.49/5.06
	M4	*0.91/0.85*	*5.52/7.28*	*4.31/6.20*	*0.90/0.85*	*5.38/7.52*	*4.40/6.16*	*0.95/0.90*	*4.28/6.02*	*3.36/4.87*

M1 consists of TP and U_{10} combinations; M2 consists of TP and SHWW combinations; M3 consists of U_{10} and SHWW combinations; M4 consists of TP, U_{10}, and SHWW combinations.

Table 3-2
The construction results of the high-quality Atlantic surface pCO_2 model based on machine learning. The model includes CNN, LSTM, ELM, BP, SVR, and XGBoost, T is the training set, V is the validation set. Numerical thickening represents the optimal performance of each model and the corresponding parameter combination. The optimal results based on the XGBoost model are highlighted in italics. All models use the same data set (Table 1) and are optimized in the test.

Approach	Variable	Copernicus			CODC-GOSD			Copernicus and CODC-GOSD		
		MAPE	CA	NSE	MAPE	CA	NSE	MAPE	CA	NSE
		T/V	T/V	T/V	T/V	T/V	T/V	T/V	T/V	T/V
CNN	M1	0.02/0.02	6.04/6.09	0.75/0.77	0.02/0.02	6.87/6.83	0.74/0.73	0.02/0.02	5.61/5.61	0.82/0.83
	M2	0.02/0.02	6.34/6.32	0.75/0.76	0.02/0.02	6.61/6.52	0.78/0.76	0.02/0.02	5.24/5.31	0.84/0.85
	M3	0.02/0.02	6.10/6.15	0.72/0.72	0.02/0.02	7.11/7.06	0.70/0.70	0.02/0.02	6.12/6.06	0.84/0.83
	M4	0.02/0.02	5.82/5.86	0.77/0.78	0.02/0.02	5.99/6.12	0.76/0.76	0.02/0.02	4.78/4.83	0.85/0.86
LSTM	M1	0.02/0.02	5.29/5.29	0.79/0.80	0.02/0.02	5.31/5.36	0.81/0.81	0.01/0.01	4.21/4.15	0.90/0.89
	M2	0.02/0.02	5.31/5.31	0.81/0.81	0.02/0.02	5.20/5.28	0.80/0.81	0.01/0.01	4.08/4.03	0.89/0.88
	M3	0.02/0.02	5.22/5.34	0.81/0.82	0.02/0.02	5.30/5.32	0.79/0.80	0.01/0.01	4.01/4.10	0.88/0.89
	M4	0.02/0.02	5.21/5.23	0.81/0.82	0.02/0.02	5.17/5.21	0.81/0.81	0.01/0.01	3.91/3.91	0.90/0.89
ELM	M1	0.02/0.02	4.96/5.11	0.82/0.83	0.02/0.02	4.92/4.98	0.84/0.84	0.01/0.01	3.83/3.89	0.89/0.90
	M2	0.02/0.02	4.95/5.02	0.83/0.84	0.02/0.02	4.93/4.96	0.83/0.84	0.01/0.01	3.81/3.89	0.89/0.90
	M3	0.02/0.02	4.95/5.02	0.83/0.83	0.02/0.02	4.90/5.00	0.83/0.84	0.01/0.01	3.81/3.88	0.89/0.90
	M4	0.02/0.02	4.90/4.98	0.83/0.84	0.02/0.02	4.85/4.92	0.83/0.84	0.01/0.01	3.78/3.84	0.90/0.90
BP	M1	0.02/0.02	4.81/4.91	0.84/0.84	0.02/0.02	4.86/4.93	0.83/0.84	0.01/0.01	3.78/3.85	0.89/0.90
	M2	0.02/0.02	4.84/4.94	0.83/0.84	0.02/0.02	4.86/4.96	0.83/0.84	0.01/0.01	3.79/3.86	0.90/0.90
	M3	0.02/0.02	4.88/4.92	0.83/0.83	0.02/0.02	4.83/4.92	0.84/0.85	0.01/0.01	3.80/3.91	0.90/0.90
	M4	0.02/0.02	4.80/4.82	0.84/0.84	0.02/0.02	4.80/4.84	0.84/0.85	0.01/0.01	3.75/3.81	0.90/0.90
SVR	M1	0.02/0.02	4.87/4.98	0.83/0.84	0.02/0.02	4.86/4.91	0.83/0.84	0.01/0.01	3.77/3.82	0.89/0.90
	M2	0.02/0.02	4.87/4.96	0.84/0.84	0.02/0.02	4.85/4.97	0.83/0.84	0.01/0.01	3.75/3.87	0.89/0.90
	M3	0.02/0.02	4.88/4.89	0.84/0.84	0.02/0.02	4.86/4.91	0.83/0.84	0.01/0.01	3.76/3.80	0.90/0.90
	M4	0.02/0.02	4.86/4.85	0.84/0.84	0.02/0.02	4.83/4.78	0.85/0.84	0.01/0.01	3.93/3.77	0.90/0.90
XGBoost	M1	0.01/0.02	3.36/4.62	0.88/0.91	0.01/0.02	3.40/4.66	0.89/0.91	0.01/0.01	2.65/3.73	0.93/0.94
	M2	0.01/0.02	3.33/4.68	0.87/0.91	0.01/0.02	3.37/4.69	0.86/0.91	0.01/0.01	2.64/3.71	0.93/0.94
	M3	0.01/0.02	3.38/4.73	0.87/0.91	0.01/0.02	3.37/4.65	0.87/0.91	0.01/0.01	2.68/3.74	0.94/0.94
	M4	*0.01/0.02*	*3.30/4.54*	*0.88/0.91*	*0.01/0.02*	*3.29/4.61*	*0.87/0.91*	*0.01/0.01*	*2.55/3.66*	*0.94/0.94*

M1 consists of TP and U_{10} combinations; M2 consists of TP and SHWW combinations; M3 consists of U_{10} and SHWW combinations; M4 consists of TP, U_{10}, and SHWW combinations.

3.36 μatm/4.84 μatm (T/V), 0.01/0.01 (T/V), and 2.55/3.66 (T/V), respectively, which ultimately establishes the M4 framework as the premier reconstruction scenario. These findings demonstrate that the synergy of the XGBoost model with U_{10}, SHWW, and TP maximises the potential of data from either the Copernicus or CODC-GOSD products, particularly when combined, which in turn positions this method as an optimal approach to sea surface pCO_2 reconstruction.

Upon completion of the model self-assessment, a comparison of accuracy between the sea surface pCO_2 reconstructed via machine learning and the SOCAT data (Table 4) reveals that the XGBoost-based pCO_2, reconstructed using the optimal XGBoost model (Table 3), exhibits the highest correlation with SOCAT, achieving an R^2 value of 0.94. This finding further confirms that the XGBoost model can effectively leverage the intricate relationships between various environmental variables, which ultimately demonstrates the significant potential of the model in sea surface pCO_2 reconstruction.

3.4. Evaluation of stability of XGBoost-based pCO_2 model

Moussa et al. (2015) highlighted that lack of data is one of the factors affecting the accuracy of North Atlantic sea surface pCO_2 reconstruction using neural networks. However, the high variability of environmental factors and their impact on model accuracy cannot be ignored. The marine environment is complex and constantly changing, with large instantaneous differences in many variables, which poses a considerable challenge for model computation. Comprehensive verification (Section 4.1) ensures the applicability of the XGBoost model and testing its stability for application in the variable marine environment is essential. To simulate sudden changes in the marine environment, the environmental parameters U_{10}, SHWW, and TP are each altered by ±20 % and the machine learning model hyperparameters remain unchanged. A difference analysis is conducted based on the set parameters. Fig. 5 shows the result of this analysis.

With the addition of 20 % uncertainty, the R^2 value of U_{10} is 0.94, the MB is 0.00 μatm, and the RMSD is 4.69 μatm. A comparison of the correlation analysis results (Table 4) indicates that the changes in R^2 and MB can be ignored and the RMSD decreases by 0.02 μatm. With the removal of the 20 % uncertainty, the R^2 value of U_{10} remains 0.94, the MB stays at 0.00 μatm, and the RMSD is 4.71 μatm. A comparison of the correlation analysis results (Table 4) indicates that the changes in R^2, RMSD, and MB can be ignored. With the addition of 20 % uncertainty, the R^2 value of SHWW is 0.94, the MB is 0.00 μatm, and the RMSD is 4.73 μatm. A comparison of the correlation analysis results (Table 4) indicates that the changes in R^2 and MB can be ignored and the RMSD increases by 0.02 μatm. With the removal of the 20 % uncertainty, the R^2 value of SHWW remains 0.94, the MB stays at 0.00 μatm, and the RMSD is 4.69 μatm. A comparison of the correlation analysis results (Table 4) indicates that the changes in R^2 and MB can be ignored and the RMSD decreases by 0.02 μatm. With the addition of 20 % uncertainty, the R^2 value of TP is 0.94, the MB is 0.00 μatm, and the RMSD is 4.69 μatm. A comparison of the correlation analysis results (Table 4) indicates that the changes in R^2 and MB can be ignored and the RMSD decreases by 0.02 μatm. With the removal of the 20 % uncertainty, the R^2 value of TP remains 0.94, the MB stays at 0.00 μatm, and the RMSD is 4.72 μatm. A comparison of the correlation analysis results (Table 4) indicates that the changes in R^2 and MB can be ignored and the RMSD increases by 0.01 μatm. Overall, with the addition of 20 % uncertainty, the RMSD of U_{10} and TP both decrease slightly, while the RMSD of SHWW increases slightly. With the removal of the 20 % uncertainty, the RMSD of U_{10} remains largely unchanged, the RMSD of TP increases slightly, and the RMSD of SHWW decreases slightly. Notably, U_{10} and TP maintain a highly consistent error trend, while SHWW does the opposite.

Given the influences of sea breeze, wave action, and precipitation—each of which can alter the characteristics of the sea–air interface to a certain extent and exhibit a strong instantaneous rate of change—considering their variability when reconstructing the model is crucial (Bates and Merlivat, 2001; Jacob et al., 2019; Turk et al., 2013). The variability error of all three variables is controlled within 1 % based on the verification above. This indicates that the XGBoost model is not highly sensitive to the uncertainty of each input environmental variable and the model tolerance for U_{10} and TP is slightly higher than that for SHWW. This finding demonstrates that the model has a certain capacity to handle abrupt environmental changes. In addition, considering the diverse data sources input into the model, some of which have undergone multiple simulation interpolations and inherently carry certain uncertainties, these uncertainties are somewhat mitigated when the high-tolerance XGBoost model is applied to similar reanalysis data products, not significantly affecting the reconstruction results (Chen et al., 2019). Thus, the XGBoost model demonstrates excellent performance in the reconstruction of Atlantic sea surface pCO_2. The broad applicability and robustness of the model help achieve ideal results in subsequent similar reconstruction tasks.

4. Discussion

4.1. Importance of incorporating environmental variables within machine learning

In this study, we selected four environmental variables—U_{10}, TP, E, and SHWW—along with Copernicus pCO_2 and CODC-GOSD pCO_2, to reconstruct the sea surface pCO_2 of the Atlantic Ocean. The selection of these variables is based on published research results and their findings. Conventional studies often use SST and SSS for regional sea surface pCO_2 reconstruction because these variables have significant advantages in capturing the thermodynamic effects of the ocean. Friedrich and Oschlies (2009) and Telszewski et al. (2009) both utilized SST and SSS to reconstruct the sea surface pCO_2 of the North Atlantic. These studies demonstrated that the variables are both important and effective in sea surface pCO_2 reconstruction. Although the SOM algorithm used at that time was not perfect, it demonstrated that machine learning algorithms are far superior to traditional regression methods in extracting the essence of environmental data. To more comprehensively account for the impact of complex environmental factors, we used Copernicus pCO_2 and CODC-GOSD pCO_2 products. These two data products contain parameters such as SST, SSS, and Chl, which are closely related to sea surface pCO_2 and are integral to model reconstruction. Thus, considering the impact of the atmosphere and waves, we introduced environmental variables that are seldom discussed in mainstream research to

Table 4
Construction results of the high-quality Atlantic surface pCO_2 model based on machine learning compared with the accuracy of SOCAT cruise survey data. Correlation analysis for Copernicus pCO_2, CODC-GOSD pCO_2, and Copernicus and CODC-GOSD pCO_2 against SOCAT pCO_2. Input parameters in all three cases include SHWW, U_{10}, TP, longitude, and latitude. Dataset quality is characterised using R^2, RMSD, and MB. Optimal results based on the XGBoost model are highlighted in italics to identify them as the best-performing model.

Approach	Copernicus pCO_2			CODC-GOSD pCO_2			Copernicus and CODC-GOSD pCO_2		
	R^2	RMSD(μatm)	MB(μatm)	R^2	RMSD(μatm)	MB(μatm)	R^2	RMSD(μatm)	MB(μatm)
CNN	0.82	9.56	−4.81	0.83	9.84	−5.51	0.90	8.12	5.05
LSTM	0.82	8.38	−0.57	0.83	8.28	−1.06	0.90	6.43	−0.83
ELM	0.83	8.01	−0.07	0.84	7.90	0.02	0.90	6.24	0.00
BP	0.85	7.69	0.04	0.84	7.79	−0.12	0.90	6.13	0.01
SVR	0.84	7.90	−0.04	0.84	7.80	0.08	0.90	6.14	0.06
XGBoost	*0.91*	*5.92*	*−0.03*	*0.91*	*5.87*	*0.03*	*0.94*	*4.71*	*0.00*

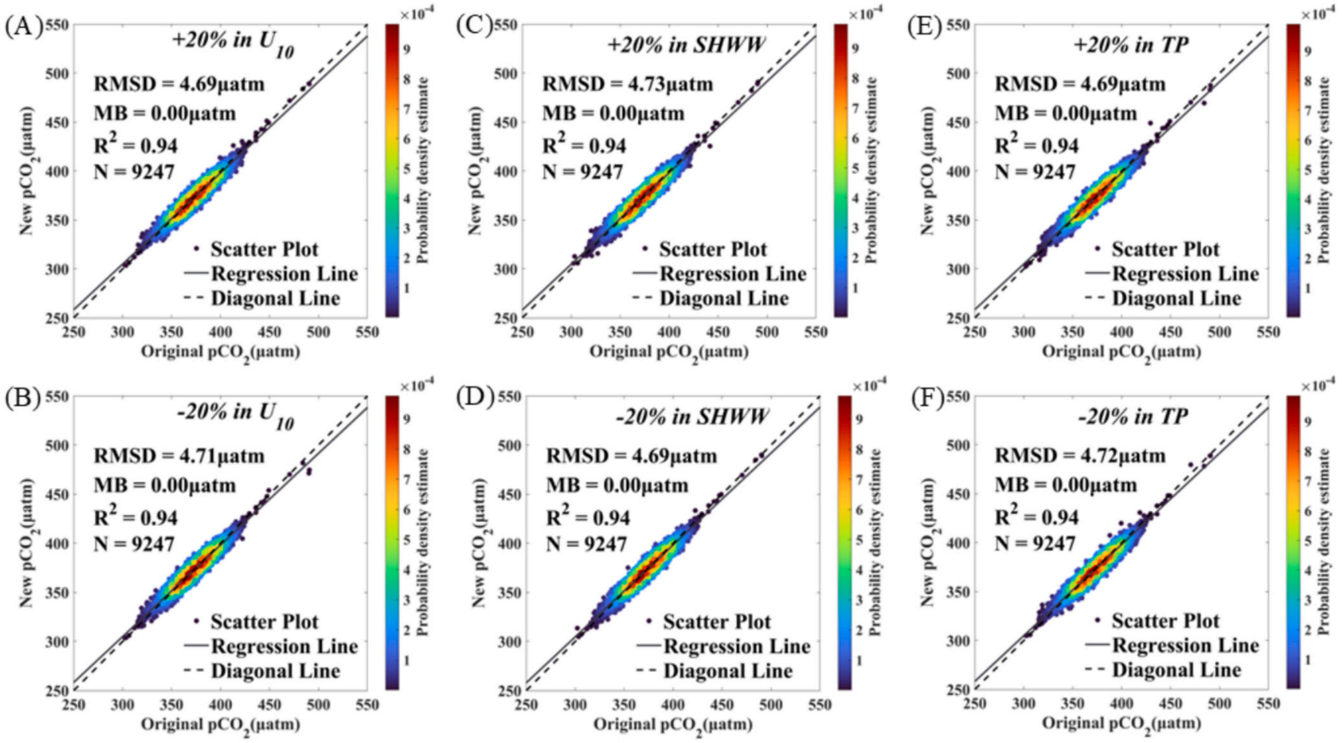

Fig. 5. Comparison between new Atlantic sea surface pCO_2 dataset, established based on original data, and original reconstructed data (Table 1). U_{10}, SHWW, and TP inputs to the XGBoost model are altered by ±20 % to simulate changes in the marine environment. As the colour of the scatter plot in the correlation analysis changes from blue to red, the data density increases progressively. X-axis represents original XGBoost-based pCO_2 data and Y-axis represents XGBoost-based pCO_2 data generated by re-learning after the change. (A) and (B) Comparison results for ±20 % changes in U_{10}. (C) and (D) Comparison results for ±20 % changes in SHWW. (E) and (F) Comparison results for ±20 % changes in TP. R^2, RMSD, and MB are used to characterise the quality of each dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

aid in the reconstruction of the Atlantic sea surface pCO_2. No clear functional relationship is present between these variables and sea surface pCO_2, so various machine learning algorithms are used to simulate the unknown relationships. After extensive data training and thorough comparison, the XGBoost-based pCO_2 model performed the best, with an overall estimation accuracy of 94 % across a wide dynamic range and low uncertainty (RMSD <5 μatm).

The model stability assessment results (Section 3.4) indicate that the XGBoost model is not highly sensitive to the uncertainty of each input variable, with all uncertainties controlled within 1 %. The model exhibits relatively higher sensitivity to SHWW and TP and lower sensitivity to U_{10}. Overall, the model is considered a relatively ideal model. Although machine learning relies on pure logical operations performed on the data, the processes controlling sea surface pCO_2—such as thermodynamics, biochemistry, ocean circulation, and sea–air exchange—are represented through the inputs of Copernicus pCO_2, CODC-GOSD pCO_2, TP, SHWW, and U_{10}. Therefore, the simulation results are credible (Turk et al., 2013; Fay and McKinley, 2017; Dixit et al., 2019; Zhong et al., 2021). However, two issues require discussion: 1) Why retain variables that the model identifies as having low sensitivity? 2) Could the insensitivity of the model to variable uncertainty result in its inability to accurately capture the characteristic information of sea surface pCO_2?

For the first question, although the sensitivity of the XGBoost model to U_{10}, SHWW, and TP is not high, with that of U_{10} being the lowest, this does not directly imply that U_{10} is unimportant for sea surface pCO_2 reconstruction. First, according to the results of the geographical detector (Fig. 4), the Q values for U_{10}, SHWW, and TP are 0.126, 0.127, and 0.151, respectively. From a quantitative perspective, the influence of U_{10}, SHWW, and TP on sea surface pCO_2 is not significantly different, particularly between U_{10} and SHWW. This is because wind, waves, and precipitation typically occur simultaneously and collectively affect the ocean surface by influencing gas exchange and water mixing. Strong winds and large waves can significantly increase the gas exchange rate, which indirectly promotes precipitation, while precipitation can also dilute surface seawater. The combination of these factors may have either additive or offsetting effects on sea surface pCO_2. Given the complexity of the mechanisms behind this cyclic pattern on sea surface pCO_2, the XGBoost model must include all variables to ensure accuracy. Second, both U_{10} and TP exhibit a significant negative correlation with sea surface pCO_2. This finding indicates that air–sea exchange is the primary factor controlling sea surface pCO_2 (Fig. 6A and B). In contrast, SHWW exhibits a strong positive correlation with sea surface pCO_2. This indicates that waves play a significant role in promoting the production of sea surface pCO_2 (Fig. 6C). The commonality among these three variables is that although their relationships with sea surface pCO_2 vary in both positive and negative directions, they exhibit a high degree of consistency across the entire Atlantic Ocean, with anomalies only observed in some coastal bays. This may explain why the XGBoost model is insensitive to variable fluctuations. Slightly different from the Q value results (Fig. 4), TP does not show the highest absolute correlation with sea surface pCO_2 across the entire Atlantic Ocean. A possible explanation is the loss of precipitation data. While spatial interpolation helps compensate for this gap, the accumulated error ultimately affects some of the research results. This also indirectly indicates that complete data is crucial for high-precision sea surface pCO_2 reconstruction. Considering these two factors, including U_{10}, SHWW, and TP in the model for sea surface pCO_2 reconstruction is essential.

For the second question, the Atlantic sea surface pCO_2 results (Tables 3-1 and 3-2) provide some insight. Notably, the XGBoost model and the other five machine learning models have issues with precision ambiguity under the M1, M2, and M3 combinations. However, all

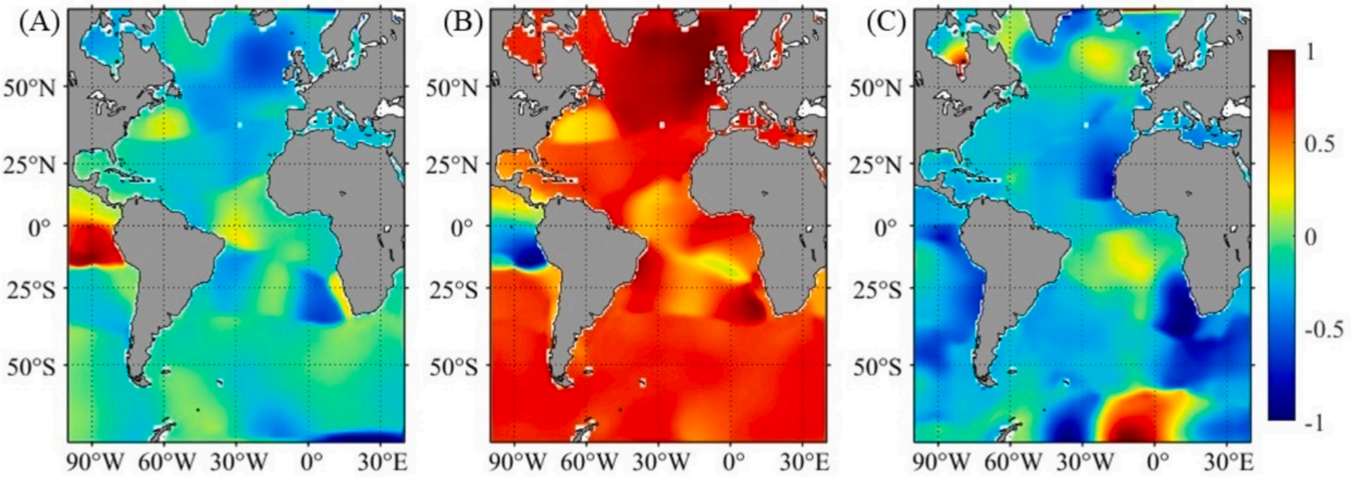

Fig. 6. Plot of correlation coefficient between environmental variables and sea surface pCO_2. (A), (B), and (C) Spatial correlation between U_{10}, SHWW, and TP, respectively, and sea surface pCO_2 in the range [−1,1].

models demonstrate significant improvement with the M4 combination. This finding indicates that inputting variables that do not characterise the various oceanic processes affecting sea surface pCO_2 from multiple perspectives into the model weakens the ability of the machine learning model to capture sea surface pCO_2 features. Although approximate results can still be obtained, the underlying computational logic remains open to discussion. However, once the complete parameters of air–sea interaction processes are incorporated into the model, the ability to capture feature information is significantly enhanced, a point corroborated by multiple machine learning models. The results demonstrate that the key to machine learning models effectively capturing sea surface pCO_2 feature information lies in whether the variables input into the model can form a logical closed-loop. For the uncertainty of variables, the condition is an obstacle to the performance of the model and an effective means of verification.

4.2. Applicability of XGBoost-based pCO_2 model in local sea areas

The results (Section 3) indicate that the XGBoost-based pCO_2 model developed for the entire Atlantic Ocean is well-suited for large-scale application. This raises an important question: Is this reconstruction method equally valuable in localized regions with anomalous values and complex environmental conditions? To assess the universal applicability of the model, we selected the northeastern sea of Canada, the eastern sea of Brazil, and the northeast of the Weddell Sea for testing the XGBoost model. The reasons for selecting these three regions with anomalous values are as follows: First, according to the results (Section 3.1), after calculating the differences between Copernicus pCO_2 and CODC-GOSD pCO_2 against SOCAT pCO_2, a concentrated distribution of anomalous points is observed in the three sea areas owing to the inherent data characteristics of Copernicus pCO_2 and CODC-GOSD pCO_2. Although the directions of the dataset deviations in the regional anomalies are not entirely consistent, the results still indicate that these three sea areas present particular challenges for model reconstruction. Therefore, conducting a detailed analysis to verify model applicability is essential. Second, all three sea areas are influenced by complex marine processes, which contrasts sharply with the more stable central ocean areas. Specifically, the northeast of Weddell Sea is an important carbon sink in the Southern Ocean and the amount of absorbed anthropogenic CO_2 has been increasing annually, reaching 4.1 mol C m^{-2} at the start of the 21st century (Mario, 2004). This has a significant impact on sea surface pCO_2, particularly in spring and summer, when human activities contribute to an increase in sea surface pCO_2 owing to rising temperatures, with an estimated increase of ~30 μatm. Meanwhile, the Weddell Sea is located near the Antarctic continent, where temperatures are relatively low, thus, the influence of sea ice on sea surface pCO_2 cannot be ignored (Margaret et al., 2020; Zemmelink et al., 2006). The northeast sea of Canada is part of the high-latitude North Atlantic, surrounded by land and numerous bays, which makes the sea area complex. Studies have demonstrated that its seasonal sea surface pCO_2 is increasing at a rate of 1.5 μatm y^{-1}. Notably, the global ocean biogeochemical model (GOBM) estimates the annual net CO_2 absorption of the Atlantic Ocean to be ~0.47 ± 0.15 Pg C yr^{-1}, while the estimate from common sea surface pCO_2 products is 0.36 ± 0.06 Pg C yr^{-1}, with the largest discrepancy occurring north of 50°N (Pérez et al., 2024). This discrepancy highlights the need to improve modeling techniques and better integrate observational data to enhance the accuracy of sea surface pCO_2 estimates in this sea area. The eastern sea of Brazil, located near the equator at the north–south divide of the Atlantic Ocean, presents a more complex marine environment. Although fewer studies have focused on sea surface pCO_2 in this area, the area remains a suitable choice for testing the applicability of the model. In summary, if the XGBoost-based pCO_2 model performs effectively in each of these three anomalous sea areas, then that provides substantial evidence of the versatility of the model.

We first evaluated the overall reconstruction performance of the model in the three outlier sea areas. The results (Table 5) reveal that the validation R^2, RMSD, and MB for the XGBoost-based pCO_2 model in the northeast sea of Canada are 0.98, 2.95 μatm, and 0.07 μatm, respectively. In the eastern sea of Brazil, the validation R^2, RMSD, and MB are 0.97, 2.81 μatm, and − 0.12 μatm, respectively. Finally, in the northeast of Weddell Sea, the validation R^2, RMSD, and MB are 0.95, 3.59 μatm, and − 0.14 μatm, respectively. Compared with the correlation analysis results (Table 4), the reconstruction performance in these three sea areas remains at a high level. Notably, the best overall R^2 and RMSD for the Atlantic in the correlation analysis results (Table 4) are 0.94 and 4.71 μatm, respectively. This finding indicates that the XGBoost-based pCO_2 model developed in this study demonstrates excellent performance in the outlier sea areas and may even surpass its overall performance across the Atlantic. To further validate the accuracy of the results in these outlier regions, this study utilizes the three most recent Atlantic voyage datasets. This approach also helps assess model adaptability over temporal scales (Chen et al., 2019; Jang et al., 2022).

Copernicus pCO_2 and CODC-GOSD pCO_2 exhibit scattered differences in the northeast sea of Canada. Copernicus pCO_2 is dominated by negative high-value difference points, while CODC-GOSD pCO_2 is dominated by positive high-value difference points. Although Copernicus pCO_2 and SOCAT pCO_2 exhibit similar trends, the overall pCO_2

Table 5
Analysis results of the correlation between Copernicus pCO_2, CODC-GOSD pCO_2, XGBoost-based pCO_2, and SOCAT pCO_2 in anomalous sea areas, characterised using R^2, RMSD, and MB to indicate dataset quality.

Region	Copernicus pCO_2			CODC-GOSD pCO_2			XGBoost-based pCO_2		
	R^2	RMSD(μatm)	MB(μatm)	R^2	RMSD(μatm)	MB(μatm)	R^2	RMSD(μatm)	MB(μatm)
Northeast sea of Canada	0.81	8.65	1.35	0.81	8.57	−0.17	0.98	2.95	0.07
Eastern sea of Brazil	0.73	8.68	0.04	0.74	8.97	−0.66	0.97	2.81	−0.12
Northeast sea of Weddell	0.74	8.53	−0.74	0.74	8.62	0.94	0.95	3.59	−0.14

values are significantly underestimated (Fig. 7C and D). In contrast, while CODC-GOSD pCO_2 maintains a similar trend to SOCAT pCO_2, it overestimates values in more than half of the points. These findings indicate good consistency from both temporal and spatial perspectives in these two scenarios. Furthermore, the XGBoost-based pCO_2 maintains a highly consistent trend with SOCAT pCO_2 and substantially reduces the overestimation of pCO_2 values, which ultimately results in better overall agreement (Fig. 7E). In the eastern sea of Brazil, both Copernicus pCO_2 and CODC-GOSD pCO_2 exhibit concentrated, complex interactions of positive and negative high-difference points. The results (Fig. 7G and H) show that both products maintain a relatively consistent trend with SOCAT pCO_2, and Copernicus pCO_2 is generally overestimated, while CODC-GOSD pCO_2 is underestimated in the first half of 2020 and overestimated in the second half. This time-varying overestimation is closely associated with the distribution of spatial difference points. In contrast, the XGBoost-based pCO_2 maintains a highly consistent trend with SOCAT pCO_2 and the average magnitude of overestimation is minimal (Fig. 7I). Similar to the eastern sea of Brazil, the northeast of Weddell Sea exhibits concentrated positive and negative high-difference points for both Copernicus pCO_2 and CODC-GOSD pCO_2, but with a larger coverage area and greater density. The comparison analysis results (Fig. 7K and L) indicate that the number of overestimation and underestimation points for Copernicus pCO_2 is comparable, with a significant increase in highly matched points. In contrast, CODC-GOSD pCO_2 is underestimated in the first half of 2020 and overestimated in the second half. The agreement between all points of XGBoost-based pCO_2 and SOCAT pCO_2 remains strong, with the average magnitude of overestimation confined to a small range (Fig. 7M).

These results effectively demonstrate that, regardless of spatial or temporal considerations, the XGBoost-based pCO_2 model has broad applicability for sea surface pCO_2 reconstruction in the Atlantic Ocean. However, whether a few navigational data points can reflect the global situation requires further exploration in future research. Although the accuracy of this theory requires substantial local research for support, the current research results indicate that the model yields accurate estimates as long as the data input encompasses sufficient temporal and spatial dimensions. Similar reconstruction studies of sea surface pCO_2 hold great potential for the global ocean.

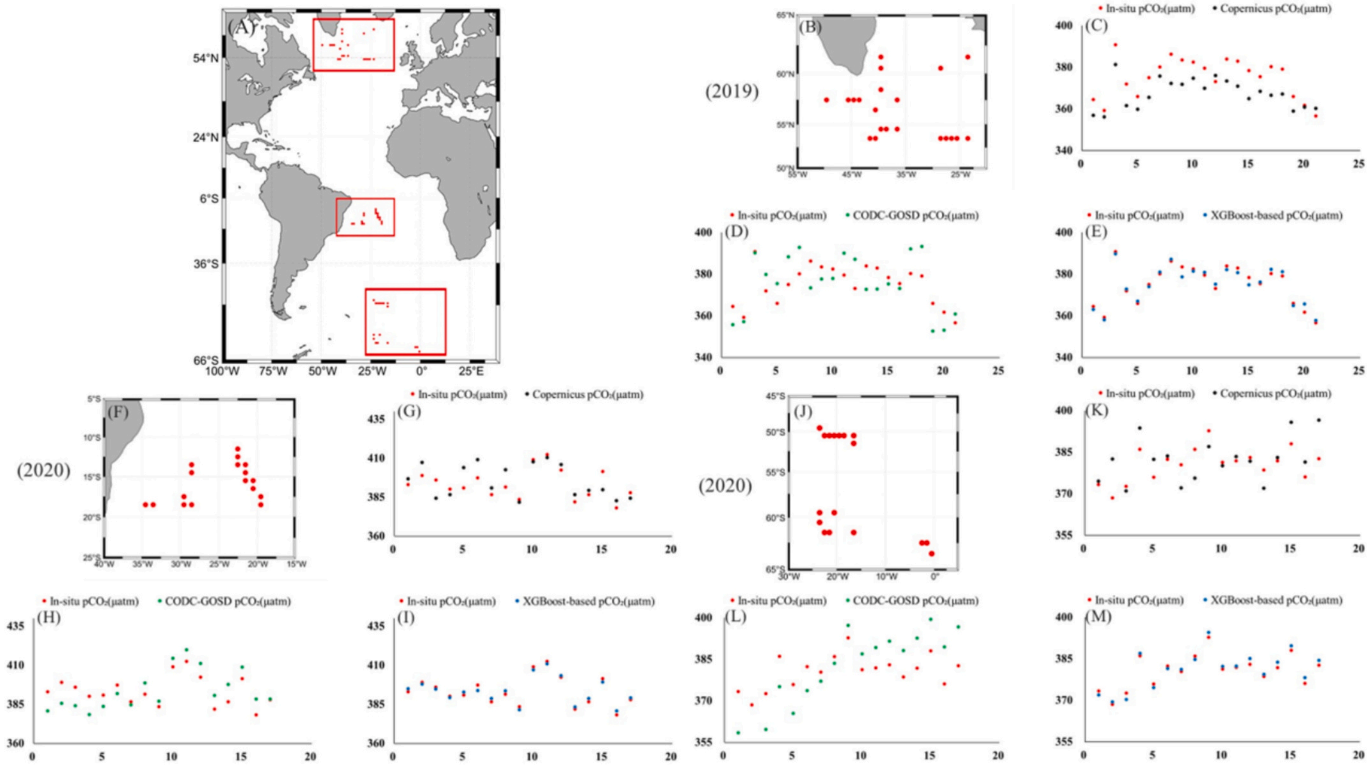

Fig. 7. Spatial and temporal distribution of the SOCAT cruise survey data is used to verify the accuracy of the anomalous sea area and to conduct a comparative analysis between the products. X-axis in the figure represents the point positions recorded according to the time sequence of the cruise survey. Accordingly, 21 points are recorded in the northeast sea of Canada, 17 points in the eastern sea of Brazil, and 17 points in the northeast of the Weddell Sea. (A) Spatial and temporal distribution of the cruise survey verification data not used for machine learning in the selected anomalous sea areas; (B), (C), (D), and (E) Comparison of SOCAT pCO_2 against Copernicus pCO_2, CODC-GOSD pCO_2, and XGBoost-based pCO_2 over time in the northeastern waters of Canada. (F), (G), (H), and (I) Comparison of the cruise survey points in the eastern sea of Brazil and SOCAT pCO_2 against Copernicus pCO_2, CODC-GOSD pCO_2, and XGBoost-based pCO_2 over time. (J), (K), (L), and (M) Comparison of the cruise survey points in the northeast of the Weddell Sea and SOCAT pCO_2 against Copernicus pCO_2, CODC-GOSD pCO_2, and XGBoost-based pCO_2 over time.

4.3. Relationship between Atlantic surface pCO_2 and global ocean acidification

Since the Industrial Revolution, the widespread use of fossil fuels has led to a rapid increase in global carbon emissions. The rising atmospheric carbon dioxide partial pressure ensures that more CO_2 is absorbed by the surface ocean. As the ocean absorbs CO_2, the acidity of seawater increases (a decrease in pH), a phenomenon known as ocean acidification (Richard et al., 2009). Research studies have demonstrated that the global ocean is currently experiencing the fastest rate of acidification in 55 million years (James et al., 2005). The current seawater pH ranges from 7.8 to 8.2 and the acidity of seawater (hydrogen ion concentration) has increased by 1–1.5 times compared with the acidity levels in 1800. Researchers have predicted that by 2100, the ocean pH will decrease by 0.3–0.4, and by 2300, the decrease could be as much as 0.7–0.8 (Siegenthaler and Sarmiento, 1993). A decrease in ocean pH will dramatically alter the chemical characteristics of seawater, which ultimately affects the physiology, growth, reproduction, and metabolism of marine organisms and threatens marine biodiversity. This will ultimately lead to irreversible changes in marine ecosystems, which disrupt their balance and their services to humans, such as a reduction in fishery resources, impeded development of the tourism industry, and decreased marine energy extraction. Therefore, ocean acidification has become the third major environmental issue that severely affects and threatens human societal development, following global change and environmental pollution (Bach et al., 2017; Sabine et al., 2004).

The enhanced XGBoost model developed in this study has been rigorously tested for its applicability and stability. The model demonstrates significant potential for Atlantic sea surface pCO_2 reconstruction. To further contextualise these findings within the marine ecological environment, the time series is extended back to 1993 by averaging the highly similar Copernicus and CODC-GOSD pCO_2 products. Furthermore, the seawater pH dataset from Copernicus Marine and the global atmospheric CO_2 concentration $2^\circ \times 2.5^\circ$ grid simulation dataset (Hou et al., 2022) are used for trend comparison, following their alignment with the research data. The results (Fig. 8) indicate that from 1993 to 2020, the curves for atmospheric CO_2 and sea surface pCO_2 exhibit a highly consistent upward trend, while the curve for sea surface pH exhibits an opposite downward trend. Over the years, the average CO_2 concentration over the Atlantic was ~382.32 ppm, increasing from 354.14 ppm in 1993 to 411.96 ppm in 2020. This represents an increase of ~57.82 ppm, with an average annual increase rate of 2.07 ppm y^{-1}. The average sea surface pCO_2 was ~342.13 μatm, increasing from 323.72 μatm in 1993 to 362.91 μatm in 2020. This represents an increase of ~39.19 μatm, with an average annual increase rate of 1.40 μatm y^{-1}. The average seawater pH was ~8.06, decreasing from 8.08 μatm in 1993 to 8.03 μatm in 2020. This represents a decrease of ~0.05, with an average annual decrease rate of 0.0018 y^{-1}. Notably, CO_2, pH, and pCO_2 did not exhibit a highly pronounced downward trend, particularly the emissions of CO_2, which continued to rise steadily. Data from the comparison of trends (Fig. 8) also reveals that for every 1 ppm increase in CO_2 emissions, pCO_2 increases by 0.8911 μatm and pH decreases by 0.0007, and this trend is intensifying.

5. Conclusions

This study fully utilized multi-source data and integrated geographic information analysis methods with various machine learning models to reconstruct sea surface pCO_2 in the Atlantic Ocean. All the machine learning models demonstrated excellent performance, with the XGBoost-based global sea surface pCO_2 model performing particularly well across different scenarios. The reconstructed XGBoost-based pCO_2 achieved an overall accuracy of 94 % in the Atlantic Ocean, with local sea areas exceeding 95 %. This model demonstrated substantially greater precision compared with the standalone Copernicus and CODC-GOSD products. The robustness and broad applicability of the model can provide more accurate information for the analysis of marine patterns in other regions and highlight the severe conditions of marine ecological environments.

Given the limitations and shortcomings of this study, future research could benefit from a more comprehensive exploration and improvement in the following areas:

1. Addressing the paucity of research areas, the reconstruction of sea surface pCO_2 in the Atlantic Ocean has been relatively successful. Future studies could consider extending this methodology to

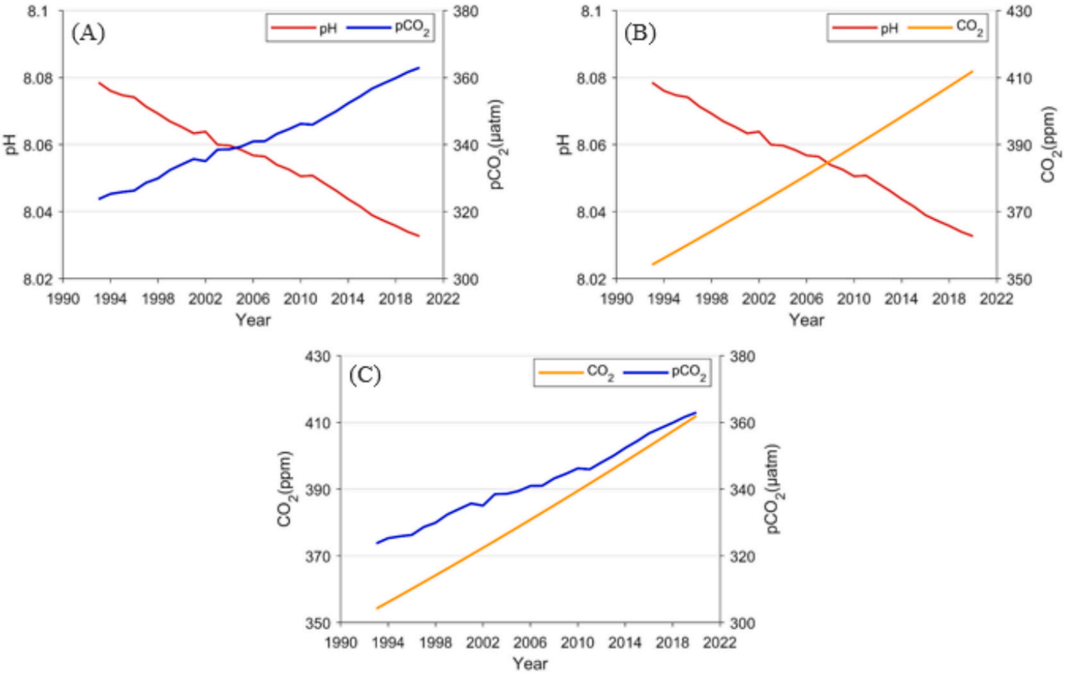

Fig. 8. Comparison of CO_2, sea surface pCO_2, and pH trends over the Atlantic. (A), (B), and (C) Trends of pH and pCO_2, pH and CO_2, and CO_2 and pCO_2 from 1992 to 2020, respectively.

reconstruct sea surface pCO_2 in other oceans. Such an expansion would facilitate the creation of continuous, large-scale data products, which ultimately provide a robust foundation for further marine scientific research.

2. The reliance on a single machine learning model presents a challenge. While this study considered multiple models, it did not combine them to evaluate their collective impact on the reconstruction. Whether such an approach would improve accuracy is an issue that requires further investigation.
3. The significant data deficiency in polar marine regions presents a challenge. However, with the anticipated generation of high-precision remote sensing data, this challenge is expected to be addressed. Meanwhile, scientific efforts from various countries will continue to gather empirical data from these polar regions. Consequently, the global ocean surface pCO_2 model is expected to undergo further refinement and improvement.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Jiaming Liu: Writing – original draft, Software, Methodology. **Jie Wang:** Writing – review & editing. **Xun Wang:** Supervision, Data curation. **Yixuan Zhou:** Supervision. **Runbin Hu:** Supervision. **Haiyang Zhang:** Supervision.

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 data we used are freely available from the following websites: (https://socat.info/) (https://marine.copernicus.eu/) (https://www.casodc.com/) (https://cds.climate.copernicus.eu/).

References

Aliakbar, M., et al., 2023. Novel integrated modelling based on multiplicative long short-term memory (mLSTM) deep learning model and ensemble multi-criteria decision making (MCDM) models for mapping flood risk. J. Environ. Manag. 345 (1), 118838. https://doi.org/10.1016/j.jenvman.2023.118838.

Bach, L.T., et al., 2017. Simulated ocean acidification reveals winners and losers in coastal phytoplankton. PLoS One 12, 188–198. https://doi.org/10.1371/journal.pone.0188198.

Bai, Y., et al., 2015. A mechanistic semi-analytical method for remotely sensing sea surface pCO_2 in river-dominated coastal oceans: a case study from the East China Sea. JGR Oceans 120, 2331–2349. https://doi.org/10.1002/2014JC010632.

Bakker, D.C.E., et al., 2016. A multi-decade record of high-quality fCO_2 data in version 3 of the Surface Ocean CO_2 atlas (SOCAT). Earth Syst. Sci. Data 8, 383–413. https://doi.org/10.5194/essd-8-383-2016.

Bates, N.R., Merlivat, L., 2001. The influence of short-term wind variability on air-sea CO_2 exchange. Geophys. Res. Lett. 28, 3281–3284. https://doi.org/10.1029/2001GL012897.

Bates, N.R., Knap, A.H., Michaels, A.F., 1998. Contribution of hurricanes to local and global estimates of air–sea exchange of CO_2. Nature 395, 58–61. https://doi.org/10.1038/25703.

Caldeira, K., Michael, E.W., 2003. Anthropogenic carbon and ocean pH. Nature 425, 365. https://doi.org/10.1038/425365a.

Cao, F., Ge, Y., Wang, J.F., 2013. Optimal discretization for geographical detectors-based risk assessment. GISci. Remote Sens. 50, 78–92. https://doi.org/10.1080/15481603.2013.778562.

Chau, T.T.T., Gehlen, M., Chevallier, F., 2022. A seamless ensemble-based reconstruction of surface ocean pCO_2 and air–sea CO_2 fluxes over the global coastal and open oceans. Biogeosciences 19, 1087–1109. https://doi.org/10.5194/bg-19-1087-2022.

Chen, T.Q., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Chen, S.L., et al., 2019. A machine learning approach to estimate surface ocean pCO_2 from satellite measurements. Remote Sens. Environ. 228, 203–226. https://doi.org/10.1016/j.rse.2019.04.019.

Chen, L., et al., 2021. Estimating soil moisture over winter wheat fields during growing season using machine-learning methods. IEEE 14, 3706–3718. https://doi.org/10.1109/JSTARS.2021.3067890.

Cho, D.J., et al., 2020. Comparative assessment of various machine learning-based Bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. Earth and space. Science 7 (4), e2019EA000740. https://doi.org/10.1029/2019EA000740.

Dickson, A., Sabine, C.L., Christian, J.R., 2007. Guide to best practices for ocean CO_2 measurements. In: Sidney: North Pacific Marine Science Organization, 191. PICES Special Publication.

Dixit, A., Lekshmi, K., Bharti, R., Chandan, M., 2019. Net sea-air CO_2 fluxes and modeled partial pressure of CO_2 in open ocean of bay of Bengal. IEEE 12, 2462–2469. https://doi.org/10.1109/JSTARS.2019.2902253.

Eray, O., Mert, C., Kisi, O., 2018. Comparison of multi-gene genetic programming and dynamic evolving neural-fuzzy inference system in modeling pan evaporation. Hydrol. Res. 49 (4), 1221–1233. https://doi.org/10.2166/nh.2017.076.

Fay, A.R., Mckinley, G.A., 2017. Correlations of surface ocean pCO_2 to satellite chlorophyll on monthly to interannual timescales. Glob. Biogeochem. Cycles 31, 436–455. https://doi.org/10.1002/2016GB005563.

Friedlingstein, P., et al., 2022. Global carbon budget 2022. Earth Syst. Sci. Data 14 (11), 4811–4900. https://doi.org/10.5194/essd-14-4811-2022.

Friedrich, T., Oschlies, A., 2009. Neural network-based estimates of North Atlantic surface pCO_2 from satellite data: a methodological study. J. Geophys. Res. Oceans 114 (C3), C03020. https://doi.org/10.1029/2007JC004646.

Hales, B., et al., 2012. Satellite-based prediction of pCO_2 in coastal waters of the eastern North Pacific. Prog. Oceanogr. 103, 1–15. https://doi.org/10.1016/j.pocean.2012.03.001.

Hou, W.Y., et al., 2022. A satellite-based dataset of global atmospheric carbon dioxide concentration with a spatial resolution of 2°×2.5° from 1992 to 2020. J. Glob. Change Data Discov. 6 (02), 191–199. +363-371.

Hu, Y.H., et al., 2023. A spatio-temporal LSTM model to forecast across multiple temporal and spatial scales. Eco. Inform. 69, 101687. https://doi.org/10.1016/j.ecoinf.2022.101687.

Jacob, M.M., et al., 2019. Salinity rain impact model (RIM) for SMAP. IEEE 12 (6), 1679–1687. https://doi.org/10.1109/JSTARS.2019.2907275.

James, C.O., et al., 2005. Anthropogenic Ocean acidification over the twenty-first century and its impact on calcifying organisms. Nature 437, 681–686. https://doi.org/10.1038/nature04095.

Jang, E., et al., 2017. Estimation of fugacity of carbon dioxide in the East Sea using in situ measurements and Geostationary Ocean color imager satellite data. Remote Sens. 9 (8), 821. https://doi.org/10.3390/rs9080821.

Jang, et al., 2022. Global sea surface salinity via the synergistic use of SMAP satellite and HYCOM data based on machine learning, 273 (2022), 112980. https://doi.org/10.1016/j.rse.2022.112980.

Jo, Y.H., et al., 2012. On the variations of sea surface pCO_2 in the northern South China Sea: a remote sensing based neural network approach. J. Geophys. Res. Oceans 117 (C8), C08022. https://doi.org/10.1029/2011JC007745.

Kevin, E.T., 2001. Climate variability and global warming. Science 293 (5527), 48–49. https://doi.org/10.1126/science.293.5527.48.

Krishna, N.K., et al., 2018. Ocean wave height prediction using ensemble of extreme learning machine. Neurocomputing 277 (14), 12–20. https://doi.org/10.1016/j.neucom.2017.03.092.

Krishna, K.V., Shanmugam, P., Nagamani, P.V., 2020. A multiparametric nonlinear regression approach for the estimation of global surface ocean pCO_2 using satellite oceanographic data. IEEE 13, 6220–6235. https://doi.org/10.1109/JSTARS.2020.3026363.

Krivoguz, D., et al., 2024. Geo-spatial analysis of urbanization and environmental changes with deep neural networks insights from a three-decade study in Kerch peninsula. Eco. Inform. 80, 102513. https://doi.org/10.1016/j.ecoinf.2024.102513.

Laith, A., et al., 2024. Particle swarm optimization algorithm: review and applications. Metaheuristic Optimiz. Algorithms 1-14. https://doi.org/10.1016/B978-0-443-13925-3.00019-4.

Landschützer, P., et al., 2016. Decadal variations and trends of the global ocean carbon sink. Glob. Biogeochem. Cycles 30 (10), 1396–1417. https://doi.org/10.1002/2015GB005359.

Lee, K., et al., 2006. Global relationships of total alkalinity with salinity and temperature in surface waters of the world's oceans. Geophys. Res. Lett. 33, L19605. https://doi.org/10.1029/2006GL027207.

Liu, Y.G., Robert, H.W., 2005. Patterns of ocean current variability on the West Florida shelf using the self-organizing map. J. Geophys. Res. 110 (C6). https://doi.org/10.1029/2004JC002786.

Liu, Q., et al., 2018. Carbon fluxes in the China seas: an overview and perspective. Sci. China Earth Sci. 61, 1564–1582. https://doi.org/10.1007/s11430-017-9267-4.

Lohrenz, S.E., et al., 2018. Satellite estimation of coastal pCO_2 and air-sea flux of carbon dioxide in the northern Gulf of Mexico. Remote Sens. Environ. 207, 71–83. https://doi.org/10.1016/j.rse.2017.12.039.

Long, J.C., et al., 2024. From meteorological to agricultural drought: propagation time and influencing factors over diverse underlying surfaces based on CNN-LSTM model. Eco. Inform. 82, 102681. https://doi.org/10.1016/j.ecoinf.2024.102681.

Luo, W., et al., 2016. Spatial association between dissection density and environmental factors over the entire conterminous United States. Geophys. Res. Lett. 43, 692–700. https://doi.org/10.1002/2015GL066941.

Ma, H.Z., Liu, S.M., 2016. The potential evaluation of multisource remote sensing data for extracting soil moisture based on the method of BP neural network. Can. J. Remote. Sens. 42 (2), 117–124. https://doi.org/10.1080/07038992.2016.1160773.

Margaret, O.O., et al., 2020. Variability of sea-air carbon dioxide flux in autumn across the Weddell gyre and offshore Dronning Maud land in the Southern Ocean. Front. Mar. Sci. 7. https://doi.org/10.3389/fmars.2020.614263.

Mario, H., 2004. Weddell Sea turned from source to sink for atmospheric CO_2 between pre-industrial time and present. Glob. Planet. Chang. 40 (3–4), 219–231. https://doi.org/10.1016/j.gloplacha.2003.08.001.

Monaco, C.L., et al., 2021. Distribution and long-term change of the sea surface carbonate system in the Mozambique Channel (1963–2019). Deep Sea Res. Part II 186-188, 104936. https://doi.org/10.1016/j.dsr2.2021.104936.

Moussa, H., et al., 2015. Satellite-derived CO_2 fugacity in surface seawater of the tropical Atlantic Ocean using a feedforward neural network. Int. J. Remote Sens. 37 (3), 580–598. https://doi.org/10.1080/01431161.2015.1131872.

Nakaoka, S., et al., 2013. Estimating temporal and spatial variation of ocean surface pCO_2 in the North Pacific using a self-organizing map neural network technique. Biogeosciences 10 (9), 6093–6106. https://doi.org/10.5194/bg-10-6093-2013.

Pérez, F.F., et al., 2024. An assessment of CO_2 storage and sea-air fluxes for the Atlantic Ocean and Mediterranean Sea between 1985 and 2018. Glob. Biogeochem. Cycles 38 (4). https://doi.org/10.1029/2023GB007862.

Pfeil, B., 2013. A uniform, quality controlled Surface Ocean CO_2 atlas (SOCAT). Earth Syst. Sci. Data Discus. 5, 125–143. https://doi.org/10.5194/essd-5-125-2013.

Rana, M.A., et al., 2019. Daily streamflow prediction using optimally pruned extreme learning machine. J. Hydrol. 577, 123981. https://doi.org/10.1016/j.jhydrol.2019.123981.

Rana, M.A., et al., 2021. Estimating reference evapotranspiration using hybrid adaptive fuzzy inferencing coupled with heuristic algorithms. Comput. Electron. Agric. 191 (106541). https://doi.org/10.1016/j.compag.2021.106541.

Rana, M.A., et al., 2024. Improved prediction of monthly streamflow in a mountainous region by metaheuristic-enhanced deep learning and machine learning models using hydroclimatic data. Theor. Appl. Climatol. 155, 205–228. https://doi.org/10.1007/s00704-023-04624-9.

Ren, Y., et al., 2014. Geographical modeling of spatial interaction between human activity and forest connectivity in an urban landscape of Southeast China. Landsc. Ecol. 29, 1741–1758. https://doi.org/10.1007/s10980-014-0094-z.

Reusch, D.B., Alley, R.B., Hewitson, B.C., 2007. North Atlantic climate variability from a self-organizing map perspective. J. Geophys. Res. 112 (D2). https://doi.org/10.1029/2006JD007460.

Reynaud, S., et al., 2003. Interacting effects of CO_2 partial pressure and temperature on photosynthesis and calcification in a scleractinian coral. Glob. Chang. Biol. 9, 1660–1668. https://doi.org/10.1046/j.1365-2486.2003.00678.x.

Richard, A.F., Scott, C.D., Sarah, R.C., 2009. Ocean acidification: present conditions and future changes in a high-CO_2 world. Oceanography 22 (4), 36–47. https://doi.org/10.5670/oceanog.2009.95.

Richardson, A.J., Risien, C., Shillington, F.A., 2003. Using self-organizing maps to identify patterns in satellite imagery. Prog. Oceanogr. 59, 223–239. https://doi.org/10.1016/j.pocean.2003.07.006.

Rödenbeck, C., et al., 2015. Data-based estimates of the ocean carbon sink variability – first results of the Surface Ocean pCO_2 mapping intercomparison (SOCOM). Biogeosciences 12, 14049–14104. https://doi.org/10.5194/bgd-12-14049-2015.

Sabine, C.L., et al., 2004. The oceanic sink for anthropogenic CO_2. Science 305, 367–371. https://doi.org/10.1126/science.1097403.

Salim, H., et al., 2023. River water temperature prediction using hybrid machine learning coupled signal decomposition: EWT versus MODWT. Eco. Inform. 78 (102376). https://doi.org/10.1016/j.ecoinf.2023.102376.

Salisbury, J.E., et al., 2008. Seasonal observations of surface waters in two Gulf of Maine estuary-plume systems: relationships between watershed attributes, optical measurements and surface pCO_2. Estuar. Coast. Shelf Sci. 77, 245–252. https://doi.org/10.1016/j.ecss.2007.09.033.

Siegenthaler, U., Sarmiento, J.L., 1993. Atmospheric carbon dioxide and the ocean. Nature 365, 119–125. https://doi.org/10.1038/365119a0.

Signorini, S.R., et al., 2013. Surface Ocean pCO_2 seasonality and sea-air CO_2 flux estimates for the north American east coast. JGR Oceans 118, 5439–5460. https://doi.org/10.1002/jgrc.20369.

Song, Z.G., et al., 2023. Construction of a high spatiotemporal resolution dataset of satellite-derived pCO_2 and Air–Sea CO_2 flux in the South China Sea (2003–2019). IEEE 61, 4207015. https://doi.org/10.1109/TGRS.2023.3306389.

Sridevi, B., Sarma, V., 2021. Role of river discharge and warming on ocean acidification and pCO_2 levels in the Bay of Bengal. Tellus Ser. B Chem. Phys. Meteorol. 73, 1–20. https://doi.org/10.1080/16000889.2021.1971924.

Sujatha, M., et al., 2023. 1D convolutional neural networks-based soil fertility classification and fertilizer prescription. Eco. Inform. 78, 102295. https://doi.org/10.1016/j.ecoinf.2023.102295.

Telszewski, M., et al., 2009. Estimating the monthly pCO2 distribution in the North Atlantic using a self-organizing neural network. Biogeosciences 6, 1405–1421. https://doi.org/10.5194/bg-6-1405-2009.

Todorova, Y., Lincheva, S., Yotinov, I., Topalova, Y., 2016. Contamination and ecological risk assessment of long-term polluted sediments with heavy metals in small hydropower cascade. Water Resour. Manag. 30, 4174–4184. https://doi.org/10.1007/s11269-016-1413-8.

Turk, D., Book, J.W., Mcgillis, W.R., 2013. pCO_2 and CO_2 exchange during high bora winds in the northern Adriatic. J. Mar. Syst. 117–118, 65–71. https://doi.org/10.1016/j.jmarsys.2013.02.010.

Wang, J.C., Wang, Y., 2022. Evaluation of the ERA5 significant wave height against NDBC buoy data from 1979 to 2019. Mar. Geod. 45 (2), 151–165. https://doi.org/10.1080/01490419.2021.2011502.

Wang, J.F., et al., 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China. Int. J. Geogr. Inf. Sci. 24, 107–127. https://doi.org/10.1080/13658810802443457.

Wang, Y.J., et al., 2021. Carbon sinks and variations of pCO_2 in the Southern Ocean from 1998 to 2018 based on a deep learning approach. IEEE 14, 3495–3503. https://doi.org/10.1109/JSTARS.2021.3066552.

Wang, J., et al., 2023. Reconstruction of surface seawater pH in the North Pacific. Sustainability 15 (7), 5796. https://doi.org/10.3390/su15075796.

Weiss, R.F., 1974. Carbon dioxide in water and seawater: the solubility of a non-ideal gas. Mar. Chem. 2, 203–215. https://doi.org/10.1016/0304-4203(74)90015-2.

Yang, B., Byrne, R.H., Wanninkhof, R., 2015. Subannual variability of total alkalinity distributions in the northeastern Gulf of Mexico: subannual GOM alkalinity variations. JGR Oceans 120, 3805–3816. https://doi.org/10.1002/2015JC010780.

Yu, S.J., et al., 2023. Satellite-estimated air-sea CO_2 fluxes in the Bohai Sea, Yellow Sea, and East China Sea: patterns and variations during 2003–2019. Sci. Total Environ. 904 (15), 166804. https://doi.org/10.1016/j.scitotenv.2023.166804.

Zafar, S., et al., 2021. Thermophysical properties using ND/water nanofluids: an experimental study, ANFIS-based model and optimization. J. Mol. Liq. 330 (115659). https://doi.org/10.1016/j.molliq.2021.115659.

Zemmelink, H.J., et al., 2006. CO_2 deposition over the multi-year ice of the western Weddell Sea. Geophys. Res. Lett. 33 (13). https://doi.org/10.1029/2006GL026320.

Zhang, H., et al., 2023. Deep learning approach for forecasting sea surface temperature response to tropical cyclones in the Western North Pacific. Deep-Sea Res. I 197, 104042. https://doi.org/10.1016/j.dsr.2023.104042.

Zhao, X.L., et al., 2021. Comparing deep learning with several typical methods in prediction of assessing chlorophyll-a by remote sensing: a case study in Taihu Lake, China. Water Supply 21 (7), 3710–3724. https://doi.org/10.2166/ws.2021.137.

Zhao, Y.F., et al., 2024. Spatio-temporal prediction of groundwater vulnerability based on CNN-LSTM model with self-attention mechanism a case study in Hetao plain, northern China. J. Environ. Sci. https://doi.org/10.1016/j.jes.2024.03.052.

Zhong, G., et al., 2021. Reconstruction of global surface ocean pCO_2 using region-specific predicators based on a stepwise FFNN regression algorithm. Copernicus GmbH 19, 1–26. https://doi.org/10.5194/bg-19-845-2022.

Marrec, P., et al., 2015. Dynamics of air–sea CO_2 fluxes in the northwestern European shelf based on voluntary observing ship and satellite observations. Biogeosciences 12 (18), 5371–5391. https://doi.org/10.5194/bg-12-5371-2015.