Reconstruction of Global Ocean Surface pCO2 and Air-Sea CO2 Flux: Based on Multigrained Cascade Forest Model

Wanqin Zhong, Xin Ma, Tianqi Shi, Ge Han, Haowei Zhang, Wei Gong

JGR Oceans

RESEARCH ARTICLE
10.1029/2024JC021483

Reconstruction of Global Ocean Surface pCO_2 and Air-Sea CO_2 Flux: Based on Multigrained Cascade Forest Model

Wanqin Zhong[1], Xin Ma[1,2], Tianqi Shi[3], Ge Han[4], Haowei Zhang[5], and Wei Gong[2,5,6]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, [2]Wuhan Institute of Quantum Technology, Wuhan, China, [3]Laboratoire des Sciences du Climat et del'Environnement, LSCE/IPSL, Guyancourt, France, [4]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, [5]School of Electronic Information, Wuhan University, Wuhan, China, [6]Luojia Laboratory, Wuhan University, Wuhan, China

Abstract Quantifying the role of air-sea CO_2 exchange is essential for accurately estimating the global carbon balance, which is dependent on the spatial and temporal resolution of ocean surface carbon dioxide partial pressure ($pCO_{2(sw)}$). When dealing with the global ocean as a vast and complex system, most existing studies tend to partition the global ocean into small-scale regions. To account for interactions of environmental variables across multiple regions, we used machine learning algorithms to holistically reconstruct a 20-year global $pCO_{2(sw)}$ map at a high resolution of 4×4 km based on products from the Moderate Resolution Imaging Spectroradiometer, reanalysis data, and Surface Ocean CO_2 Atlas. Three machine learning methods were compared, with multigrained cascade forest (gcForest) demonstrating the highest accuracy in global reconstruction (r^2 of 0.92, root mean square error of 13.46, and mean absolute error of 7.34 μatm). The global $pCO_{2(sw)}$ has shown a steady increase at an average annual growth rate of 1.95 ± 0.05 μatm yr^{-1}, controlled mainly by sea surface temperature and chlorophyll concentration. This study covers an ocean area of approximately 335×10^6 km^2, encompassing over 95% of the annual average carbon sink area. During 20 years, the daily CO_2 flux decreased by 0.44 mmol m^{-2} d^{-1}, while the proportion of carbon sink area remained constant, indicating ocean's carbon uptake capacity per unit area has been increasing.

Plain Language Summary We utilized machine learning algorithms to holistically reconstruct $pCO_{2(sw)}$ data on a global scale, using key environmental variables including absorption due to gelbstoff and detritus at 443 (Adg), chlorophyll-a concentration (Chl-a), sea surface temperature (SST), mixed layer depth, and remote sensing reflectance (R_{rs}). The multigrained cascade forest (gcForest) reconstruction method demonstrated high accuracy and applicability, improving the spatiotemporal resolution of the $pCO_{2(sw)}$ data set. This data product effectively captured the spatial variability, asymmetric bimodal seasonal characteristics, and the 20-year growth trend of $pCO_{2(sw)}$. Furthermore, the global air-sea CO_2 flux was estimated using the reconstructed $pCO_{2(sw)}$ data set, with additional analyses performed on the monthly maxima of global CO_2 source and sink flux.

Key Points:

- Reconstruct monthly maps of global ocean surface pCO_2 holistically with a resolution of 4×4 km using multigrained cascade forest
- Global ocean surface pCO_2 is controlled mainly by sea surface temperature and chlorophyll concentration holistically
- The increase in carbon absorbed by the ocean was linked with enhancement in carbon uptake capacity instead of expansion of carbon sinks areas

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

X. Ma,
maxinwhu@whu.edu.cn

Citation:

Zhong, W., Ma, X., Shi, T., Han, G., Zhang, H., & Gong, W. (2025). Reconstruction of global ocean surface pCO_2 and air-sea CO_2 flux: Based on multigrained cascade forest model. *Journal of Geophysical Research: Oceans*, *130*, e2024JC021483. https://doi.org/10.1029/2024JC021483

Received 20 JUN 2024
Accepted 1 FEB 2025

Author Contributions:

Conceptualization: Tianqi Shi
Data curation: Wanqin Zhong, Xin Ma, Ge Han, Haowei Zhang
Formal analysis: Wanqin Zhong
Funding acquisition: Xin Ma, Wei Gong
Investigation: Wanqin Zhong, Wei Gong
Methodology: Wanqin Zhong, Tianqi Shi
Resources: Ge Han, Haowei Zhang
Visualization: Wanqin Zhong
Writing – original draft: Wanqin Zhong, Xin Ma, Tianqi Shi
Writing – review & editing: Wanqin Zhong, Xin Ma, Tianqi Shi, Haowei Zhang, Wei Gong

1. Introduction

The oceans, as the largest and most persistent carbon sink, play an irreplaceable and prominent role in the global carbon budget by absorbing approximately a quarter of the anthropogenic atmospheric CO_2 (Friedlingstein et al., 2022), which helps mitigate global warming. The annual mean oceanic CO_2 uptake was estimated at -2.7 ± 0.3 Pg C yr^{-1} during the period 1990 through 2019 (Gruber et al., 2023), increasing to -2.9 ± 0.4 Pg C yr^{-1} in 2021 (Friedlingstein et al., 2022). The absorption of atmospheric CO_2 leads to changes in ocean chemistry that may profoundly affect marine ecosystems and contribute to extreme weather events (Caldeira & Wickett, 2005; Fangohr et al., 2008). Besides, the role of the oceans in the climate-carbon feedback has become increasingly important over time and may outweigh the impact of terrestrial sources (Randerson et al., 2015).

Air-sea CO_2 flux is a key metric for quantifying the exchange of carbon dioxide between the oceans and the atmosphere. The difference between atmospheric CO_2 partial pressure ($pCO_{2(atm)}$) and ocean surface CO_2 partial pressure $pCO_{2(sw)}$) controls the air-sea CO_2 gas exchange and consequently regulates the air-sea CO_2 flux (Arrigo & Van Dijken, 2007; Fennel et al., 2008; Takahashi et al., 2014). Positive and negative $pCO_{2(sw)}$ values determine

the direction of CO_2 transfer between the atmosphere and seawater. Positive values indicate that seawater is oversaturated with CO_2 relative to the atmosphere, resulting in a net release of CO_2 to the atmosphere, and vice versa (Rippeth et al., 2014). Direct and indirect $pCO_{2(sw)}$ observations have traditionally been obtained through ships, moorings, drifters, and autonomous surface platforms. However, the sparsity of observations, particularly in the Southern Hemisphere, has resulted in significant spatial and temporal uncertainties, as well as discrepancies in the estimation of air-sea CO_2 flux (Ishii et al., 2014; Mackay & Watson, 2021).

Existing studies have shown that $pCO_{2(sw)}$ is influenced by the biological processes, the interplay of oceanic vertical mixing, the equilibrium dynamics between the atmosphere and ocean, and thermodynamic processes (Bai et al., 2015; Sharp et al., 2022). These factors can be captured by chlorophyll (Chl-a), mixed layer depth (MLD), absorption due to gelbstoff and detritus (Adg), and sea surface temperature (SST), respectively (Tu et al., 2021). The rapid advancement of satellite remote sensing technology provides multitimescale global satellite data on environmental variables, bringing a new direction to address the sparsity of $pCO_{2(sw)}$ observations (S. L. Chen et al., 2019; Marrec et al., 2015). Remote sensing reflectance ($R_{rs}(\lambda)$), defined as the ratio of water-leaving radiance below the surface of water to downwelling irradiance above the surface, is utilized to characterize the spectral properties of the upper water layers (Mobley, 1999). Bio-optical algorithms are then applied to the $R_{rs}(\lambda)$ to generate robust estimates of colored dissolved organic matter (CDOM), suspended sediment (SS), and Chl-a at a global scale (Clark, 1981; McClain et al., 2006; Pabi & Arrigo, 2006). Thus, $R_{rs}(\lambda)$ in various spectral bands, characterized as apparent optical properties and observed directly by remote sensors, can serve as environmental variables for reconstructing $pCO_{2(sw)}$.

Traditional empirical regressions and machine learning algorithms have been employed to simulate and predict $pCO_{2(sw)}$ based on satellite-observed oceanic environmental variables. Within the developed machine learning models, the root mean square error (RMSE) of satellite-estimated $pCO_{2(sw)}$ have decreased from ~20 to ~10 μatm by incorporating Chl-a with other physical variables such as SST, MLD, and sea surface salinity (SSS) (Friedrich & Oschlies, 2009; Landschutzer et al., 2013; Lefevre et al., 2005; Song et al., 2023). After lots of studies conducted on various small-scale marine areas with good degrees of fitting, researchers gradually paid attention to global $pCO_{2(sw)}$. Landschützer et al. (2014) utilized a feed-forward network method to establish a nonlinear relationship between $pCO_{2(sw)}$ and environmental variables in oceanic biogeochemical provinces partitioned by a self-organizing map. The oceans are commonly divided into biogeochemical subregions for a separate study, with models trained within these individual subregions (Laruelle et al., 2017). This is because $pCO_{2(sw)}$ is generally controlled by one or two physicochemical processes in different oceanic regions (Bai et al., 2015; Yu et al., 2023). This partitioned reconstruction method is widely used to generate global $pCO_{2(sw)}$ products with RMSEs of 20.48 μatm (Chau et al., 2022) and 17.99 μatm (Zhong et al., 2022). However, the nongeneralizability of regional models often leads to inaccuracies when applied outside their training regions. While the prediction error of this method is low, the accuracy is strongly dependent on the delineation of biogeochemical subregions (Zhong et al., 2022). Furthermore, thermohaline circulation interactions are often neglected when reconstructing $pCO_{2(sw)}$ in subregions. Wu et al. (2024) recently introduced a global holistic reconstruction approach based on optimized random forest and estimated the ORF-pCO_2 of 0.25° × 0.25° with an RMSE of 15.34 μatm, demonstrating interannual variations in $pCO_{2(sw)}$ over the past 10 years.

This study advances existing research by generating a high-resolution $pCO_{2(sw)}$ product spanning the past two decades, with a focus on interdecadal variability. The aim is to apply a machine learning model to accurately reconstruct global $pCO_{2(sw)}$ holistically while accounting for the interactions of environmental variables across multiple regions. Absorption due to gelbstoff and detritus (Adg), chlorophyll (Chl-a), MLD, temperature (SST), and R_{rs} were selected as input variables to reconstruct the global $pCO_{2(sw)}$ product using a machine learning model. A comparative analysis with other machine learning models, including eXtreme Gradient Boosting (XGBoost) and random forest, demonstrated that the multigrained cascade forest (gcForest) model achieved the highest accuracy in reconstructing global data. Subsequently, a 20-year monthly gridded data set of $pCO_{2(sw)}$ and air-sea CO_2 flux with a spatial resolution of 4 × 4 km was established. Finally, the spatiotemporal patterns and driving factors of $pCO_{2(sw)}$ and air-sea CO_2 flux calculated by $pCO_{2(sw)}$ were analyzed.

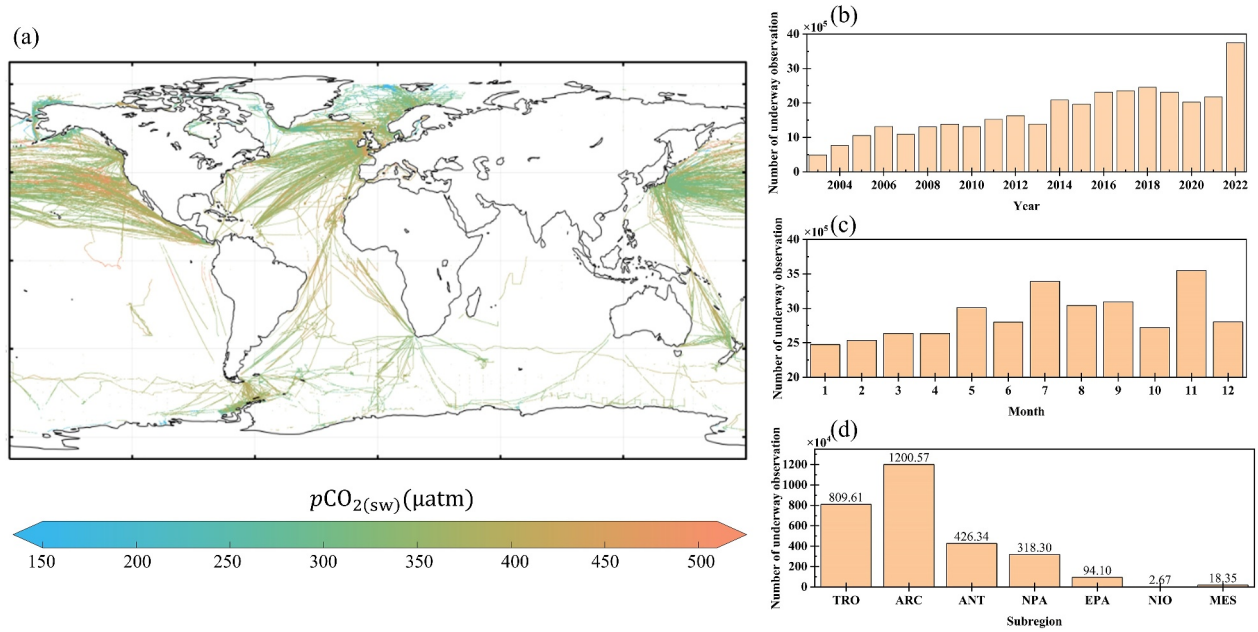

Figure 1. Visualization of cruise tracks and bar charts for observation statistics: (a) Scatter plots of SOCAT observations. (b), (c), and (d) Bar charts presenting number statistics for underway seawater pCO_2 observations categorized by year, month, and subregion.

2. Materials

2.1. Underway Observation Data of SOCAT

This paper utilized underway $pCO_{2(sw)}$ observation data from the Surface Ocean CO_2 Atlas (SOCAT) to construct empirical relationship between $pCO_{2(sw)}$ and environmental variables using machine learning–based models. The SOCAT version 2023 used in the study was released on 20 June 2023 (Bakker et al., 2023). Seawater pCO_2 data were collected by the international marine carbon research community along the cruise tracks shown in Figure 1a. The density of cruise tracks varies across different oceanic regions. More in situ observations were measured in the Northern Hemisphere than in the Southern Hemisphere. The data distribution of $pCO_{2(sw)}$ across years and months is shown in Figures 1b and 1c. With updated versions of the SOCAT data set, the amount of underway observation data has gradually increased, nearly doubling by 2022. This surge in data indicates that monthly in situ observations are now more abundant. Additionally, in situ measurements with an accuracy below 10 μatm were excluded to guarantee data reliability.

2.2. Environmental Variables

The study considered five environmental variables: absorption due to gelbstoff and detritus at 443 (Adg) to represent dissolved organic matter effects, chlorophyll-a concentration (Chl-a) for biological absorption effects on $pCO_{2(sw)}$, SST for thermodynamic effects, MLD for physical vertical mixing effects (Tu et al., 2021), and remote sensing reflectance $R_{rs}(\lambda)$ to capture optical information observable in visible light. The Moderate Resolution Imaging Spectroradiometer (MODIS) provides $R_{rs}(\lambda)$ level 3 products spanning 10 bands within the visible wavelength range (412 ~ 678 nm). The spectral bands of $R_{rs}(\lambda)$ were selected based on two criteria: (a) their relevance to estimating three water constituents (Chl-a, CDOM, and SS) and (b) their coverage from the blue to red bands provided by MODIS. These criteria ensured that the selected bands could serve as environmental variables for reconstructing $pCO_{2(sw)}$. Specifically, the $R_{rs}(\lambda)$ at blue wavelengths (443 and 488 nm) and green wavelengths (555 nm) are strongly correlated with ocean surface Chl-a (Hu et al., 2012; O'Reilly & Werdell, 2019; Reynolds et al., 2001). Additionally, R_{rs} peaks in the 448–555 nm bands correspond to different SSS ranges, and $R_{rs}(667)$ from MODIS has been used as a surrogate for sediment concentration in the water column, both affecting SSS-CDOM retrievals (S. L. Chen & Hu, 2017). Moreover, $R_{rs}(488)$ and $R_{rs}(678)$ are sensitive and important to both Chl-a and SSS estimation compared to other MODIS bands (He et al., 2020; Wang &

Deng, 2018). Five visible spectral bands matching these criteria, including $R_{rs}(443)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, and $R_{rs}(678)$, were selected.

The monthly data set for Adg, Chl-a, SST, and $R_{rs}(\lambda)$ from January 2003 to December 2022 were derived from the MODIS level 3 products at a spatial resolution of 4 km (NASA Ocean Color). The 8-day MLD data set, derived from the Hybrid Coordinate Ocean Model (HYCOM), was provided by Ocean Productivity at a spatial resolution of 0.125° (Ocean Productivity: Input HYCOM MLD data oregonstate.edu).

2.3. Matching Up and Classification of Data Sets

To reduce noise of global environmental variables, median filtering with a 3 × 3-pixel box was applied prior to data set matching. To achieve a consistent spatial resolution, MLD data were bilinearly interpolated to match the spatial resolution of 4 km. Regarding temporal resolution, the 8-day MLD was accumulated in monthly scale and averaged when multiple MLD values occurred within the same grid. SOCAT underway observations were averaged within the same month and 4-km grid to construct matchup pairs with environmental variables. Ultimately, 1,876,199 valid matchup pairs of observed $pCO_{2(sw)}$ and satellite observations from 2003 to 2022 were available for model reconstruction. To standardize the magnitudes of environmental variables, MLD was $\log_{10}$-transformed before predicting using the machine learning–based model. Due to the limitation of MODIS satellite data, only summer data (from May to September) for the ARC (Arctic Ocean) and winter data (from December to February) for the ANT (Antarctic) were included in the analysis.

After matching the data, a total of 1,876,199 matchup pairs were randomly divided into three groups: training group (75%), validation group (15%), and test group (10%). Before running the models, a fixed test group was extracted to evaluate the performance of XGBoost, random forest, and gcForest. To avoid the contingency effects from a single division of training and validation groups, each model's performance was assessed using 10-fold cross-validation.

3. Methods and Calculations

3.1. Machine learning Models

The multigrained cascade forest (gcForest) is a novel method that builds a decision tree ensemble inspired by the convolutional deep neural networks (CNN) (Zhou & Feng, 2017). Random forest excels at handling complex interactions among classification features and is robust against noise and incomplete data. Deep neural networks operate through layer-by-layer processing for feature generation, processing, and transmission. gcForest adopts a cascade structure similar to CNN, where each layer receives feature vectors from the previous layer using a decision tree ensemble and forwards the results to the next layer. To enhance model's generalization, gcForest employs two types of random forest—completely random tree forest and standard random forest—within the decision tree forest ensemble of its cascade structure. After extending a new level, the performance of the entire cascade is estimated on the validation group, and the iterative process of training is terminated if there is no significant performance gain. Therefore, the model automatically determines the number of cascade levels based on validation group accuracy. In summary, gcForest combines the advantages of CNN and random forest to optimize model complexity performance based on its cascade structure. Figure 2 outlines the process of training the empirical model with $pCO_{2(sw)}$ underway observations and satellite data using gcForest.

Extreme Gradient Boosting (XGBoost), a gradient-boosting decision tree algorithm, has been widely applied in the prediction of geological variables (Chen & Guestrin, 2016; Yu et al., 2023). A shared characteristic among XGBoost, random forest, and gcForest is that the number of trees directly influences their fitting accuracy. However, XGBoost adopts the "boosting" method, sampling data sets according to the error rate, whereas random forest adopts the "bagging" method, assigning equal weights to all training data. In this study, XGBoost, random forest, and gcForest were employed to reconstruct global seawater $pCO_{2(sw)}$ and compare their model performances. All models were trained with identical input parameters and data sets, ensuring consistency across the training, validation, and testing groups.

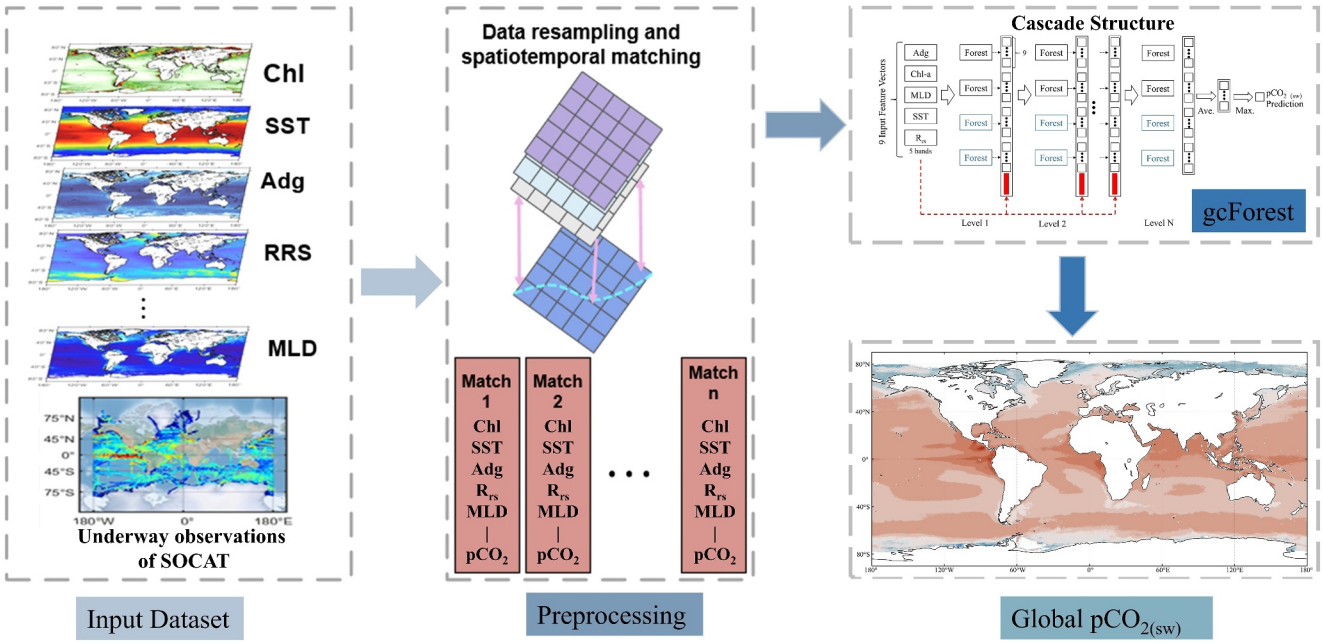

Figure 2. Flowchart of the data processing procedure in this study. Each level of the cascade forest consists of two standard random forests (shown in black) and two completely random tree forests (shown in blue).

3.2. Flux Calculation

The air-sea CO_2 fluxes across the air-water interface (FCO_2) were computed using the following equation (Frankignoulle, 1988):

$$FCO_2 = k_w \times K_0 \times \Delta pCO_2 = k_w \times K_0 \times \left(pCO_{2(sw)} - pCO_{2(atm)}\right) \quad (1)$$

where k_w (cm h^{-1}) refers to the gas transfer velocity, K_0 $\left(\text{mol } L^{-1} \text{ atm}^{-1}\right)$ refers to the CO_2 solubility constant, and ΔpCO_2 refers to the difference between $pCO_{2(sw)}$ and $pCO_{2(atm)}$ (Takahashi et al., 2009). k_w was calculated based on a quadratic dependence on wind speed from ERA5 reanalysis (Copernicus Interactive Climate Atlas) and SST from MODIS (Wanninkhof, 2014). K_0 was determined using SSS estimates from the NASA Estimating the Circulation and Climate of the Ocean project (APDRC Datadoc | ECCO2 Cube92 model output (hawaii.edu)) and SST (Weiss, 1974). After obtaining k_w, K_0, and $pCO_{2(sw)}$, $pCO_{2(atm)}$ was calculated using (Dickson et al., 2007):

$$pCO_{2(atm)} = XCO_2 \left(P_{baro} - P_{sw}\right) \quad (2)$$

where XCO_2 (Global Monitoring Laboratory-Carbon Cycle Green house Gases (noaa.gov)) refers to atmospheric CO_2 mole fraction, P_{baro} refers to sea level pressure, which comes from ERA5 reanalysis, and P_{sw} refers to water vapor pressure (Takahashi et al., 1993). Negative values of F indicate CO_2 transfer from the atmosphere to seawater, signifying that the region acts as a CO_2 sink.

3.3. Performance Evaluation

The criteria for evaluating the model's performance and reliability include the coefficient of determination (r^2; Equation 3), root mean square error (RMSE; Equation 4), mean absolute error (MAE; Equation 5), and mean bias (MB; Equation 6). Each criterion is used to assess an efficiency index from the observed and predicted values.

Table 1
Model Performance and Uncertainty of $pCO_{2(sw)}$ in Global and 7 Regions

Region	Latitude	Longitude	r^2	RMSE (μatm)	MAE (μatm)	$\theta_{pCO_{2(sw)}}$ (μatm)
Global	All	All	0.92	13.46	7.34	8.08
Tropical ocean (TRO)	$30° S - 30° N$	All	0.90	10.40	6.11	6.25
Arctic Ocean (ARC)	$60° N - 90° N$	All	0.92	19.39	10.4	11.48
Antarctic (ANT)	$90° S - 60° S$	All	0.93	13.45	7.38	8.14
North Pacific (NPA)	$30° N - 50° N$	$180° W - 120° W, 140° E - 180° E$	0.89	12.94	7.26	7.71
Equatorial Pacific (EPA)	$10° S - 10° N$	$180° W - 80° W$	0.87	19.94	11.74	11.69
North Indian Ocean (NIO)	$0° N - 25° N$	$50° E - 110° E$	0.53	16.51	11.40	10.22
Mediterranean Sea (MES)	$30° N - 45° N$	$0° E - 30° E$	0.96	8.59	5.30	5.83

$$r^2 = \left(\frac{Cov(pCO_{2obs}, pCO_{2pred})}{\sqrt{Var[pCO_{2obs}]Var[pCO_{2pred}]}} \right)^2 \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(pCO_{2obs,i} - pCO_{2pred,i}\right)^2} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|pCO_{2obs,i} - pCO_{2pred,i}\right| \tag{5}$$

$$MB = \frac{1}{n}\sum_{i=1}^{n}\left(pCO_{2obs,i} - pCO_{2pred,i}\right) \tag{6}$$

where pCO_{2obs} and pCO_{2pred} refer to the in situ values and the corresponding predicted values produced by gcForest for each "i" observation, respectively. Cov is the covariance, and Var is the variance.

The MB provides a measure of the direction of the discrepancy: positive bias values indicate overestimation, while negative bias values indicate underestimation. It is important to note that a bias value close to zero does not indicate the absence of systematic bias, as a significant positive offset at one point in space or time may cancel out a significant negative shift elsewhere (Gloege et al., 2022).

3.4. Uncertainty Calculation

The uncertainty in $pCO_{2(sw)}$ is a composite of observation uncertainty (θ^2_{obs}), mapping uncertainty (θ^2_{map}), and griding uncertainty (θ^2_{grid}):

$$\theta_{pCO_{2(sw)}} = \sqrt{\theta^2_{obs} + \theta^2_{map} + \theta^2_{grid}} \tag{7}$$

The process for calculating these three components was detailed in Landschützer et al. (2018) and Sharp et al. (2022). Briefly, θ^2_{obs} was evaluated as the weighted average of observations assigned to each quality control flag in SOCAT. θ^2_{map} was evaluated using the RMSE for the test group, while θ^2_{grid} was evaluated as the unweighted standard deviation within each grid cell containing two or more observations. To better assess the global distribution and errors of $pCO_{2(sw)}$ and FCO_2, the ocean was divided into seven representative oceans with reference to existing studies (Jang et al., 2022; Olmedo et al., 2020, 2021), as shown in Table 1. In addition, although the Mediterranean Sea (MES) serves as an important carbon sink connected to the Atlantic Ocean, observations in this region are limited in existing studies (D'Ortenzio et al., 2008).

4. Results

4.1. Model Calibration and Validation

The gcForest model was trained using Adg, Chl-a, SST, MLD, R_{rs} (488), and R_{rs} (555) to reconstruct $pCO_{2(sw)}$ holistically, generating a 20-year global $pCO_{2(sw)}$ time series data set.

4.1.1. Overall Model Performances

After selection of model, the matchup database, with a volume of 1,876,199 records, was separated into a training group (80%) and a validation group (20%). The seven selected environmental variables were used to train the gcForest model, and the accuracy for both the training and verification groups is displayed in Figure 8. Overall, the gcForest model achieved a test RMSE of approximately 13.46 μatm and an MAE of 7.34 μatm. Moreover, MB is clearly related to $pCO_{2(sw)}$, indicating overestimation below 200 μatm and underestimation above 500 μatm. However, MB shows no direct relationship with the year or the number of matchup pairs (Table S1 in Supporting Information S1). Therefore, when discussing the temporal pattern of $pCO_{2(sw)}$, it can be considered that the temporal changes do not appear to introduce the deviation of systematic error of $pCO_{2(sw)}$. These results confirm that the gcForest is capable of establishing 20-year monthly averaged gridded products at a spatial resolution of 4 × 4 km.

4.1.2. Uncertainty Analysis

The $\theta_{pCO_{2(sw)}}$ was evaluated separately for each of the eight regions listed in Table 1. The global $pCO_{2(sw)}$ data set has an uncertainty of 8.08 μatm, with higher than average uncertainties in the MES, tropical ocean (TRO), and North Pacific (NPA) regions. Field measurement for the North Indian Ocean (NIO) are lacking in the SOCAT, resulting in an unsatisfactory predictive capacity ($r^2 = 0.53$, $\theta_{pCO_{2(sw)}} = 10.22$ μatm). Due to the spatial correlation of $pCO_{2(sw)}$ and the autocorrelation within the gcForest, regional or temporal averaged uncertainties do not scale exactly with $\theta_{pCO_{2(sw)}}$ (Landschützer et al., 2014).

4.2. Spatiotemporal Pattern of $pCO_{2(sw)}$ and Its Global Trends

The spatial pattern of the global $pCO_{2(sw)}$ distribution remained consistent each year. The spatial distribution of the average $pCO_{2(sw)}$ from 2003 to 2022 is illustrated in Figure 4, and its monthly climatology variation is depicted in Figure 5. This distribution pattern revealed a gradual decrease in $pCO_{2(sw)}$ values from the highest levels in the TRO toward lower latitudes in both the Northern and Southern Hemispheres, with no significant relationship observed with longitude (Figures 4b and 4c). This spatial pattern aligned with findings from previous studies (Landschutzer et al., 2016; Takahashi et al., 2009). In February, the highest $pCO_{2(sw)}$ values were observed in the Eastern Equatorial Pacific (EPA) and NIO, while lower $pCO_{2(sw)}$ were located in the MES and Antarctic (ANT) regions. By May, these differences in $pCO_{2(sw)}$ values between regions had significantly narrowed. In the summer, the maximum occurred in the MES where $pCO_{2(sw)}$ peaked at approximately 414.60 μatm in July, while the lowest values shifted to the Arctic Ocean (ARC). In September, a significant area of high $pCO_{2(sw)}$ emerged on the western side of the NIO, attributed to limited field measurements exceeding 400 μatm during that period. As the season progressed into winter, the distribution patterns in the ANT, a region with the lowest $pCO_{2(sw)}$, became more distinct.

The seasonal variations of $pCO_{2(sw)}$ across the globe and seven subregions over 20 years are illustrated in Figures 5 and 6. $pCO_{2(sw)}$ showed a bimodal distribution, similar to the observed pattern with CO_2. Specifically, $pCO_{2(sw)}$ increased during the spring, reaching a peak in March, followed by a decrease to a minimum around June. Subsequently, it began to rise again from summer through fall, continuing until around November. Beyond seasonal variations, a notable interannual trend was identified during the study period. Between 2003 and 2022, global $pCO_{2(sw)}$ increased from 355.24 to 387.91 μatm (Figure 6), with an average growth rate of 1.95 ± 0.05 μatm yr^{-1}. This rate was determined through regression analysis with seasonal adjustment (Takahashi et al., 2009). In five subregions (excluding the polar regions), the growth rates were higher than the global average. The MES exhibited the highest growth rate at 2.71 ± 0.25 μatm yr^{-1}, followed by the NIO at 2.38 ± 0.06 μatm yr^{-1}. The TRO, NPA, and EPA showed slightly higher rates than the global average, at

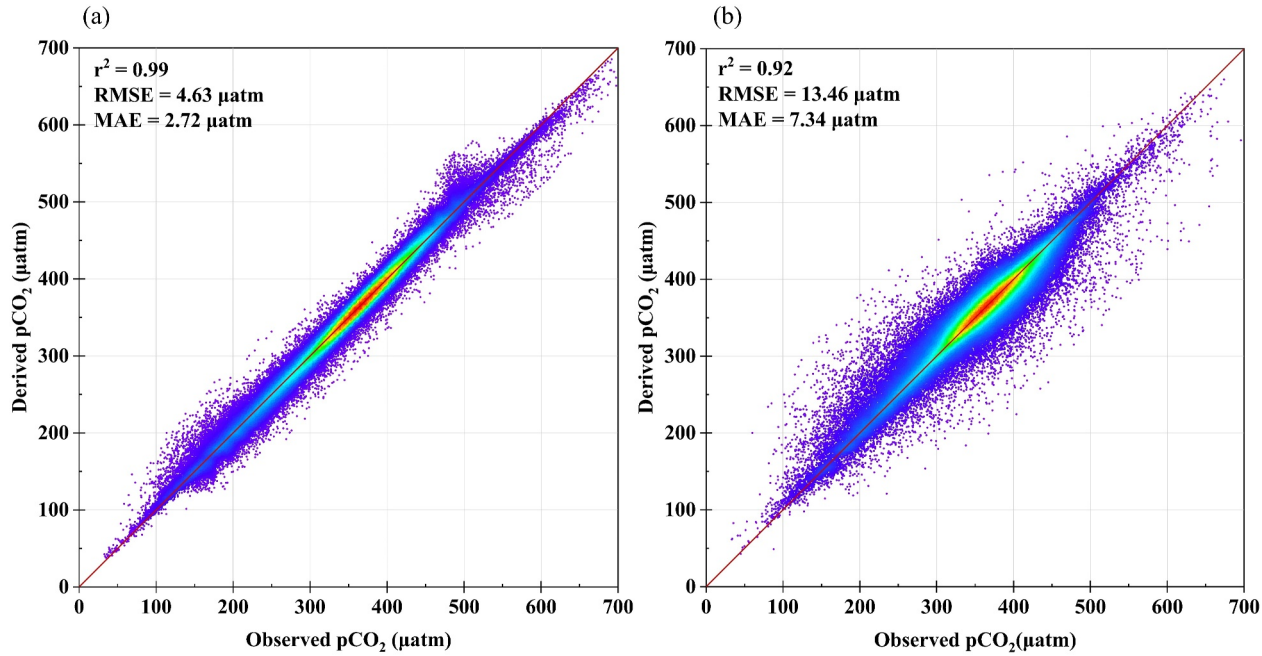

Figure 3. The performance of the gcForest method in estimating $pCO_{2(sw)}$ during (a) the model training phase and (b) the independent validation phase.

2.13 ± 0.04, 2.13 ± 0.11, and 2.10 ± 0.10 µatm yr^{-1}, respectively. Due to the limitation of MODIS satellite data, only summer data (from May to September) for the ARC and winter data (from December to February) for the ANT were analyzed, potentially affecting the representativeness of these regions. Consequently, the mean growth rates for ARC and ANT were calculated as 1.05 ± 0.20 and 1.33 ± 0.12 µatm yr^{-1}, respectively. Takahashi et al., 2009 estimated the global ocean's mean growth rate from 1970 to 2007 as 1.69 ± 0.51 µatm yr^{-1}. In

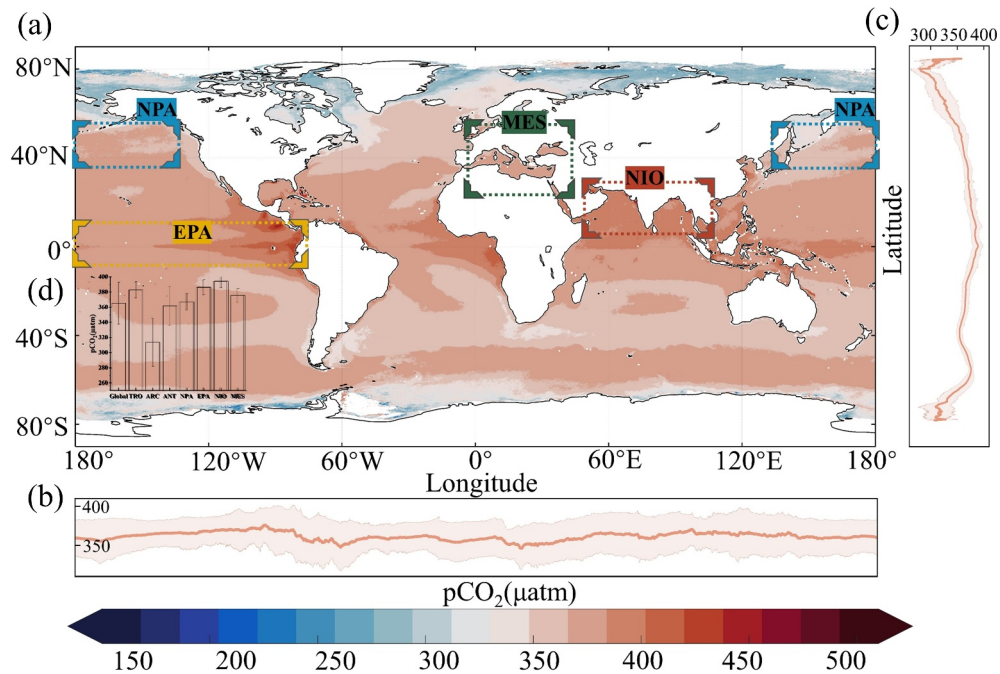

Figure 4. Spatial distribution of global $pCO_{2(sw)}$ from 2003 to 2022 (a). The longitudinal and latitudinal profiles' averaging $pCO_{2(sw)}$ are shown on (b) and (c). Statistics for $pCO_{2(sw)}$ in seven representative oceans, which are detailed in Table 1, are shown in (d).

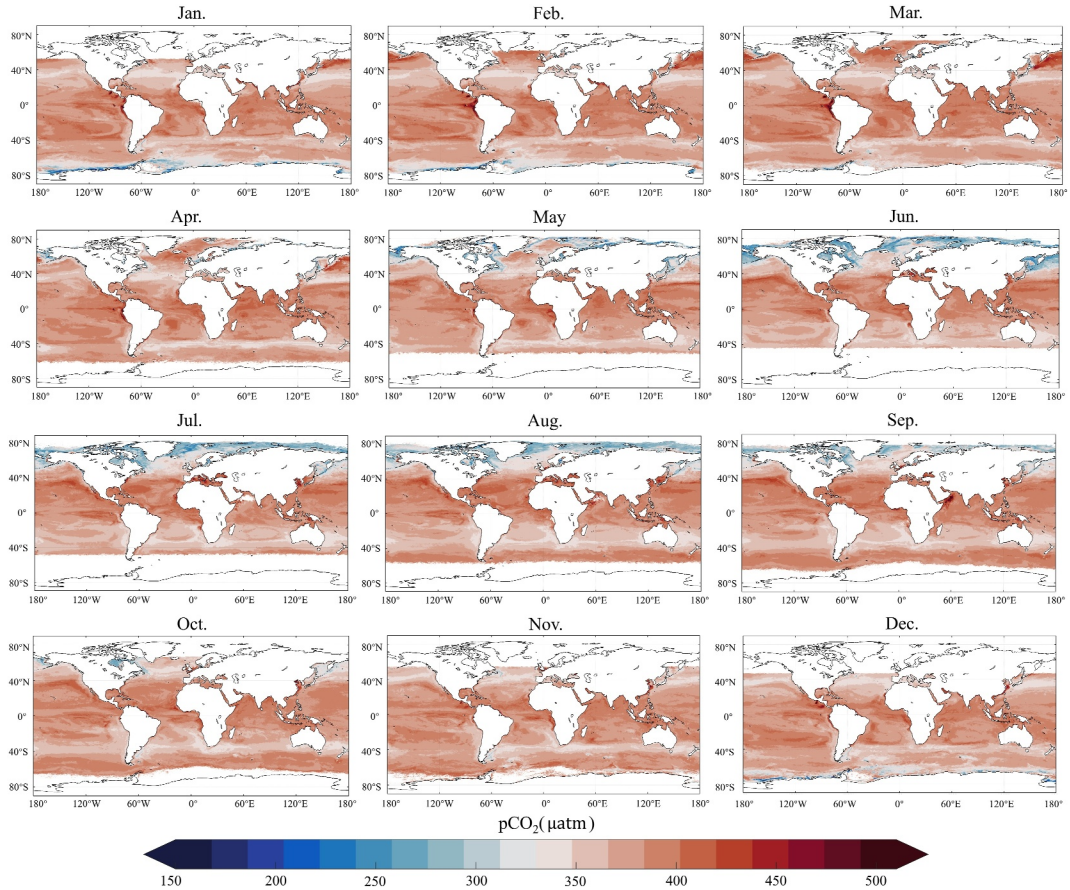

Figure 5. Monthly climatology $pCO_{2(sw)}$ from gcForest over the period from 2003 to 2022.

comparison to 30 years ago, the $pCO_{2(sw)}$ growth rate over the past 20 years has increased by 15%, highlighting an acceleration in recent decades. Notably, the $pCO_{2(sw)}$ growth rate rose from 1.02 ± 0.11 µatm yr^{-1} in the first decade (2003–2012) to 2.48 ± 0.13 µatm yr^{-1} in the second decade (2013–2022), exceeding the atmospheric annual pCO_2 growth rate of 2.21 ± 0.02 µatm yr^{-1}. This indicated the intensifying impact of anthropogenic CO_2 emissions on oceanic carbon dynamics.

In conclusion, the ARC and ANT regions consistently exhibited below-average $pCO_{2(sw)}$ values across all observational months. Conversely, the NIO, TRO, and EPA regions exhibited $pCO_{2(sw)}$ values that fluctuated

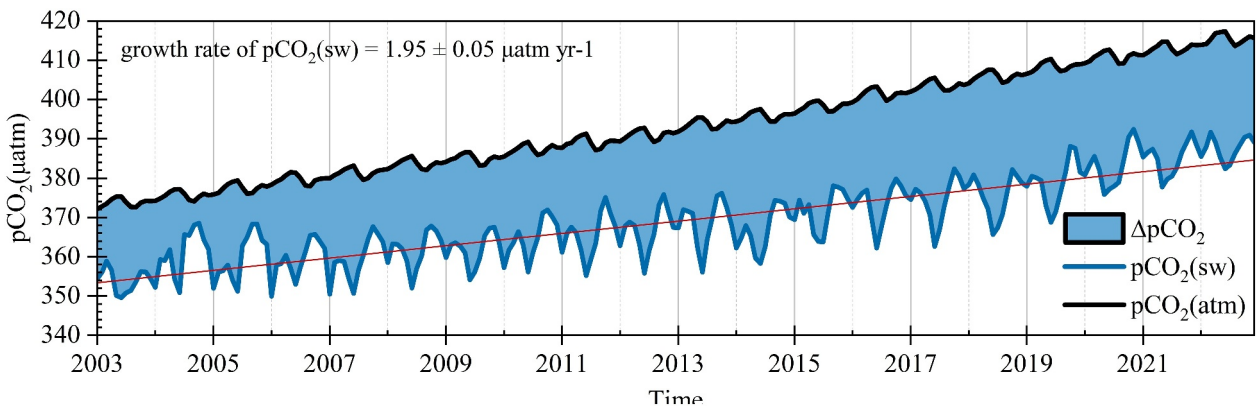

Figure 6. Time series of global $pCO_{2(sw)}$. The blue lines represent the monthly averaged $pCO_{2(sw)}$ predicted with gcForest. The black lines represent calculated monthly averaged $pCO_{2(atm)}$. And the red line indicates the mean growth rate estimated by linearly regressing with seasonal adjustment.

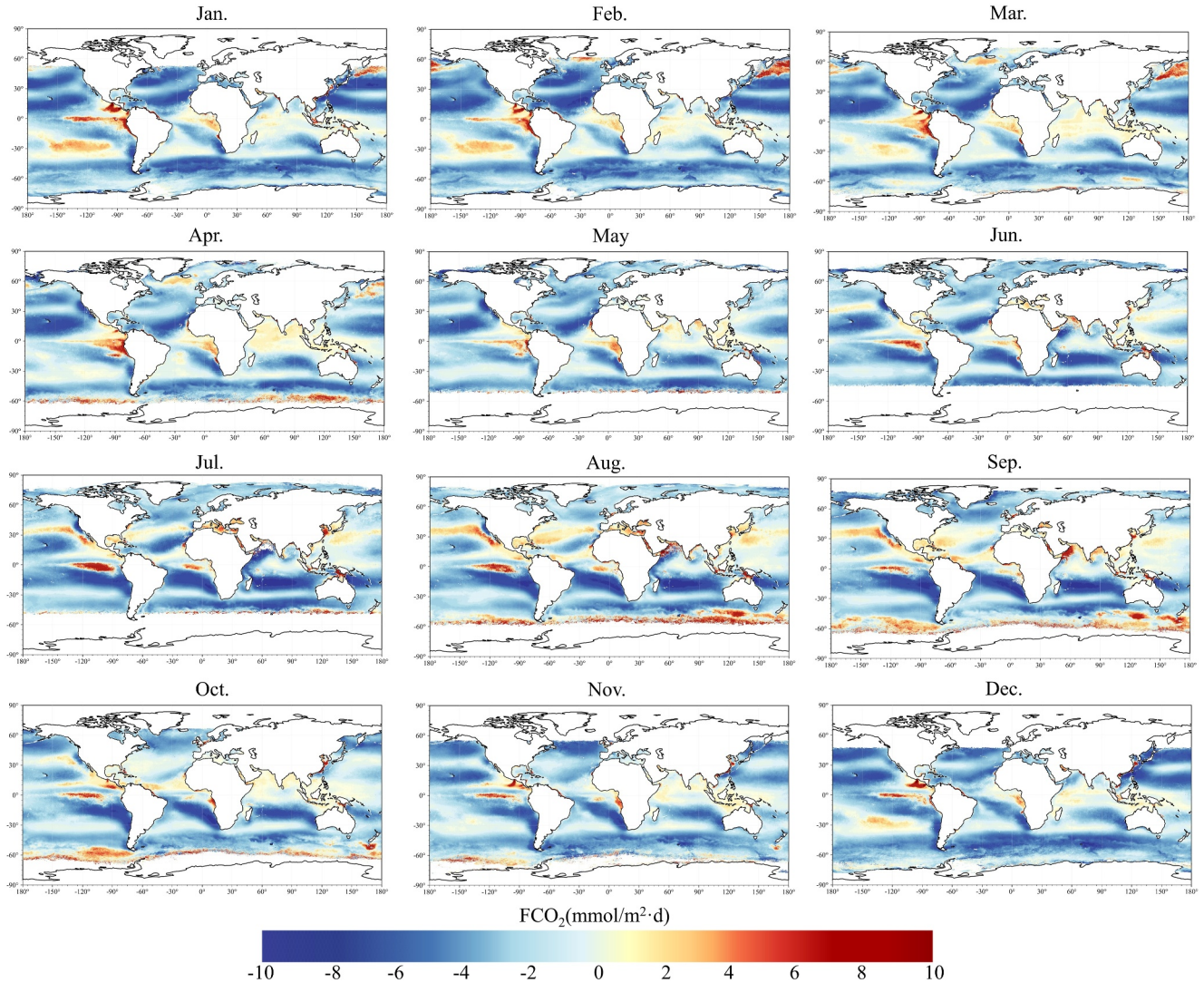

Figure 7. Climatological monthly spatial distribution of global air-sea CO_2 flux from 2003 to 2022. Blue indicates areas acting as carbon sinks, red represents carbon sources, and yellow represents areas where the average daily CO_2 flux is slightly below zero.

above the global average. Notably, the MES region uniquely exhibited several months where ΔpCO_2 significantly exceeded zero, indicating its role as a CO_2 source from June to September.

4.3. Spatial Distribution of Air-Sea CO_2 Flux

Using the gcForest-derived $pCO_{2(sw)}$ products, we revisited the global carbon uptake and outgassing capacity by calculating the air-sea CO_2 flux (FCO_2). Figure 7 presents the monthly spatial distribution of global FCO_2 in 2022. The majority of the global area is blue and yellow in color, indicating that the oceans act as a major carbon sink in the carbon cycle. As with $pCO_{2(sw)}$, the availability of FCO_2 data shifted from the Northern Hemisphere to the Southern Hemisphere from spring to winter, due to the characteristics of MODIS data. Between 40°N and 40°S, oceanic CO_2 sources are primarily located near coastlines, with sporadic occurrences in the open ocean throughout the year. In spring and summer, most of the ARC region acts as an oceanic carbon sink, except for the NPA region near Asia, which is outgassing CO_2 in early March. In fall and winter, a banded CO_2 source region is clearly observed in the ANT region near 60°S. Our results show that the MES region acts as a carbon source from June to September and as a carbon sink from October to May. The spatial distribution of FCO_2 does not fully align with that of $pCO_{2(sw)}$, because factors such as wind speed and SSS also affect the magnitude of air-sea CO_2 flux.

Table 2
The Performance of XGBoost, Random Forest, and gcForest in Training $pCO_{2(sw)}$

Input Parameter	Model	Training group			Test group			Model parameters
		r^2	RMSE	MAE	r^2	RMSE	MAE	
Adg, Chl-a, SST, and MLD, $\mathbf{R_{rs}(443)}$, $\mathbf{R_{rs}(488)}$, $\mathbf{R_{rs}(555)}$, $\mathbf{R_{rs}(667)}$, $\mathbf{R_{rs}(678)}$.	XGBoost	0.73	24.82	16.96	0.72	25.41	17.25	n_trees = 1,200, max_depth = 6, learning_rate = 0.05
	Random forest	0.96	8.42	5.06	0.81	19.45	12.44	n_trees = 100
	gcForest	0.98	6.82	4.00	0.85	17.35	10.70	n_trees = 100, optimal n_layer = 3
Adg, Chl-a, SST, and MLD, $\mathbf{R_{rs}(488)}$, $\mathbf{R_{rs}(555)}$.	XGBoost	0.72	25.63	17.28	0.70	26.11	17.57	n_trees = 1,200, max_depth = 6, learning_rate = 0.05
	Random forest	0.98	6.01	3.26	0.89	15.49	8.60	n_trees = 100
	gcForest	0.99	4.63	2.72	0.92	13.46	7.34	n_trees = 100, optimal n_layer = 3

5. Discussion

5.1. ML Model and Input Parameter Selection

We tested the XGBoost, random forest, and gcForest models with the same model inputs of Adg, Chl-a, SST, MLD, and R_{rs} including $R_{rs}(443)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, and $R_{rs}(678)$ based on the same training group, validation group, and testing group. Error statistics, including R^2, RMSE, and MAE, for estimating $pCO_{2(sw)}$ before input parameter selection are presented in Table 2. The gcForest model achieved the best performance with a testing RMSE of 17.35 μatm, an MAE of 10.70 μatm, and an R^2 of 0.85 μatm, followed by random forest. The gcForest model, benefiting from its robust cascade structure, demonstrated high effectiveness in accurately reconstructing global $pCO_{2(sw)}$ (Zhou & Feng, 2017). And it was subsequently trained to predict $pCO_{2(sw)}$ after input parameter selection.

Variable contribution reflects the extent to which environmental variables influence $pCO_{2(sw)}$. Figure 8a presents a matrix of correlation coefficients, indicating that higher absolute values of the coefficient denote stronger the correlation between variables. Adg, Chl-a, SST, and MLD were independent of each other, while there was a strong correlation between R_{rs} (443) and R_{rs} (488), as well as among R_{rs} (555), R_{rs} (667), and R_{rs} (678). Figure 8b showed that SST contributes the most to $pCO_{2(sw)}$ with 0.22, followed by Chl-a at 0.17 and MLD at 0.16. Among the five $R_{rs}(\lambda)$ variables, the green light band R_{rs} (555) had the highest contribution, at 0.09. In summary, SST is one of the most crucial factors in determining $pCO_{2(sw)}$, and the impact of the R_{rs} (555) is more significant than other spectral bands. Considering both variable contributions and correlations, R_{rs} (488) and R_{rs} (555) were selected to participate in the prediction of $pCO_{2(sw)}$. The performance of gcForest was better than that of XGBoost and random forest after excluding R_{rs} in three spectral bands (Table 2), and the results of gcForest was shown in Figure 3. Figure 8c shows that the ranking of environmental variable contributions did not change after the selection processes, with SST remaining the highest contributor.

5.2. Comparison With Other Climatological Products

This study presented a holistic approach to generate a monthly 4 × 4 km global $pCO_{2(sw)}$ distribution from 2003 to 2022 using the gcForest model. For comparison, we analyzed our gcForest product alongside three SOM-FFN products: the Zhong-FFN product (1° × 1°) (Zhong et al., 2022), MPI-FFN product (1° × 1°) (Jersild et al., 2023; Landschützer et al., 2016), and MOB-product (0.25° × 0.25°) (Chau et al., 2022; Denvil-Sommer et al., 2019). Additionally, the climatological gridded SOCAT observation product (1° × 1°) was included for performance benchmarking. Because the Zhong-FFN product only includes data available before 2021, we extracted data from the overlapping period (2003–2020) to ensure comparability among all products. To provide a more detailed evaluation of data set differences, the climatological monthly $pCO_{2(sw)}$ estimates from the four

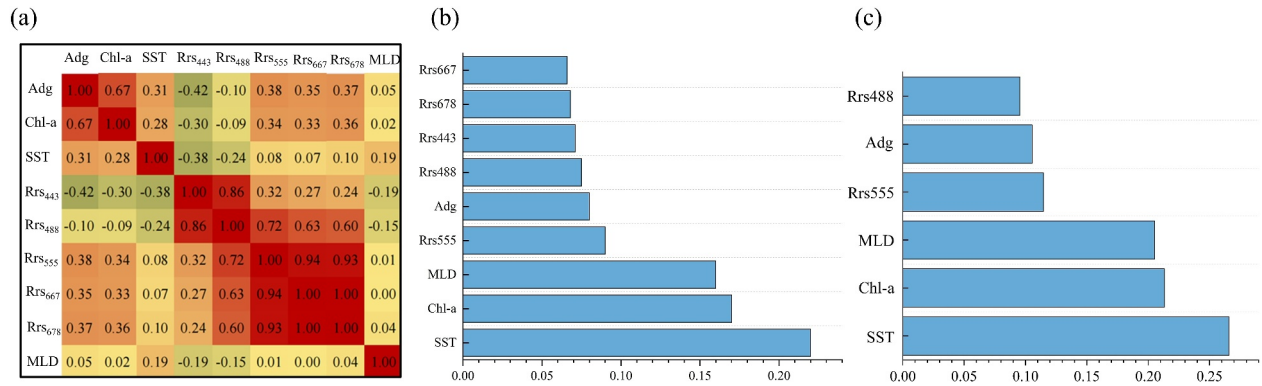

Figure 8. Correlation analysis of matchup pairs (a), and contributions of independent variables before (b) and after (c) input parameter selection.

reconstruction products were compared with gridded SOCAT observations in six representative regions (Table 1). Sudden spikes in $pCO_{2(sw)}$ values observed in February and September in the EPA region (Figure 9j), as well as in February in the MES region (Figure 9k), were attributed to measurement routes rather than seasonal variations. Additionally, model performance between the monthly $pCO_{2(sw)}$ estimates and SOCAT underway observations from 2003 to 2020 are presented in Figures S1–S6.

In terms of spatial variability, the gcForest product (Figure 9b) generally aligned more closely with the gridded SOCAT observations (Figure 9a) than three SOM-FFN products (Figures 9c–9e) in the Southern Ocean. In regions with sufficient observations, such as TRO and NPA, all four reconstruction products aligned well with the gridded SOCAT observations. However, in the ARC region, the gcForest product demonstrated better agreement with the gridded SOCAT observations, while the SOM-FFN product showed a notable tendency to overestimate $pCO_{2(sw)}$ values in summer and underestimate them in winter (Figure 9g). In the ANT region, the monthly variation of the gcForest product fluctuated around the gridded SOCAT, effectively capturing the seasonal variation in the observed data. However, during the Southern Hemisphere summer (December to March), all reconstruction products overestimated $pCO_{2(sw)}$, indicating persistent uncertainties in south of 60°S due to the limited spatial and temporal coverage of in situ observations. However, the improvement of gcForest product was not consistent across all regions. In the EPA region, the climatological gcForest product matched the gridded SOCAT observations in some months but showed a slightly underestimation. A similar pattern was observed in smaller oceanic regions like NPA and MES. In contrast, three SOM-FFN products tended to slightly overestimated $pCO_{2(sw)}$ values in these smaller regions, particularly in the EPA. The scatterplots (Figures S1-S6 in Supporting Information S1) illustrate that the higher resolution of the gcForest products allows for better capture of temporal and spatial variations in $pCO_{2(sw)}$ values observed in SOCAT data across various regions. However, in regions with high $pCO_{2(sw)}$ values like TRO, EPA, and MES, the spatial distribution was influenced by surrounding lower $pCO_{2(sw)}$ values due to the holistic nature of global reconstruction, resulting in underestimation. This limitation is inherent in the holistic reconstruction method. Conversely, zoning approaches that focus solely on internal variations within small areas neglect interactions with surrounding regions, often resulting in overestimations in high-value areas (e.g., Figure 9j) and underestimation in low-value areas (e.g., Figure 9g). The SOM-FFN products showed no consistent biases across regions when compared with the gridded SOCAT but generally captured consistent seasonal variations across different oceanic regions. This was partially attributed to differences in the zonal boundaries used for reconstruction, which inevitably impacted the regional $pCO_{2(sw)}$ estimates.

5.3. Drivers to the Variation of $pCO_{2(sw)}$

Distinct regional differences in $pCO_{2(sw)}$ trends, driven by geographic and natural characteristics, were evident (Takahashi et al., 2009). From Figure 8a, we infer that variations in ocean surface temperature and chlorophyll are the two primary drivers of elevated $pCO_{2(sw)}$ on a global scale, with MLD playing a crucial role in air-sea

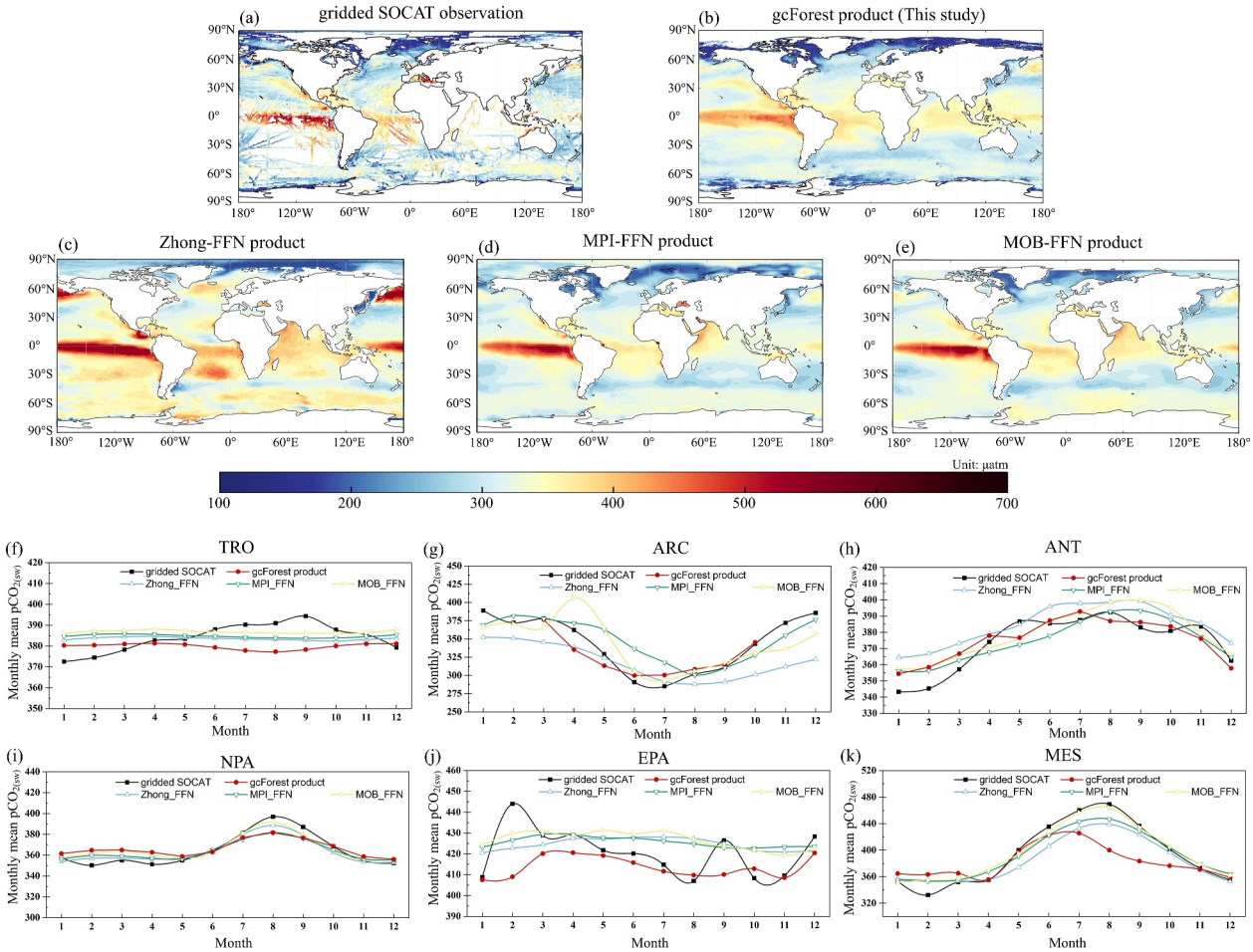

Figure 9. The climatological monthly $pCO_{2(sw)}$ distribution of gridded SOCAT observations, the gcForest product in this study, the Zhong-FFN product, the MPI-FFN product, and the MOB-FFN product are depicted in panels (a), (b), (c), (d), and (e), respectively. The seasonal variations and model performances of the five data sets across six representative regions are shown in panels (f)–(k).

interactions, ranking just behind SST and Chl-a in terms of impact. In the fall, global MLD values typically reached their maximum, suggesting that upward mixing greatly affected the $pCO_{2(sw)}$.

In the MES, the pattern of $pCO_{2(sw)}$ differed from other regions, as shown in Figure 10g; it increased during summer and decreased in winter. As inferred form Figure 10, this unique behavior was primarily driven by variations in SST and R_{rs} (488), indicating that the thermodynamic effects and phytoplankton influence dominated the changes in $pCO_{2(sw)}$ there. In both polar regions, $pCO_{2(sw)}$ rose with increasing SST, as high summer temperatures lead to sea ice melting, exposing $pCO_{2(sw)}$ primarily to thermodynamic effects. While $pCO_{2(sw)}$ generally decreased with increasing Chl-a, the substantial increase in summertime primary production in polar regions significantly contributed to the rise in $pCO_{2(sw)}$ (Tu et al., 2021). This effect is attributed to massive algal blooms at high Chl-a concentrations, which release CO_2 as the organic matter from these blooms is decomposed by bacteria (Sunda & Cai, 2012). The Chl-a trend observed in the polar region in Figure 10b was primarily due to missing MODIS observational data, resulting in regional averages that do not accurately represent the actual conditions. However, in the global ocean, especially in the EPA and MES, where Chl-a observations are sufficient, Chl-a variations aligned with changes in $pCO_{2(sw)}$. The significant contribution of Chl-a to $pCO_{2(sw)}$, as illustrated in Figure 8c, coupled with the negative correlation between Chl-a and $pCO_{2(sw)}$ depicted in Figures 10b and 10g, confirms that increasing Chl-a results in decreased $pCO_{2(sw)}$.

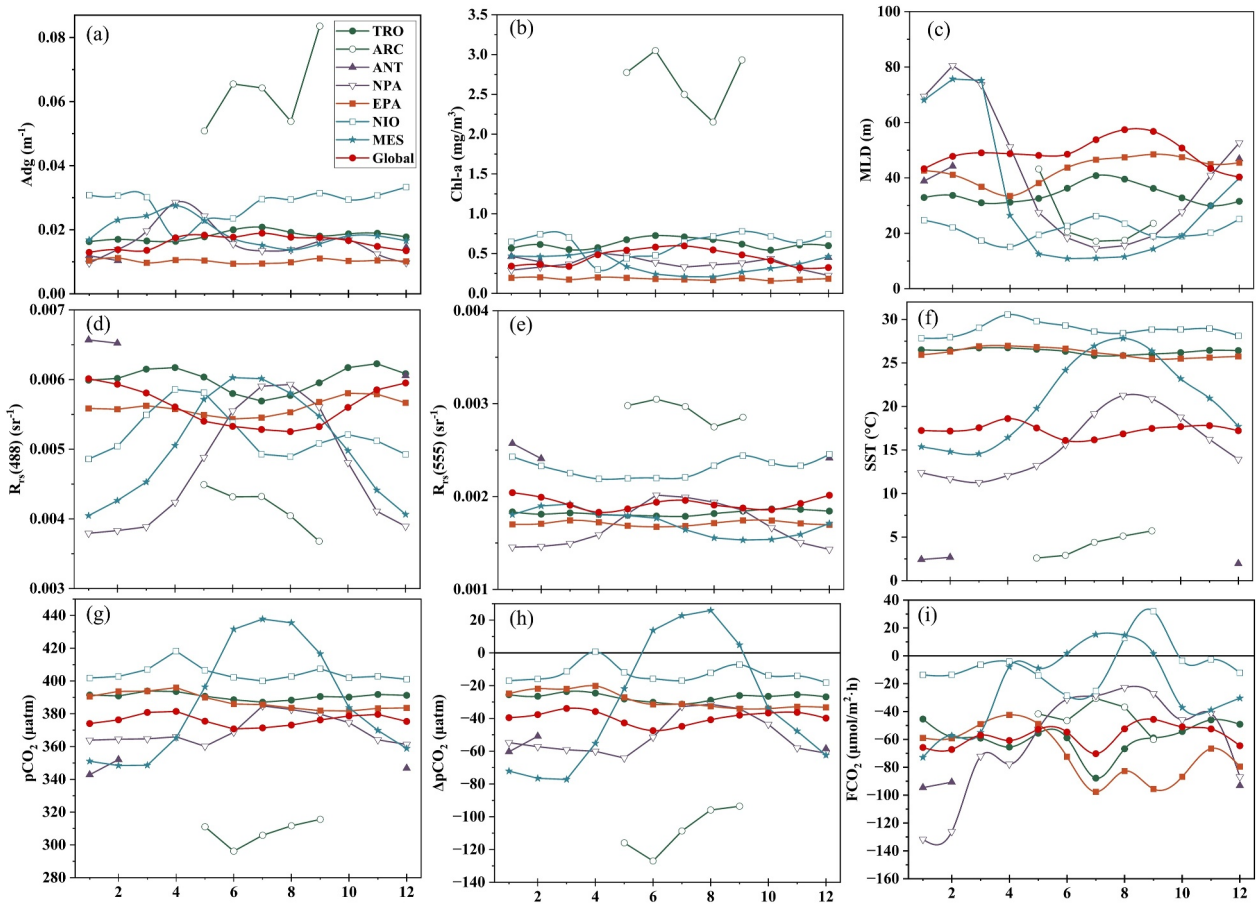

Figure 10. Seasonal changes of the mean value of (a) Adg, (b) Chl-a, (c) mixed layer depth, (d) $R_{rs}(488)$, (e) $R_{rs}(555)$, (f) sea surface temperature, (g) $pCO_{2(sw)}$, (h) ΔpCO_2, and (i) FCO_2 in global and 7 regions in 2022. The horizontal line in (h) and (i) indicates the direction CO_2 exchange. In all subgraphs, the solid red circle represents the change in the global oceans, and the rest of the conformity is detailed in the top right corner of (a).

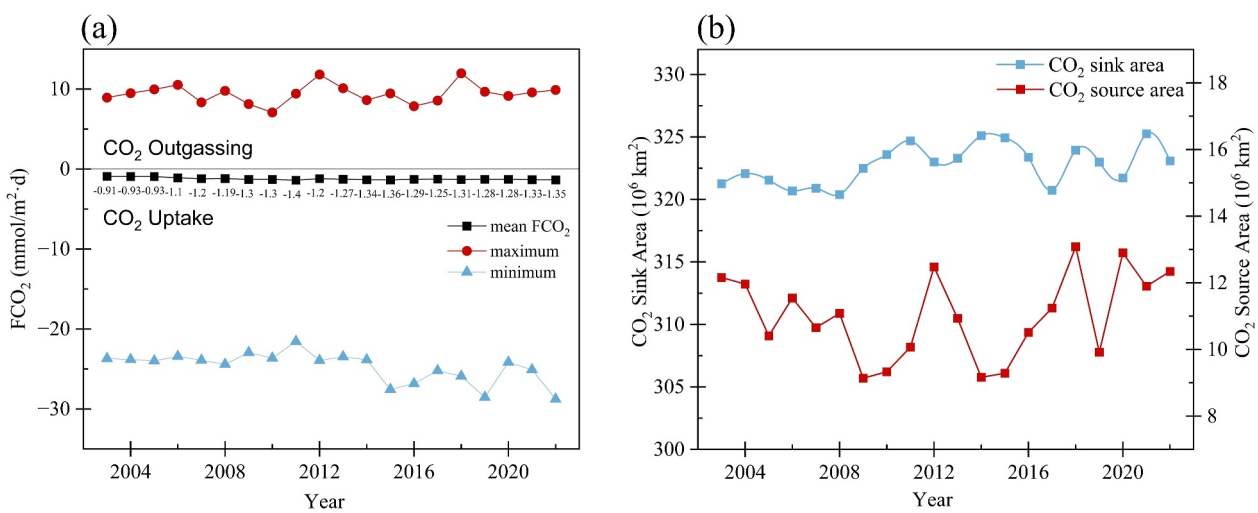

Figure 11. Annual changes in CO_2 flux (a) and CO_2 sink/source area (b). In panel (a), the maximum is shown in red at the top, the minimum in blue at the bottom, and the average in black in the middle. In panel (b), the blue line represents the CO_2 sink area, and the red line represents the CO_2 source area.

5.4. Global-Scale Long-Term Trends of CO_2 Source and Sink Flux

Quantifying air-sea CO_2 flux and analyzing the long-term changes in oceanic CO_2 uptake capacity are essential for assessing and predicting the potential impacts of the ocean carbon cycle. Detailed interannual variations in air-sea CO_2 fluxes are illustrated in Figure 11. In 2022, the average daily CO_2 flux was -1.35 mmol m^{-2} d^{-1}, a decrease of approximately 0.44 mmol m^{-2} d^{-1} from 2003 (-0.91 mmol m^{-2} d^{-1}). The maximum CO_2 source and sink fluxes reflect changes in the intensities of CO_2 sources and sinks (Zhang & Fischer, 2014). The maximum daily CO_2 source flux varied from 8.94 to 9.90 mmol m^{-2} d^{-1} during 2003–2022, averaging around 9.42 mmol m^{-2} d^{-1} over this period. However, the intensity of the CO_2 source flux has remained relatively stable, showing no significant increasing trend over the 20 years. Conversely, the maximum daily CO_2 sink flux increased by 5.08 mmol m^{-2} d^{-1}, at an annual rate of 0.20 mmol m^{-2} d^{-1} yr^{-1}. The maximum daily CO_2 sink flux was almost twice that of the maximum daily CO_2 source flux. These results indicate that the ocean's carbon uptake capacity exceeds its carbon emission capacity and has increased markedly during the study period.

Figure 11b showed no substantial trend in the CO_2 sink or source areas. The sink area varied from 320.68×10^6 to 325.25×10^6 km^2, with a deviation of 1.56×10^6 km^2. The source area ranged from 9.14×10^6 to 13.08×10^6 km^2, with a deviation of 1.27×10^6 km^2. Only minimal changes in area size were observed. In conclusion, the increase in carbon absorbed by the oceans was driven by enhanced carbon uptake capacity, not by an expansion of carbon sink areas.

6. Conclusions

This study primarily focuses on the distribution and variation of global ocean CO_2 flux over the last two decades through a holistic research approach. Using satellite observation data and the gcForest algorithm, we reconstruct global $pCO_{2(sw)}$ holistically to account for the interaction effects of environmental variables across regions. Five environmental variables affecting $pCO_{2(sw)}$ were selected to train the machine learning models: Adg for the dissolved organic matter effects, Chl-a for biological absorption effects, SST for thermodynamic effects, MLD for physical vertical mixing effects, and remote sensing reflectance (R_{rs}) to complement the optical information observable. After comparing three machine learning models, gcForest demonstrated superior performance in holistically reconstructing $pCO_{2(sw)}$. A 20-year global monthly $pCO_{2(sw)}$ product with a spatial resolution of 4×4 km was reconstructed, yielding an RMSE of 13.46 μatm. Finally, global air-sea CO_2 fluxes from 2003 to 2022 were calculated based on the $pCO_{2(sw)}$ data product. The spatiotemporal distribution patterns and trends of the reconstructed products are consistent with existing studies and products, confirming that the results of this study are reliable while improving spatial resolution and accuracy.

From 2003 to 2022, $pCO_{2(sw)}$ increased from 355.24 to 387.91 μatm, accompanied by a rise in CO_2 absorbed by the ocean. Given the variability in geographic and physical characteristics, $pCO_{2(sw)}$ within different marine regions is influenced by one or more environmental factors. Feature importance analysis indicated that SST and chlorophyll concentration were the primary drivers of global $pCO_{2(sw)}$ dynamics. Remarkably, the MES showed a seasonal variation in $pCO_{2(sw)}$, increasing in summer and decreasing in winter, likely influenced by SST and R_{rs} (488). A preliminary analysis of the relationship between CO_2 flux and CO_2 flux area was conducted. It was concluded that CO_2 uptake was gradually increasing, while the regional area of global CO_2 uptake remained stable.

Analyzing the relationship between regional $pCO_{2(sw)}$ and environmental variables underscores the need to understand the contributions of these variables to air-sea CO_2 flux. Compared to the zoning approach, the holistic approach provides a comprehensive perspective that effectively eliminates extreme $pCO_{2(sw)}$ values and generally align more closely with the gridded SOCAT observations in the Southern Ocean where measurements are scarce. However, we acknowledge the limitations of our study. In regions with relatively high $pCO_{2(sw)}$ values, the surrounding lower $pCO_{2(sw)}$ values in the overall reconstruction may lead to underestimation. Additionally, the $pCO_{2(sw)}$ retrieval approach based on machine learning methods relies heavily on the accuracy and distribution of underway observations as seen in existing studies. Given the sparse filed measurements in the NIO, significant uncertainty and potential bias may arise due to mutual influences around this region. Nonetheless, our holistic approach provides new insights into reconstructing global $pCO_{2(sw)}$. And we recommend using remote sensing reflectance to improve the accuracy of quantifying regional or global ocean fluxes.

Data Availability Statement

The underway observations of $pCO_{2(sw)}$ are available via http://www.socat.info (Bakker et al., 2023). The global ocean Adg, Chl-a, and SST estimates, based on MODIS satellite data, are obtained from the Ocean Productivity website (NASA Ocean Color), and the MLD data set is available online at Ocean Productivity: Input HYCOM MLD data (oregonstate.edu). Wind speed and sea level pressure data are accessible at the ERA5 reanalysis product (Copernicus Interactive Climate Atlas) (Hersbach et al., 2020). Sea surface salinity data from the ECCO2 is downloaded from the APDRC Datadoc | ECCO2 Cube92 model output (hawaii.edu) (Menemenlis et al., 2005) and xCO_2 is available at Global Monitoring Laboratory-Carbon Cycle Greenhouse Gases (noaa.gov) (Lan et al., 2023). Data are processed with https://github.com/kingfengji/gcForest.git and analyzed via MATLAB R2020a (MathWorks, 2020).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant 42171464), the Key Research and Development Project of Hubei Province (2022BCA057), supported by the Fundamental Research Funds for the Central Universities (2042023kf0216, 2042024kf0034, ZNJC202415, and 413000028), and a preresearch project (D040107), LIESMARS Special Research Funding. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Arrigo, K. R., & Van Dijken, G. L. (2007). Interannual variation in air-sea CO2 flux in the ross sea, Antarctica: A model analysis. *Journal of Geophysical Research*, *112*(C3), 16. https://doi.org/10.1029/2006jc003492

Bai, Y., Cai, W. J., He, X. Q., Zhai, W. D., Pan, D. L., Dai, M. H., & Yu, P. S. (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. *Journal of Geophysical Research-Oceans*, *120*(3), 2331–2349. https://doi.org/10.1002/2014jc010632

Bakker, D. C. E., Alin, S. R., Bates, N., Becker, M., Feely, R. A., Gkritzalis, T., et al. (2023). Surface Ocean CO2 Atlas database version 2023 (SOCATv2023) (NCEI accession 0278913). *NOAA National Centers for Environmental Information*. [Dataset]. https://doi.org/10.25921/R7XA-BT92

Caldeira, K., & Wickett, M. E. (2005). Ocean model predictions of chemistry changes from carbon dioxide emissions to the atmosphere and ocean. *Journal of Geophysical Research*, *110*(C9). https://doi.org/10.1029/2004jc002671

Chau, T. T. T., Gehlen, M., & Chevallier, F. (2022). A seamless ensemble-based reconstruction of surface ocean pCO2 and air–sea CO2 fluxes over the global coastal and open oceans. *Biogeosciences*, *19*(4), 1087–1109. https://doi.org/10.5194/bg-19-1087-2022

Chen, S. L., & Hu, C. M. (2017). Estimating sea surface salinity in the northern Gulf of Mexico from satellite ocean color measurements. *Remote Sensing of Environment*, *201*, 115–132. https://doi.org/10.1016/j.rse.2017.09.004

Chen, S. L., Hu, C. M., Barnes, B. B., Wanninkhof, R., Cai, W. J., Barbero, L., & Pierrot, D. (2019). A machine learning approach to estimate surface ocean pCO2 from satellite measurements. *Remote Sensing of Environment*, *228*, 203–226. https://doi.org/10.1016/j.rse.2019.04.019

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Clark, D. K. (1981). Phytoplankton pigment algorithms for the nimbus-7 CZCS. *Oceanography from Space*, 227–237. https://doi.org/10.1007/978-1-4613-3315-9_28

Denvil-Sommer, A., Gehlen, M., Vrac, M., & Mejia, C. (2019). LSCE-FFNN-v1: A two-step neural network model for the reconstruction of Surface Ocean pCO2 over the global ocean. *Geoscientific Model Development*, *12*(5), 2091–2105. https://doi.org/10.5194/gmd-12-2091-2019

Dickson, A. G., Sabine, C. L., & Christian, J. R. (Eds.) (2007)., *Guide to best practices for ocean CO2 measurements* (Vol. 3). PICES Special Publication.

D'Ortenzio, F., Antoine, D., & Marullo, S. (2008). Satellite-driven modeling of the upper ocean mixed layer and air-sea CO2 flux in the Mediterranean Sea. *Deep Sea Research Part I: Oceanographic Research Papers*, *55*(4), 405–434. https://doi.org/10.1016/j.dsr.2007.12.008

Fangohr, S., Woolf, D. K., Jeffery, C. D., & Robinson, I. S. (2008). Calculating long-term global air-sea flux of carbon dioxide using scatterometer, passive microwave, and model reanalysis wind data. *Journal of Geophysical Research*, *113*(C9). https://doi.org/10.1029/2005jc003376

Fennel, K., Wilkin, J., Previdi, M., & Najjar, R. (2008). Denitrification effects on air-sea CO2 flux in the coastal ocean: Simulations for the northwest North Atlantic. *Geophysical Research Letters*, *35*(24), 5. https://doi.org/10.1029/2008gl036147

Frankignoulle, M. (1988). Field-measurements of air sea CO2 exchange. *Limnology & Oceanography*, *33*(3), 313–322. https://doi.org/10.4319/lo.1988.33.3.0313

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Gregor, L., Hauck, J., et al. (2022). Global carbon budget 2022. *Earth System Science 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. *Journal of Geophysical Research*, *114*(C3), 12. https://doi.org/10.1029/2007jc004646

Gloege, L., Yan, M., Zheng, T., & McKinley, G. A. (2022). Improved quantification of ocean carbon uptake by using machine learning to merge global models and pCO(2) data. *Journal of Advances in Modeling Earth Systems*, *14*(2), 19. https://doi.org/10.1029/2021ms002620

Gruber, N., Bakker, D. C. E., DeVries, T., Gregor, L., Hauck, J., Landschuetzer, P., et al. (2023). Trends and variability in the ocean carbon sink. *Nature Reviews Earth and Environment*, *4*(2), 119–134. https://doi.org/10.1038/s43017-022-00381-x

He, J. Y., Chen, Y. J., Wu, J. P., Stow, D. A., & Christakos, G. (2020). Space-time chlorophyll-a retrieval in optically complex waters that accounts for remote sensing and modeling uncertainties and improves remote estimation accuracy. *Water Research*, *171*, 115403. https://doi.org/10.1016/j.watres.2019.115403

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hu, C. M., Lee, Z., & Franz, B. (2012). Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *Journal of Geophysical Research*, *117*(C1). https://doi.org/10.1029/2011jc007395

Ishii, M., Feely, R. A., Rodgers, K. B., Park, G. H., Wanninkhof, R., Sasano, D., et al. (2014). Air-sea CO2 flux in the pacific ocean for the period 1990-2009. *Biogeosciences*, *11*(3), 709–734. https://doi.org/10.5194/bg-11-709-2014

Jang, E., Kim, Y. J., Im, J., Park, Y.-G., & Sung, T. (2022). Global sea surface salinity via the synergistic use of SMAP satellite and HYCOM data based on machine learning. *Remote Sensing of Environment*, *273*, 112980. https://doi.org/10.1016/j.rse.2022.112980

Jersild, A., Landschützer, P., Gruber, N., & Bakker, D. C. E. (2023). An observation-based global monthly gridded sea surface pCO2 and air-sea CO2 flux product from 1982 onward and its monthly climatology (NCEI Accession 0160558). Index of /data/oceans/ncei/ocads/data/0160558/MPI_SOM-FFN_v2022.

Lan, X., Tans, P., Thoning, K., & Laboratory, N. (2023). Trends in globally-averaged CO2 determined from NOAA global monitoring laboratory measurements. NOAA GML. [Dataset]. https://doi.org/10.15138/9N0H-ZH07

Landschutzer, P., Gruber, N., & Bakker, D. C. E. (2016). Decadal variations and trends of the global ocean carbon sink. *Global Biogeochemical Cycles*, *30*(10), 1396–1417. https://doi.org/10.1002/2015gb005359

Landschützer, P., Gruber, N., Bakker, D. C. E., & Schuster, U. (2014). Recent variability of the global ocean carbon sink. *Global Biogeochemical Cycles*, *28*(9), 927–949. https://doi.org/10.1002/2014gb004853

Landschützer, P., Gruber, N., Bakker, D. C. E. (2018). Strengthening seasonal marine CO_2 variations due to increasing atmospheric CO_2. *Nature Climate Change*, *8*(2), 146–150. https://doi.org/10.1038/s41558-017-0057-x

Landschutzer, P., Gruber, N., Bakker, D. C. E., Schuster, U., Nakaoka, S., Payne, M. R., et al. (2013). A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink. *Biogeosciences*, *10*(11), 7793–7815. https://doi.org/10.5194/bg-10-7793-2013

Laruelle, G. G., Landschützer, P., Gruber, N., Tison, J. L., Delille, B., & Regnier, P. (2017). Global high-resolution monthly pCO2 climatology for the coastal ocean derived from neural network interpolation. *Biogeosciences*, *14*(19), 4545–4561. https://doi.org/10.5194/bg-14-4545-2017

Lefevre, N., Watson, A. J., & Watson, A. R. (2005). A comparison of multiple regression and neural network techniques for mapping in situ pCO (2) data. *Tellus Series B Chemical and Physical Meteorology*, *57*(5), 375–384. https://doi.org/10.1111/j.1600-0889.2005.00164.x

Mackay, N., & Watson, A. (2021). Winter Air-Sea CO2 fluxes constructed from summer observations of the polar Southern Ocean suggest weak outgassing. *Journal of Geophysical Research-Oceans*, *126*(5). https://doi.org/10.1029/2020jc016600

Marrec, P., Cariou, T., Mace, E., Morin, P., Salt, L. A., Vernet, M., et al. (2015). Dynamics of air-sea CO2 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

MathWorks. (2020). MATLABr2020a. *R2020a - Updates to the MATLAB and Simulink product families - MATLAB & Simulink*. [Software]. http://mathworks.com/products/new_products/release2022a.html

McClain, C., Hooker, S., Feldman, G., & Bontempi, P. (2006). Satellite data for ocean biology, Biogeochemistry,and climate research. Eos. *Transactions - American Geophysical Union*, *87*. https://doi.org/10.1029/2006eo340002

Menemenlis, D., Fukumori, I., & Lee, T. (2005). Using green's functions to calibrate an ocean general circulation model. *Monthly Weather Review*, *133*(5), 1224–1240. https://doi.org/10.1175/mwr2912.1

Mobley, C. D. (1999). Estimation of the remote-sensing reflectance from above-surface measurements. *Applied Optics*, *38*(36), 7442–7455. https://doi.org/10.1364/ao.38.007442

Olmedo, E., Gonzalez-Gambau, V., Turiel, A., Guimbard, S., Gonzalez-Haro, C., Martinez, J., et al. (2020). Toward an enhanced SMOS level-2 ocean salinity product. *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6434–6453. https://doi.org/10.1109/jstars.2020.3034432

Olmedo, E., Gonzalez-Haro, C., Hoareau, N., Umbert, M., Gonzalez-Gambau, V., Martinez, J., et al. (2021). Nine years of SMOS sea surface salinity global maps at the Barcelona Expert Center. *Earth System Science Data*, *13*(2), 857–888. https://doi.org/10.5194/essd-13-857-2021

O'Reilly, J. E., & Werdell, P. J. (2019). Chlorophyll algorithms for ocean color sensors-OC4, OC5 and OC6. *Remote Sensing of Environment*, *229*, 32–47. https://doi.org/10.1016/j.rse.2019.04.021

Pabi, S., & Arrigo, K. R. (2006). Satellite estimation of marine particulate organic carbon in waters dominated by different phytoplankton taxa. *Journal of Geophysical Research*, *111*(C9), 8. https://doi.org/10.1029/2005jc003137

Randerson, J. T., Lindsay, K., Munoz, E., Fu, W., Moore, J. K., Hoffman, F. M., et al. (2015). Multicentury changes in ocean and land contributions to the climate-carbon feedback. *Global Biogeochemical Cycles*, *29*(6), 744–759. https://doi.org/10.1002/2014gb005079

Reynolds, R. A., Stramski, D., & Mitchell, B. G. (2001). A chlorophyll-dependent semianalytical reflectance model derived from field measurements of absorption and backscattering coefficients within the Southern Ocean. *Journal of Geophysical Research*, *106*(C4), 7125–7138. https://doi.org/10.1029/1999jc000311

Rippeth, T. P., Lincoln, B. J., Kennedy, H. A., Palmer, M. R., Sharples, J., & Williams, C. A. J. (2014). Impact of vertical mixing on sea surface pCO2 in temperate seasonally stratified shelf seas. *Journal of Geophysical Research-Oceans*, *119*(6), 3868–3882. https://doi.org/10.1002/2014jc010089

Sharp, J. D., Fassbender, A. J., Carter, B. R., Lavin, P. D., & Sutton, A. J. (2022). A monthly surface pCO2 product for the California Current Large Marine Ecosystem. *Earth System Science Data*, *14*(4), 2081–2108. https://doi.org/10.5194/essd-14-2081-2022

Song, Z. G., Yu, S. J., Bai, Y., Guo, X. H., He, X. Q., Zhai, W. D., & Dai, M. H. (2023). Construction of a high spatiotemporal resolution dataset of satellite-derived pCO2 and Air-Sea CO2 flux in the south China sea (2003-2019). *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–15. https://doi.org/10.1109/tgrs.2023.3306389

Sunda, W. G., & Cai, W. J. (2012). Eutrophication induced CO2-acidification of subsurface coastal waters: Interactive effects of temperature, salinity, and atmospheric PCO2. *Environmental Science and Technology*, *46*(19), 10651–10659. https://doi.org/10.1021/es300626f

Takahashi, T., Olafsson, J., Goddard, J. G., Chipman, D. W., & Sutherland, S. C. (1993). Seasonal variation of CO2 and nutrients in the high-latitude surface oceans: A comparative study. *Global Biogeochemical Cycles*, *7*(4), 843–878. https://doi.org/10.1029/93gb02263

Takahashi, T., Sutherland, S. C., Chipman, D. W., Goddard, J. G., Ho, C., Newberger, T., et al. (2014). Climatological distributions of pH, pCO2, total CO2, alkalinity, and CaCO3 saturation in the global surface ocean, and temporal changes at selected locations. *Marine Chemistry*, *164*, 95–125. https://doi.org/10.1016/j.marchem.2014.06.004

Takahashi, T., Sutherland, S. C., Wanninkhof, R., Sweeney, C., Feely, R. A., Chipman, D. W., et al. (2009). Climatological mean and decadal change in surface ocean pCO(2), and net sea-air CO2 flux over the global oceans. *Deep Sea Research Part II: Topical Studies in Oceanography*, *56*(8–10), 554–577. https://doi.org/10.1016/j.dsr2.2008.12.009

Tu, Z., Le, C., Bai, Y., Jiang, Z., Wu, Y., Ouyang, Z., et al. (2021). Increase in CO2 uptake capacity in the arctic chukchi sea during summer revealed by satellite-based estimation. *Geophysical Research Letters*, *48*(15). https://doi.org/10.1029/2021gl093844

Wang, J., & Deng, Z. Q. (2018). Development of a MODIS data based algorithm for retrieving nearshore sea surface salinity along the northern Gulf of Mexico coast. *International Journal of Remote Sensing*, *39*(11), 3497–3511. https://doi.org/10.1080/01431161.2018.1445880

Wanninkhof, R. (2014). Relationship between wind speed and gas exchange over the ocean revisited. *Limnology and Oceanography: Methods*, *12*(6), 351–362. https://doi.org/10.4319/lom.2014.12.351

Weiss, R. F. (1974). Carbon dioxide in water and seawater: The solubility of a non-ideal gas. *Marine Chemistry*, *2*(3), 203–215. https://doi.org/10.1016/0304-4203(74)90015-2

Wu, H., Wang, L., Ling, X., Cui, L., Sun, R., & Jiang, N. (2024). Spatiotemporal reconstruction of global ocean surface pCO2 based on optimized random forest. *Science of the Total Environment*, *912*, 169209. https://doi.org/10.1016/j.scitotenv.2023.169209

Yu, S. J., Song, Z. G., Bai, Y., Guo, X. H., He, X. Q., Zhai, W. D., et al. (2023). Satellite-estimated air-sea CO2 fluxes in the bohai sea, yellow sea, and east China sea: Patterns and variations during 2003-2019. *Science of the Total Environment*, *904*, 166804. https://doi.org/10.1016/j.scitotenv.2023.166804

Zhang, J. Z., & Fischer, C. J. (2014). Carbon dynamics of Florida bay: Spatiotemporal patterns and biological control. *Environmental Science and Technology*, *48*(16), 9161–9169. https://doi.org/10.1021/es500510z

Zhong, G., Li, X., Song, J., Qu, B., Wang, F., Wang, Y., et al. (2022). Reconstruction of global surface ocean pCO2 using region-specific predictors based on a stepwise FFNN regression algorithm. *Biogeosciences*, *19*(3), 845–859. https://doi.org/10.5194/bg-19-845-2022

Zhou, Z. H., & Feng, J. (2017). *Deep forest: Towards an alternative to deep neural networks*. Paper presented at the 26th International Joint Conference on Artificial Intelligence (IJCAI). https://doi.org/10.48550/arXiv.1702.08835