Global Air-sea CO_2 Flux Inversion Based on Multi-Source Data Fusion and Machine Learning

Yongqiang Chen[1], Siyi Wang[1], Wenping He[1*]

[1]Scholl of Atmospheric Sciences, Sun Yat-sen University, Zhuhai, 519082, China

Correspondence to: Wenping He (wenping_he@163.com)

Abstract. The global gridded dataset of partial pressure of CO_2 (pCO_2) in the surface ocean and the associated air-sea CO_2 flux are crucial for studying climate change and global carbon cycle. However, the complex nonlinear dynamics of atmospheric and marine systems, along with limited observational data bring significant challenges to the inversion of these data. To address these challenges, a two-stage machine learning algorithm was developed. This algorithm incorporates a replacement method for missing ocean data by introducing ocean model simulations to fill these gaps and a machine learning model of dimensionality reduction-clustering-regression to manage system nonlinearity. By integrating in-situ observations, satellite observations and reanalysis datasets, this study reconstructs the global sea surface pCO_2 data at monthly 1° ×1° grid from 1993 to 2020, and then derives the corresponding air-sea CO_2 flux through the bulk flux formulation. The results demonstrate that the new inversion method can effectively capture the complex relationship between pCO_2 observations and other oceanic characteristics data in the surface ocean, allowing for extrapolation to global ocean regions. Compared to other data-based spatio-temporal interpolation methods, the global gridded dataset obtained in this study shows leading performance in terms of root mean square error (RMSE) and the coefficient of determination (R^2). Specifically, the average RMSE of the new dataset is reduced by approximately 42% and 45% in the Southern Ocean and Arctic Ocean regions comparing with the optimal results from other inversion datasets. Additionally, the new global pCO_2 dataset successfully reconstructs the time series close to the observations in coastal and coral reef regions, indicating that the machine learning algorithm can effectively reproduce the time variation characteristics of complex and highly heterogeneous waters. This study successfully applied a multi-source data fusion approach, offering an alternative solution to address the issue of missing ocean observational data, and providing a new perspective for the inversion research of oceanic carbon flux.

Keywords. pCO2, Air-sea CO_2 flux, Machine learning, Data fusion, Inversion

1 Introduction

The ocean plays a vital role in the global carbon cycle, with the air-sea CO_2 flux being one of the primary focuses of marine carbon cycle research (DeVries, 2022a). As a significant carbon sink within the earth system, the ocean absorbs approximately 25% of anthropogenic CO_2 emissions in the atmosphere annually (Friedlingstein et al., 2023; Intergovernmental Panel on Climate Change (IPCC), 2022, Chap. 5). Currently, the oceans have stored about 30% of the CO_2 emitted by human activities since the Industrial Revolution, playing a critical role in mitigating climate change (Khatiwala et al., 2013; Sabine et al., 2004). A comprehensive global dataset of sea surface partial pressure of CO_2 (pCO_2) is essential for studying oceanic carbon sinks and climate change. However, both the ocean and atmosphere operate as a highly complex and nonlinear dynamical system, characterized by phenomena such as stochastic process, phase shifts and abrupt changes (Liu et al., 2024; Liu Qun-Qun et al., 2015; Mei et al., 2024). Together, these nonlinear features make pCO_2 inversion especially challenging, as it involves analysing complex, non-stationary, and nonlinear data(He et al., 2021). Additionally, conducting extensive observations across the vast oceans poses significant challenges and incurs high economic costs. Therefore, the available oceanic observation data are sparse and unevenly distributed in time and space. The scarcity of observations further complicates the reconstruction of accurate global sea surface pCO_2 field, and imposes substantial technical difficulties in reconstructing complete ocean carbon sink dataset.

Currently, there are two main approaches for reconstructing the spatio-temporal field of sea surface pCO_2 and air-sea CO_2 flux: data-based spatio-temporal inversion methods, and numerical model assimilation methods. Data-based inversion methods primarily use observational data of sea surface pCO_2, combined with climate and reanalysis data products. By employing statistical models or machine learning techniques, the observational data is interpolated and extrapolated to generate comprehensive spatio-temporal fields of sea surface pCO_2(Fay et al., 2014; Friedrich and Oschlies, 2009; Jones et al., 2015; Landschützer et al., 2013; Ono et al., 2004). On this basis, these methods use the reconstructed sea surface pCO_2 data and the bulk flux formulation to calculate the global sea-air CO_2 flux (Wanninkhof, 2014). The empirical bulk flux formulation takes into account the concentration gradients of CO_2 between the surface seawater and the adjacent atmosphere, as well as the intensity of turbulent exchange processes between the sea and air interface. Using the reconstructed sea surface pCO_2 fields and satellite-retrieved sea surface temperature, wind speed, and other relevant data, the global sea-air CO_2 flux is calculated (Land et al., 2015).

However, the accuracy of results from this data-based inversion method largely depends on the quantity of observational data (Hauck et al., 2020). In some oceanic regions, due to the lack of sufficient training data, there may be significant epistemic uncertainty in the reconstruction results (Siddique et al., 2022). Another major defect of this method is that the proxy variables used to establish reconstruction relationship with sea surface pCO_2 observational data, especially those derived from satellite remote sensing, are often affected by missing data due to factors such as cloud cover and seasonal missing orbits, and so on. (IOCCG, 2000; Shutler et al., 2020, 2024). This problem is especially prominent in seasonally ice-covered areas such as the Southern Ocean and the Arctic Ocean, as well as in coastal regions. Therefore, most data-based inversion datasets have only realized the pseudo-global reconstruction of ocean carbon flux (Fay et al., 2021). This data gap further increases the uncertainty in estimating the global ocean carbon sink. To overcome this issue, the current inversion methods primarily adopt two approaches. One approach is to simply exclude regions with severe data gaps from the inversion range of the model. The other method is to use a temporary alternative method, which typically involve filling in the missing years or months of proxy variables with their corresponding monthly climatology calculated from existing data. In cases where the initial period of the inversion target time is missing data, due to the observational instruments were not yet established, the temporary alternative method usually backfills the missing data with its monthly climatological data superimposed with its linear trend. As for the periodic data gaps on large regional scale, such as missing chlorophyll concentrations in high-latitude regions during winter, the temporary alternative method often imputes them the minimum value of existing observational data from other times for the missing grid points or fills them with random white noise with an amplitude close to the actual values (Brewin et al., 2021; Fay et al., 2021; Gregor et al., 2019). Although these temporary substitution methods enable the inversion model to cover as large a spatial range as possible, they may result in distorted information being input into the inversion model, thereby increasing the error of reconstructed pCO_2. And if the spatial coverage of the inverted sea surface pCO_2 field is incomplete, the pCO_2 results need to be scaled proportionally when calculating the global sea-air carbon flux, which further increases the uncertainty of the carbon flux results (Friedlingstein et al., 2023; Hauck et al., 2020).

Numerical model assimilation methods refer to solving the ocean state by invoking numerical assimilation techniques into global ocean biogeochemical models (GOBMs)(Dowd et al., 2014). Based on initial conditions of a series of physical and biochemical parameters, combined with the constraints of meteorological data and satellite-derived data, GOBMs simulate the physical, chemical, and biological processes affecting sea surface CO_2 concentrations and the corresponding spatio-temporal variations of carbon cycle components within the ocean, thus constraining the results of surface pCO_2 and air-sea carbon flux (Carroll et al., 2020; Séférian et al., 2020; Wanninkhof et al., 2013). Simulated state variables within the model include marine carbon cycle components such as pCO_2, DIC (dissolved inorganic carbon), nutrients and their reservoirs at different depths in the ocean interior. The advantage of numerical model assimilation method is that it can be used in evaluating the change of marine carbonate system parameters and their roles in climate change or extreme climate events (Burger et al., 2020; Gruber et al., 2021; Hauri et al., 2013). The ensemble results of state-of-art models are often used to evaluate the trend of variables related to air-sea carbon cycle (Friedlingstein et al., 2023). However, some studies have shown that the simulation errors of numerical model method are considerably higher than those of data-based inversion methods (Gruber et al., 2009; Hauck et al., 2020; Verdy and Mazloff, 2017). Therefore, in order to accurately assess the response and feedback of global ocean carbon sinks to global climate change, it is necessary to combine the advantages of data interpolation-extrapolation methods and numerical model methods. By addressing their respective shortcomings, a comprehensive and reliable global sea-air CO_2 flux dataset with complete temporal and spatial coverage can be established.

Previous studies have widely applied the concept of data fusion to high-quality datasets inversion and data prediction in fields of oceanography, meteorology and atmospheric chemistry (Geng et al., 2021; Salcedo-Sanz et al., 2020; Sauzède et al., 2020; Vafaei et al., 2022; Wang et al., 2023). Data fusion schemes can integrate datasets from various sources, including conventional observations, satellite remote sensing data, reanalysis data, and three-dimensional model outputs. By effectively utilizing the useful information provided by different sources, reconstructed results can meet the requirements of constructing highly accurate datasets with full spatial coverage and long temporal spans. Therefore, the concept of multi-source data fusion can also be applied to the reconstruction of sea surface pCO_2 and sea-air CO_2 flux. The existing data-based spatiotemporal interpolation and extrapolation methods have not utilized the simulation results of GOBMs. However, GOBMs can simulate the spatiotemporal variations of physical and biogeochemical components within the ocean, which effectively compensate for the insufficiencies of ocean observational data. Therefore, by increasing data sources and integrating in-situ observational data, satellite remote sensing data and GOBMs model outputs, a more complete proxy variables dataset can be provided for the inversion model of air-sea carbon flux. This helps to establish datasets with comprehensive temporal and spatial coverage and better reconstruction results, thereby improving the data-based spatio-temporal interpolation and extrapolation methods in terms of data coverage and reconstruction accuracy.

In this context, this study attempts to introduce GOBMs simulation into the inversion of sea-air CO_2 fluxes, combining conventional in-situ observations, satellite remote sensing, and reanalysis data to construct a more complete proxy variables dataset. Based on machine learning methods, we explore new methods for reconstructing global sea surface pCO_2 and sea-air CO_2 fluxes. Furthermore, we systematically compare the data quality of this new approach with existing data products, evaluating the effectiveness of multi-source data fusion and machine learning methods in improving the performance of sea-air CO_2 flux inversions. The remaining sections of this paper will detail the data, methods and models used (Section 2), present and evaluation the inversion results (Section 3), and discuss the inversion results (Section 4).

2 Data and Methods

This study employs a two-stage inversion model based on machine learning methods to reconstruct the sea surface pCO_2 field used for air-sea CO_2 flux calculation. Firstly, using data from historical simulations from multiple model and marine elements that characterize the ocean's physical and chemical state, a machine learning model is used to filling the non-random missing values in the proxy variables dataset. Then, a three-step machine learning model involving dimensionality reduction, clustering, and regression is used to establish the nonlinear relationship between filled marine proxy variables and the sea surface pCO_2 observational data, and realize the reconstruction of a globally comprehensive sea surface pCO_2 field.

2.1 Data

Table 1 summarizes all the input data used in this study, including in situ observational data, satellite-based data, reanalysis data. The SeaFlux dataset (Gregor, 2023) is also used here to provide information for flux calculation and uniform comparison between our and other inversion datasets. The ocean surface CO_2 observational dataset is SOCAT v2022 (Bakker et al., 2022). SOCAT v2022 is a global ocean database that provides important data resources for studying the ocean carbon cycle. It collects approximately 33.7 million observations of the global oceans and coastal seas from 1957 to 2021, quality-controlled by over 100 international marine carbon research groups, with an observational accuracy better than 5 μatm. Since the direct observational variable of SOCAT is sea surface fugacity of CO_2, it can be converted to partial pressure of CO_2 according to formula (1).

$$pCO_2 = fCO_2 \times exp\,[-\frac{P_{atm}(B+2\delta)}{RT}], \quad (1)$$

Because of the sparse spatial and temporal distribution of observational data in global sea surface pCO_2 datasets including SOCAT, it is necessary to learn the relationship between proxy elements and in-situ observations to fill these gaps. Studies have shown that the increasing trend in ocean carbon sinks is mainly driven by anthropogenic CO_2 forcing in the atmosphere and regulated by natural variability within the ocean (DeVries, 2022b; DeVries et al., 2023; Rohr et al., 2023; Wanninkhof et al., 2013). Therefore, the proxy elements selected for this study are categorized into three groups based on the physical state of the ocean, the biochemical state of the ocean, and the exchange processes between the atmosphere and the ocean. The proxy elements include: sea surface temperature (SST), sea surface salinity (SSS), mixed layer depth (MLD), sea level anomaly (SLA), and sea surface eddy kinetic energy (EKE) that reflect the physical state of the ocean; chlorophyll-a concentration (Chl-a) and dissolved inorganic nutrients that reflect the biochemical state of the ocean surface and interior; and atmospheric CO_2 mixing ratio (xCO_2), 10m wind speed, and sea level pressure (SLP) that indicate the atmospheric state and air-sea exchange processes. SST, SSS, MLD, and Chl-a have been proven to have a close relationship with sea surface pCO_2 and are commonly used as feature variables in inversion products (Woolf et al., 2016; Yang et al., 2024). Sea surface eddy kinetic energy (EKE) refers to the energy of eddy movements in the ocean, which is mainly related to the eddy structures in water currents and plays a key role in the horizontal transport of heat and salinity. Also, studies have shown that incorporating EKE into the inversion of sea surface pCO_2 can improve the quality of inversion results (Gregor et al., 2019). And sea level anomaly (SLA) is used to represent the heat and mass transfer in the ocean. Additionally, dissolved inorganic nutrients (DIC) play a vital role in primary production and carbon uptake in marine ecosystem. Since nutrients are monthly climate data, they are extended to all months within the inversion range based on the mass conservation assumption and subjected to principal component analysis (PCA), using only the first two principal components as feature variables to reduce the model's complexity.

Features	Description	Dataset	Datatype	Resolution	Time span	Source
fCO_2	Fugacity of CO_2 in surface water	SOCAT version 2022	In-situ observations	Gridded to 1°, Monthly	1957 to 2021	https://socat.info/index.php/version-2022/
SST	Sea surface temperature	NOAA Optimum Interpolation (OI) SST V2	Reanalysis Data	0.25°, Monthly	1981 to date	https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.html
SSS	Sea surface salinity	EN.4.2.2	Reanalysis Data	1°，Monthly	1900 to date	https://www.metoffice.gov.uk/hadobs/en4/
MLD	Mixing layer depth	Global Ocean Ensemble Physics Reanalysis	Reanalysis Data	0.25°, Monthly	1993 to 2020	https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_ENS_001_031
SLA	Sea surface anomaly	CDS dataset	Satellite-based data	0.25°, Monthly	1993 to date	https://cds.climate.copernicus.eu/doi/10.24381/cds.4c328c78
EKE	Eddy kinetic energy	Copernicus-GlobCurrent	Reanalysis Data	0.25°, Monthly	1993 to 2022	https://data.marine.copernicus.eu/product/MULTIOBS_GLO_PHY_MYNRT_015_003
Ice	Ice concentration	NOAA Optimum Interpolation (OI) SST V2	Reanalysis Data	0.25°, Monthly	1981 to date	https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.html
Chl-a	Chl-a concentration	Copernicus-GlobColour	Satellite-based data	4 km, Monthly	1997 to date	https://data.marine.copernicus.eu/product/OCEANCOLOUR_GLO_BGC_L4_MY_009_104
DIC	phosphate, nitrate and silicate	World Ocean Altas 2018	Reanalysis Data	1°，Monthly climatology	-	https://www.ncei.noaa.gov/products/world-ocean-atlas
xCO_2	Atmospheric CO_2 mixing ratio	Jena CarboScope Atmospheric CO_2 Inversion	Inversion data	5°, 6 hourly	1993 to 2022	https://www.bgc-jena.mpg.de/CarboScope/
U_{10}	10m windspeed	ERA5	Reanalysis Data	0.25°, Monthly	1940 to date	https://cds.climate.copernicus.eu/doi/10.24381/cds.f17050d7
SLP	Sea level pressure	ERA5	Reanalysis Data	0.25°, Monthly	1940 to date	https://cds.climate.copernicus.eu/doi/10.24381/cds.f17050d7

Table 1 Summary of Datasets and Products Used for Machine Learning Model.

2.2 Methods

Figure 1 illustrates the data and workflow for reconstructing the sea surface partial pressure of carbon dioxide (pCO_2). The inversion workflow includes obtaining data from multiple sources (Fig1.step1), performing data preprocessing including features engineering and gap-filling chlorophyll-a concentration field utilizing model historical simulation data (Fig1.step2) and the finally performing the three-step pCO_2 field inversion algorithm involves dimensionality reduction, clustering, and regression (Fig1.step3,4,5).

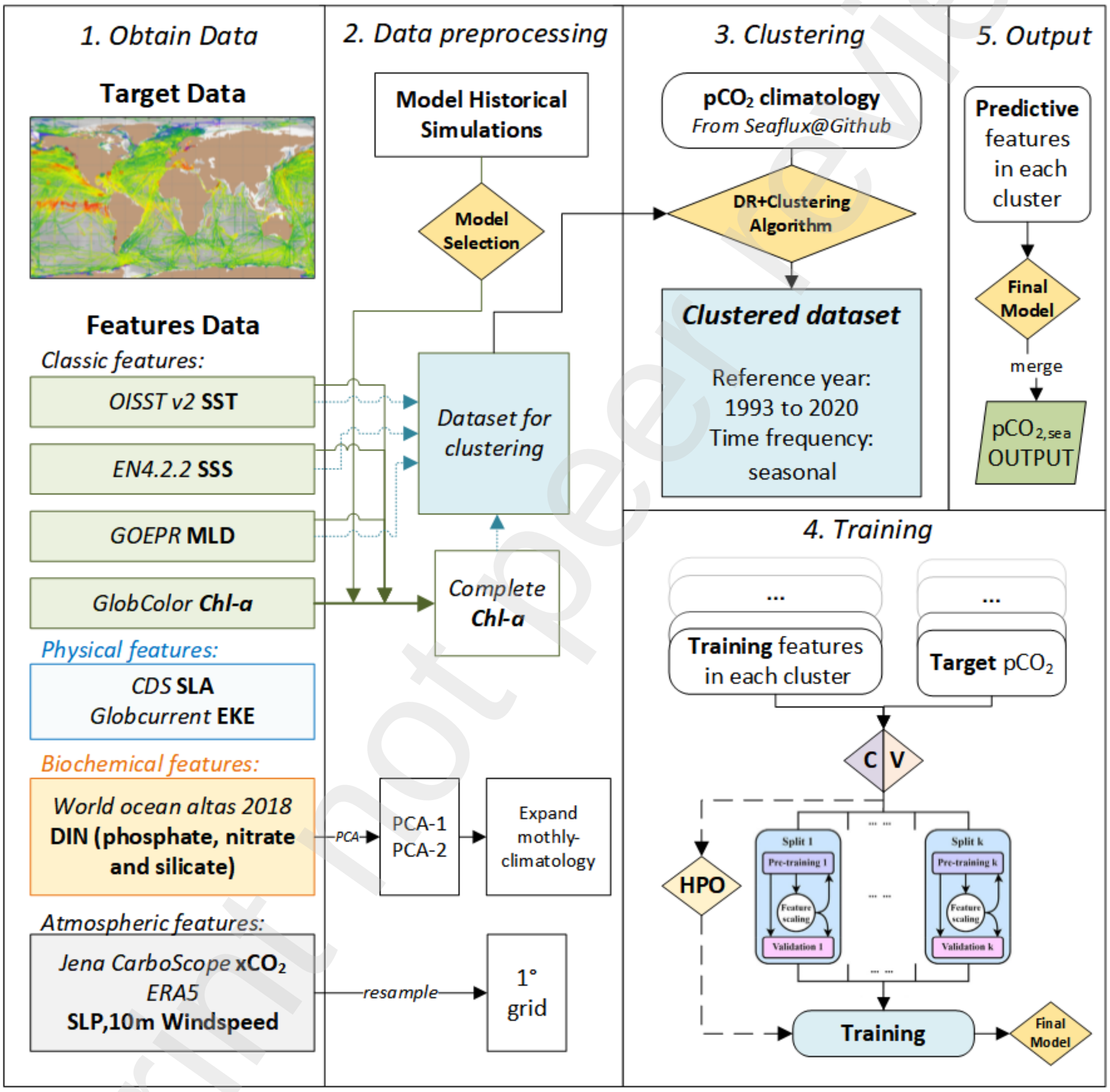


Figure 1 Flowchart of Sea Surface Partial Pressure of CO_2 Inversion

Data preprocessing mainly includes the following steps: first, using model historical simulations and other feature elements as machine learning inputs to fill in the missing areas in the Chl-a field; second, temporally extending the climatological data of dissolved inorganic nutrients and performing principal component analysis (PCA) to reduce dimensionality; third, applying a log10 transformation to the features of MLD, EKE, and Chl-a, which exhibit skewed distributions, and then standardizing all feature variables. During the data filling process, the historical simulations of sea surface chlorophyll concentration from 18 CMIP6 models were quantitatively evaluated to select 4 models with relatively good simulation performance: EC-Earth3-CC, IPSL-CM6A-LR-INCA, KIOST-ESM, and MIROC-ES2L (the evaluating result is later discussed in section 3.1). Then, using the model simulated chlorophyll concentration and the other variables of sea surface temperature, sea surface salinity, and mixed layer depth as mentioned in section 2.1, the relationship between these proxy variables and the satellite-derived GlobColour chlorophyll-a concentration was established using the XGBoost machine learning regression algorithm. Filled missing values of the GlobColour dataset mainly cover the Southern Ocean, Arctic Ocean, and some coastal areas. At the same time, in order to prevent the common issue of data leakage[1] in machine learning, this study independently preprocessed the training dataset with log10 transformation and other feature scaling processing, then applied the same preprocessing parameters to transform the validation and test sets, ensuring the isolation of information between the training data and other data in the training process.

Within the framework of machine learning, we propose a revised three-step algorithm consist of dimensionality reduction, clustering, and regression. This algorithm is capable of learning the complex relationship between observed pCO_2 and selected feature variables. Then，by applying the trained pCO_2 inversion model to a complete feature dataset, we can obtain a comprehensive and reconstructed global sea surface pCO_2 dataset，and can be used in subsequent calculations. The pCO_2 inversion algorithm draws inspiration from the clustering-regression approach used in inversion products such as CSIR-ML6 and SOM-FFN and optimizes upon it (Gregor et al., 2019; Landschutzer et al., 2016). Clustering is a commonly used unsupervised machine learning technique in oceanic researches (Solidoro et al., 2007). This study firstly subjects all the data to dimensionality reduction before the traditional clustering-regression approach. This is because applying clustering algorithms to high-dimensional data often encounters the so-called "curse of dimensionality," where the increase in data dimensions leads to sparse samples in high-dimensional space, severely impacting the performance of clustering algorithms. Moreover, traditional clustering algorithms such as K-means are not suitable for complex and non-linear Earth system data with non-Gaussian distribution (Sonnewald et al., 2020, 2019). Therefore, this study employs the t-distributed stochastic neighbor embedding (t-SNE) method for nonlinear dimensionality reduction. This method measures the similarity between data points using Gaussian probability distributions in high-dimensional space and t-distributions in low-dimensional space. By minimizing the difference in similarities between the high-dimensional and low-dimensional spaces, it effectively preserves local features among data points when projecting high-dimensional data into two- or three-dimension map, aiding subsequent clustering (Maaten and Hinton, 2008). This dimensionality reduction algorithm has been used in the partitioning of marine ecoregions and has shown that combining clustering algorithms with t-SNE yields better clustering results than using clustering algorithms alone in Earth system applications (Azeem et al., 2023; Balamurali et al., 2019). Based on the t-SNE dimensionality reduction combined with the spectral clustering algorithm, using sea surface pCO_2 data form SeaFlux, sea surface temperature form OISST V2, sea surface salinity form EN4.2.2 s, mixed layer depth from GOEPR, and chlorophyll concentration data supplemented by

[1] The unintentional leakage of data distribution characteristics and other information from the validation and test sets into the training set during the preprocessing stage, leading to reduced generalization performance of the model in machine learning.

model data, the global ocean can be divided into multiple regions with similar thermodynamic and biological characteristics.

After clustering, XGBoost regression algorithm is used for building inversion relationship within each region. XGBoost model is an optimized tree model based on Gradient Boosting Tree (GBT) with several advantages, including high precision, strong noise resistance, and parallel processing capabilities (Chen and Guestrin, 2016). XGBoost model uses efficient learning strategies to handle sparse features, making it effectively address missing data in remote sensing datasets. Additionally, the XGBoost model's high interpretability and training efficiency make it a powerful tool for feature learning and numerical regression. The regression model undergoes 10-fold cross-validation and Bayesian hyperparameter optimization (TPE) to optimize the regression model's hyperparameters, completing the inversion of global monthly 1°×1°sea surface pCO_2 field from 1993 to 2020.

2.3 Calculation of air-sea CO_2 flux

The carbon dioxide flux between the sea and the atmosphere interface is commonly calculated using the bulk flux formulation. This method primarily relies on the difference in partial pressure of CO_2 between the atmospheric boundary layer and the surface layer of seawater, with the influence of other factors represented parametrically. Here, the calculation from sea surface pCO_2 to air-sea CO_2 flux is performed by the pySeaFlux library (Gregor and Humphreys, 2021). This library provides the necessary auxiliary data and standard parameters required for the calculation, as well as a standard procedure for filling in the missing values in pCO_2 data products. As a result, differences in air-sea flux calculated by different inversion products are solely due to the differences in the sea surface pCO_2 they infer, thus enabling standardized comparisons of air-sea flux between different products. The formula for calculating air-sea CO_2 flux is as follows:

$$F\mathrm{CO}_2 = K_0 \cdot k_w \cdot (1 - ice) \cdot (p\mathrm{CO}_2^{sea} - p\mathrm{CO}_2^{atm}), \quad (2)$$

In Equation (2), K_0 represents the solubility of CO_2 in seawater calculated using the Weiss (1980) formulation. The other variables used for this calculation are salinity from EN4.2.2, sea surface temperature from OISSTv2, and sea level pressure from ERA5. K_w represents the gas transfer velocity, calculated using the Wanninkhof (1992) formulation:

$$k_w = a{\cdot}{U_{10}}^2{\cdot}Sc660, \quad (3)$$

Here, the gas transfer velocity parameter a is scaled to a global value of 16.5 cm/hr; U_{10} is the 10-meter wind speed, with data from the ERA5 wind speed product; $Sc660$ is the Schmidt number, a dimensionless number used to describe the mass transfer properties of CO_2 in seawater, calculated using sea surface temperature data from OISSTv2. ice represents the sea ice coverage. $p\mathrm{CO}_2^{sea}$ is the inversion result of partial pressure of CO_2 in sea surface; $p\mathrm{CO}_2^{atm}$ is the partial pressure of carbon dioxide in the marine boundary layer at atmosphere, derived from the dry air mole fraction of CO_2 product provided by ESRL. In this study, the flux is defined as positive when CO_2 is released from the ocean to the atmosphere and negative when the ocean uptakes CO_2 from the atmosphere.

Currently, there are two main methods to evaluate the uncertainty of gridded pCO_2: the standard deviation method and the error decomposition method. The error decomposition method divides uncertainty and error into three parts: instrumental measurement error, spatial variability, and error due to insufficient spatio-temporal sampling (Wang et al., 2014). Theoretically, the error decomposition method is more scientific, but its application is limited. The main reason is that, in addition to the monitoring data that need to be assessed, it also requires data with high spatial resolution and high spatial coverage, such as model simulations or remotely sensed pCO_2 data, to assess spatial variability. However, currently, model or remotely sensed pCO_2 data are still in the research stage, and no mature products are available, making the error decomposition method difficult to apply in practice. Therefore, this paper uses the standard deviation method to evaluate the uncertainties of $p\mathrm{CO}_2^{sea}$ and $F\mathrm{CO}_2$. Given the high certainty of K_w, sea ice coverage, and the Schmidt number, the uncertainty of the carbon dioxide flux is determined by the uncertainties of ΔpCO_2 and U_{10}. Through the derivation of error propagation, we obtain the standard deviation propagation formulation:

$$Uncertainty = |FCO_2| \times \sqrt{\left(\frac{2 \times \delta U_{10}}{U_{10}}\right)^2 + \left(\frac{\delta \Delta pCO_2}{\Delta pCO_2}\right)^2}, \tag{4}$$

2.4 Metrics for evaluation

In this study, we selected bias, Root Mean Squared Error (RMSE), latitude-weighted RMSE, and the coefficient of determination (R^2) as metrics to evaluate the accuracy of the inversion result. Bias is the mean of the errors between the model predictions and the target values (Eq.5). RMSE is the square root of the arithmetic mean of the squared residuals between the model predictions and the target values (Eq.6). Latitude-weighted RMSE is a widely used evaluation metric in geosciences and climate research. Compared to conventional RMSE, it replaces the weighting factor with the ratio of the cosine value of the grid point's latitude to the average latitude cosine value of all grid points, which better assesses the average estimation error when data distribution is wide (Eq.7). The coefficient of determination (R^2) measures the goodness of fit of the model and is determined by the ratio of the regression sum of squares to the total sum of squares (Eq.8). In the formulas for each evaluation metric, y_i represents the true value of the target variable; $\bar{y}$ represents the mean of the true values; and $\hat{y}_i$ represents the model predictions.

$$Bias = \sum_{i=0}^{n} \frac{\widehat{y_i} - y_i}{n}, \tag{5}$$

$$RMSE(\hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}, \tag{6}$$

$$latitude-weighted\ RMSE(\hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} w_i (\hat{y}_i - y_i)^2},\ w_i = \frac{\cos lat(i)}{\frac{1}{n} \sum_{i=1}^{n} \cos lat(i)}, \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{8}$$

3 Model building and training

3.1 Evaluation and Selection of GOBM Products

First, as introduced in Section 2.2, data form GOBMs was selected to fill the gaps in chlorophyll data. Sea surface chlorophyll concentration is one of the primary factors controlling the variability of sea surface partial pressure of CO_2 (pCO_2), reflecting the total amount of phytoplankton. It is also one of the factors with the most missing observation data globally. Within the scope of this study, the GlobColour chlorophyll-a inversion product, which combines data from multiple ocean color remote sensing satellites, covered only about 66% of the global ocean grid points. The ocean is a complex system with interrelated variables. To achieve a comprehensive inversion of global sea surface carbon flux, a complete set of feature variable fields is required. The coupled model intercomparison project (CMIP), organized by the World Climate Research Programme (WCRP), provides valuable data and platforms for climate-related research (Eyring et al., 2016). In CMIP6, coupled ocean models' simulations of marine biogeochemical cycles offer crucial support for studying climate change and ocean acidification. Thus, using model-simulated sea surface chlorophyll concentrations to fill in the gaps in observational data is a practical solution. To select the best model data to fill these gaps, this study first evaluated the historical simulation data of sea surface chlorophyll concentration from 18 CMIP6 models. All model and observation data were processed to match the temporal and spatial resolution of the inversion outputs, i.e., global monthly data at 1°×1° resolution.

Figure 2 presents the evaluation results of the sea surface chlorophyll concentration simulated by the 18 CMIP6 models. As is shown in Figure 2a, the remote sensing inversion data for sea surface chlorophyll-a concentration show significant spatial heterogeneity. High chlorophyll concentration areas are concentrated in regions of high primary productivity, such as temperate seas and coastal upwelling areas, while low concentrations are found in the subtropical gyres (Fig.2a). Figures 2b-s show the distribution of differences between each model's simulation and satellite inversion data. Most models exhibit small estimation biases, except for MPI-ESM-1-2-HAM, MPI-ESM1-2-LR, and MPI-ESM1-2-HR, which overestimate sea surface chlorophyll concentration (Figures 2n-p). Additionally, many models tend to underestimate chlorophyll concentration in coastal high-value areas in the remote sensing inversion data.

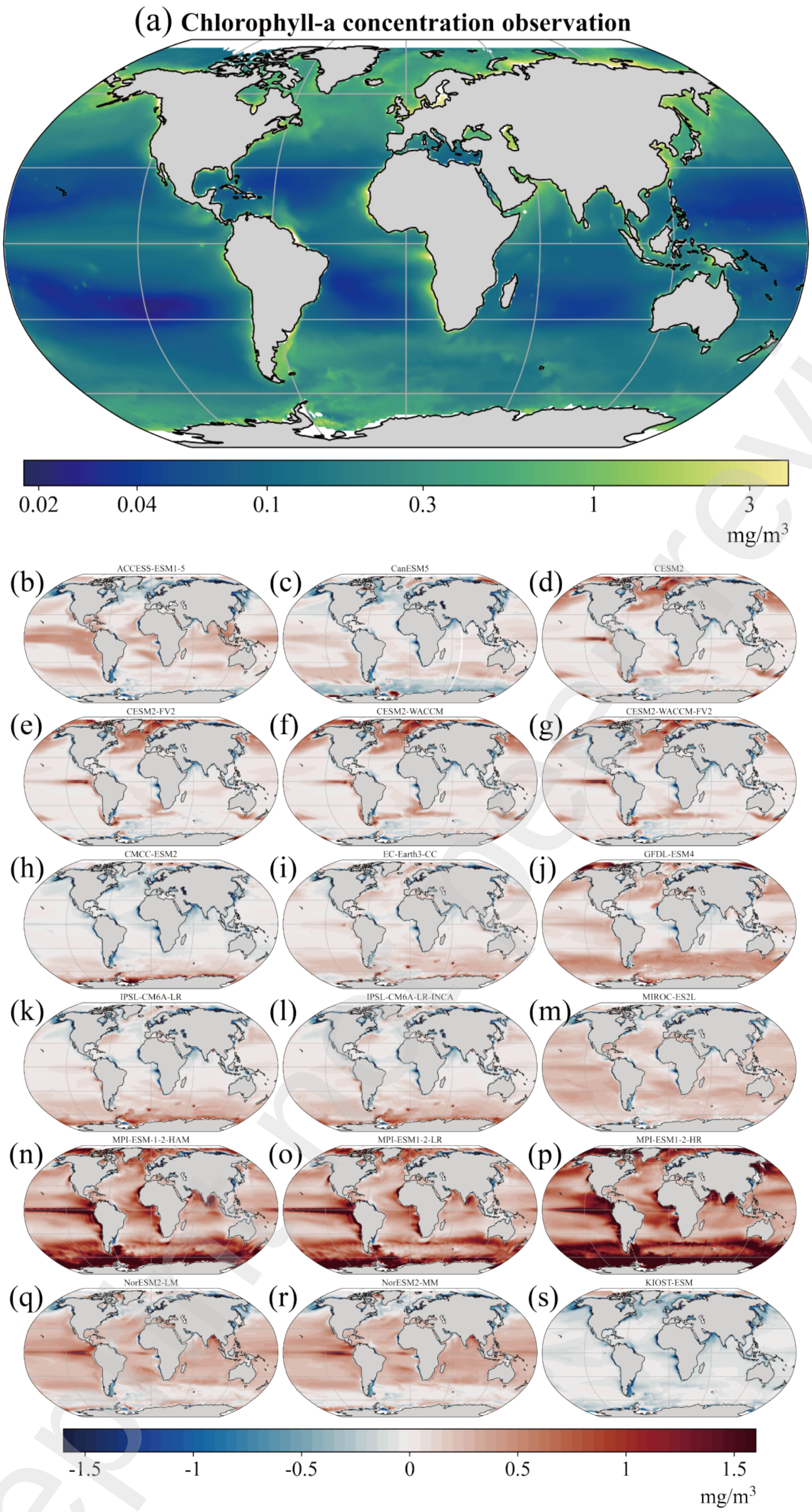


Figure 2 Evaluation of Modeled Sea Surface Chlorophyll Quality. (a) Average spatial field of GlobColour sea surface chlorophyll-a concentration product; (b) Average spatial field of the difference between ACCESS-ESM1-5 model sea surface chlorophyll concentration and GlobColour chlorophyll-a concentration product; (c) Same as (b) but for the CanESM5 model; (d) CESM2; (e) CESM2-FV2; (f) CESM2-WACCM; (g) CESM2-WACCM-FV2; (h) CMCC-ESM2; (i) EC-Earth3-CC; (j) GFDL-ESM4; (k) IPSL-CM6A-LR; (l) IPSL-CM6A-LR-INCA; (m) MIROC-ES2L; (n) MPI-ESM-1-2-HAM; (o) MPI-ESM1-2-LR; (p) MPI-ESM1-2-HR; (q) NorESM2-LM; (r) NorESM2-MM; (s) KIOST-ESM

Comparing model-estimated chlorophyll concentrations with satellite-derived chlorophyll-a concentrations remains a challenging task in ocean modeling. Therefore, this study uses three indicators for quantitative evaluation: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE), and the interannual variability index R_{iav} (Jönsson et al., 2023). R_{iav} quantifies the model's performance in simulating chlorophyll concentration on an interannual scale, with values closer to 0 indicating better simulation results (Rödenbeck et al., 2015). Based on the RMSE, MAE, and MBE evaluation results of the model-simulated sea surface chlorophyll concentration, it is evident that 15 of the 18 models have small simulation errors, except for MPI-ESM-1-2-HAM, MPI-ESM1-2-HR, and MPI-ESM1-2-LR (Table 2). Overall, the evaluation results show that EC-Earth3-CC, IPSL-CM6A-LR-INCA, KIOST-ESM, and MIROC-ES2L have better overall simulation performance compared to the other 14 CMIP6 models. Therefore, in the subsequent experiments of this study, the simulation data from these four models were used to fill the gaps in chlorophyll observations. These data, combined with sea surface temperature, sea surface salinity, and mixed layer depth as feature variables, were used to construct the relationship between these variables and satellite-derived GlobColour chlorophyll concentrations using the XGBoost machine learning regression algorithm. This approach filled the GlobColour data gaps to generate a complete feature variable dataset.

Model	**RMSE**	**MAE**	**MBE**	**R^{IAV}**
ACCESS-ESM1-5	0.64	0.28	0.01	0.9
CanESM5	0.74	0.24	-0.07	1.05
CESM2	0.72	0.25	0.03	1.32
CESM2-FV2	0.7	0.24	0.02	1.13
CESM2-WACCM	0.72	0.25	0.04	1.13
CESM2-WACCM-FV2	0.71	0.25	0.03	1.18
CMCC-ESM2	0.7	0.2	-0.09	0.82
EC-Earth3-CC	0.63	0.21	0	0.77
GFDL-ESM4	0.67	0.29	0.11	0.88
IPSL-CM6A-LR	0.68	0.23	0	0.63
IPSL-CM6A-LR-INCA	0.66	0.22	0	0.59
KIOST-ESM	0.62	0.19	-0.18	1.07
MIROC-ES2L	0.55	0.27	0.1	0.77
MPI-ESM-1-2-HAM	1.57	0.7	0.53	8.05
MPI-ESM1-2-HR	2.04	0.94	0.8	12.06
MPI-ESM1-2-LR	1.45	0.63	0.45	6.28
NorESM2-LM	0.7	0.36	0.14	0.89
NorESM2-MM	0.69	0.36	0.13	1.09

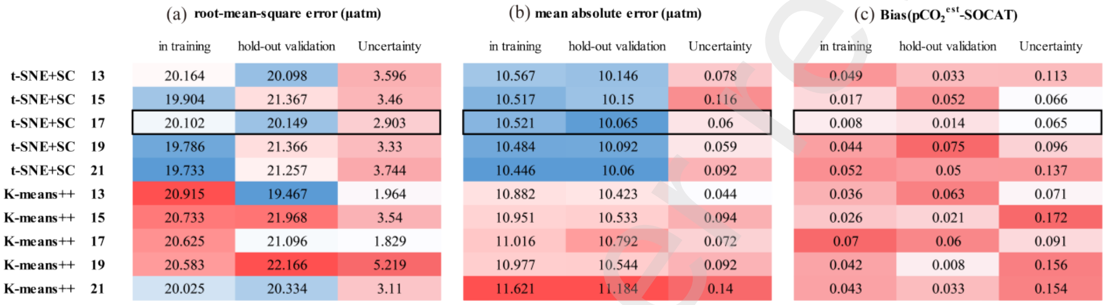
Table 2 Evaluation Results of Modeled Sea Surface Chlorophyll Concentration

3.2 Training of pCO_2 Inversion model

After constructing a complete feature variables dataset, this study applied a three-step algorithm of dimension reduction, clustering, and regression to learn the complex relationships between pCO_2 observation data and feature variables, thus training the pCO_2 inversion model. In this stage, sensitivity experiments were conducted to test different cluster counts to determine the optimal parameter. Since clustering algorithms are unsupervised and lack post-hoc validation, traditional metrics such as silhouette coefficient and Davies-Bouldin index have limitations when applied to complex data. Therefore, multiple cluster numbers and different clustering algorithms were used for clustering and regression calculations. The final choice of the cluster number parameter was based on the inversion quality of the pCO_2 field models obtained from different cluster numbers.

For each clustering configuration, a 10-fold cross-validation combined with a hyperparameter optimization algorithm was used to train the regression model's hyperparameters. In 10-fold cross-validation, the dataset is randomly divided into ten parts, with nine parts used as the training set and one as the test set in rotation. Based on this, the Tree-structured Parzen Estimators (TPE) Bayesian hyperparameter optimization algorithm was used for 200 iterations, using the average error from cross-validation as the loss function to find the global optimal solution of different hyperparameter combinations. During the experiments, the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Bias were used to evaluate the uncertainty of the machine learning model's inversion (Figure 3). Lower values of these indicators indicate higher limit of the model's inversion capability. To more accurately assess model performance, the holdout method was further employed on training dataset with different clustering configurations and above-mentioned trained hyperparameters. Specifically, the original training dataset was randomly split into 90% training data and 10% validation data, repeated ten times. After each random sampling, the model with pre-trained hyperparameters was used to learn the training data and evaluated on the corresponding validation data. The average of these ten validation results was taken as the final validation result, while the standard deviation of these ten results was calculated to indicate the uncertainty of model performance. When the holdout method's error is similar to the optimal training result, it indicates low uncertainty in the model's generalization process.

Figure 3 shows the comparison of different clustering schemes and cluster numbers. It can be observed that during training, the t-SNE dimension reduction combined with the spectral clustering algorithm used in this study performed slightly better than the k-means++ clustering algorithm. Specifically, after same procedure of hyperparameters optimization, the RMSE and MAE of the five models with cluster numbers 13 to 21 were lower than those of the corresponding k-means++ results (Figures 3a and 3b). This indicates that dimension reduction before clustering benefits subsequent pCO_2 field calculations. Compared to k-means++, the t-SNE combined with spectral clustering algorithm produced smoother cluster boundaries and fewer instances of intermixing at cluster borders. In addition, both clustering algorithms showed a positive bias in inversion results, suggesting a tendency to overestimate sea surface pCO_2 (Figure 3c). Comparing the training errors and the average validation errors from holdout method revealed that the t-SNE non-linear dimension reduction combined with the spectral clustering algorithm had validation errors close to training errors, indicating good generalization performance. In contrast, the inversion results based on k-means++ clustering were more sensitive to different cluster numbers, with larger variations in different sampling test (Figure 3a). These results suggest that the t-SNE non-linear dimension reduction combined with spectral clustering improves the inversion capability of subsequent machine learning models, producing results with smaller errors and less dependence on the cluster number. Finally, comparing the configurations showed that using the t-SNE non-linear dimension reduction algorithm with 17 clusters yielded the smallest bias and lowest uncertainty among the ten holdout validations, indicating high robustness. Therefore, in the subsequent study, the machine learning model using t-SNE combined with spectral clustering and 17 clusters will be trained on the entire dataset, and this TSSCXG-17 inversion results will be used for evaluation and carbon flux calculation, ultimately producing the global monthly 1°×1° sea surface pCO_2 field inversion results from 1993 to 2020.



		(a) root-mean-square error (µatm)			(b) mean absolute error (µatm)			(c) Bias(pCO_2^{est}-SOCAT)		
		in training	hold-out validation	Uncertainty	in training	hold-out validation	Uncertainty	in training	hold-out validation	Uncertainty
t-SNE+SC	13	20.164	20.098	3.596	10.567	10.146	0.078	0.049	0.033	0.113
t-SNE+SC	15	19.904	21.367	3.46	10.517	10.15	0.116	0.017	0.052	0.066
t-SNE+SC	17	20.102	20.149	2.903	10.521	10.065	0.06	0.008	0.014	0.065
t-SNE+SC	19	19.786	21.366	3.33	10.484	10.092	0.059	0.044	0.075	0.096
t-SNE+SC	21	19.733	21.257	3.744	10.446	10.06	0.092	0.052	0.05	0.137
K-means++	13	20.915	19.467	1.964	10.882	10.423	0.044	0.036	0.063	0.071
K-means++	15	20.733	21.968	3.54	10.951	10.533	0.094	0.026	0.021	0.172
K-means++	17	20.625	21.096	1.829	11.016	10.792	0.072	0.07	0.06	0.091
K-means++	19	20.583	22.166	5.219	10.977	10.544	0.092	0.042	0.008	0.156
K-means++	21	20.025	20.334	3.11	11.621	11.184	0.14	0.043	0.033	0.154

Figure 3 Heatmap Showing Average (a) RMSE, (b) MAE, and (c) Bias Under Different Clustering Configurations. Clustering configurations consist of two algorithms (t-SNE combined with spectral clustering and K-means++) and 13 to 21 clusters. Rows represent the number of clusters, and columns represent training error and the average error and uncertainty of 10 hold-out validations.

Overall, the sea surface pCO_2 reconstructed in this study fits the SOCAT dataset used for training well and accurately reproduces the widely recognized spatial distribution pattern of pCO_2 (Takahashi et al., 2009). The spatial distribution of model estimation residuals is shown in Figure 4. The temporal-averaged estimation residuals are generally small, and their spatial distribution somewhat reflects the spatial characteristics of different clusters. Additionally, in open ocean areas, the model's estimation error is relatively small, with larger errors concentrated in the equatorial regions, especially the eastern equatorial Pacific and equatorial Atlantic, and also in the Southern Ocean and Antarctic coastal areas (Figure 4a). The inversion-observation plot shows some overestimation of low values and underestimation of high values by the machine learning model (Figure 4b). Overall, the dimension reduction-clustering-regression approach and XGBoost machine learning regression algorithm effectively construct the non-linear relationship between feature variables and observations. The reconstructed pCO_2 field shows minor differences from the learned SOCAT pCO_2 observation field, with an overall RMSE of only 5.302 µatm and a correlation coefficient of 0.993, demonstrating the machine learning model's capability to learn the relationship between sea surface pCO_2 and atmospheric and oceanic feature variables.

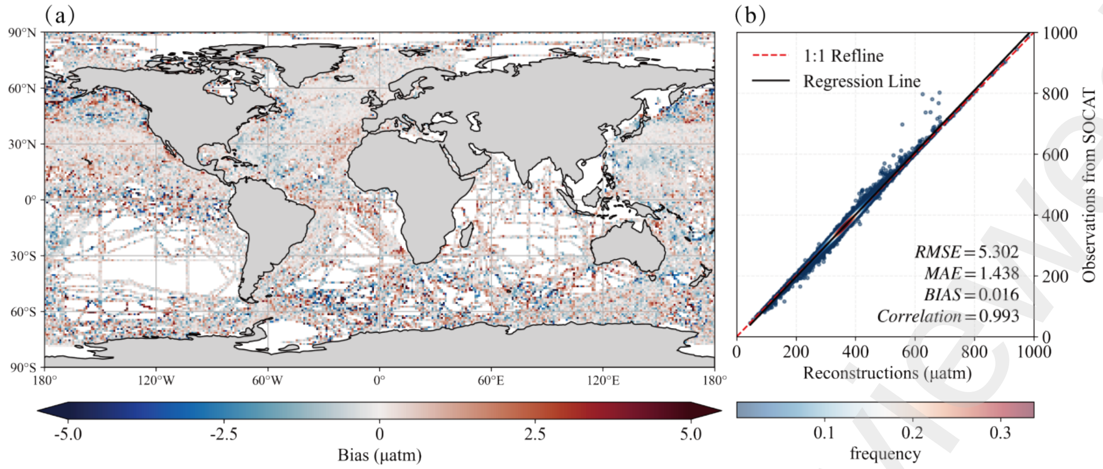


Figure 4 Grid Bias (a) and Scatter Density Plot (b) of TSSCXG-17 pCO_2 Inversion Values

4 Results and Evaluation

4.1 Evaluation with Independent Ocean Products

To evaluate the quality of the sea surface pCO_2 inversion, it is necessary to objectively evaluate the inversion results with observation data independent of SOCAT data set. Therefore, the LDEO (Lamont-Doherty Earth Observatory, v2019 version) dataset was used to assess the spatio-temporal reconstruction of sea surface pCO_2 in this study (Takahashi et al., 2017). The 2019 version of the LDEO dataset collected approximately 14.2 million global sea surface pCO_2 observations, covering the period from 1957 to 2019 and including both open ocean and coastal areas. The LDEO dataset, which has undergone quality control, provides monthly 1°×1° gridded pCO_2 observational data with an uncertainty of ±2.5 μatm. To validate the effectiveness of our proposed three-step reconstruction algorithm of dimension reduction-clustering-regression, we further compared the results with other data-based inversion datasets, which were all trained using SOCAT data. Additionally, to assess the impact of completing the chlorophyll feature variable field on inversion results, the global ocean was divided into seven regions: Southern Ocean, Arctic Ocean, North Pacific, South Pacific, North Atlantic, South Atlantic, and Indian Ocean. Since different pCO_2 products cover different time spans, we selected the pCO_2 inversion results from 1993 to 2019 for comparison.

Dataset	**Version**	**Method**	**Time coverage**	**Reference**
LSCE-FFNN	v2022	Deep learning	1985 to 2021	Denvil-Sommer et al., 2019
SOM-FFN	v2022	Deep learning	1981 to 2021	Landschützer et al., 2016
CSIR-ML6	v2021	Machine learning	1982 to 2020	Gregor et al., 2019
NIES-NN	v2020	Deep learning	1985 to 2019	Zeng et al., 2014
JMA-MLR	v2020	Multiple linear regression	1990 to 2019	Iida et al., 2021
JENA-MLS	v2023	Diagnostic model	1957 to 2022	Rödenbeck et al., 2013

Table 3 Datasets Used in Comparative Evaluation

Figure 5 presents the relative simulation skill of each inversion dataset to the LDEO observational data, with Figures 5a-h showing the evaluation results for the global ocean and the seven sub-regions, respectively. From Figure 5, it can be seen that the scatter points of the six inversion datasets and TSSCXG-17 reconstruction are close to each other on the Taylor diagram, reflecting similar errors, correlation coefficients, and standard deviations for the LDEO dataset. On a global scale, all inversion datasets underestimated the standard deviation of the sea surface pCO_2 , revealing a shortcoming in capturing the range of sea surface pCO_2 variability (Figure 5a). Regionally, the TSSCXG-17 inversion results were closer to the LDEO observational data in most areas, especially in the Southern Ocean, Arctic Ocean, North Pacific, and South Pacific, where TSSCXG-17 achieved higher correlation coefficients and standard deviations that were very close to the observational data (Figures 5b-e). Notably, in the Southern Ocean and Arctic Ocean, where chlorophyll field gap-filling was crucial, the evaluation results of TSSCXG-17 dataset outperformed the average inversion result of the other models (Figures 5b, c). According to the LDEO dataset, all models performed worst in reconstructing pCO_2 in the South Atlantic among the seven regions. This may be due to the lack of observational data, uneven sampling points, or different time spans of sampling observations in the SOCAT and LDEO datasets within this region.

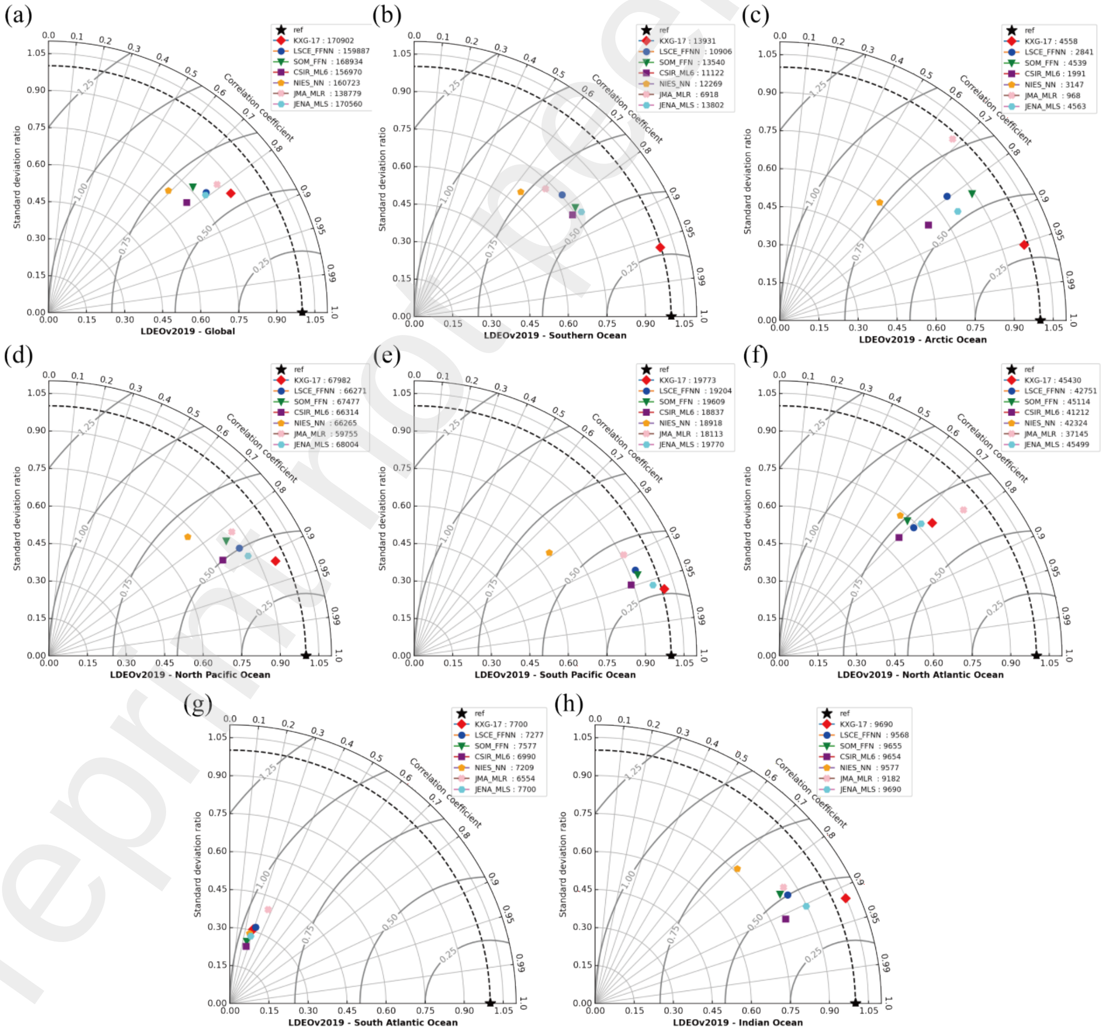


Figure 5 Comparison of Sea Surface pCO_2 Estimates from Different Inversion Datasets and pCO_2 Observational Data from the LDEOv2019 Dataset. (a) Global ocean evaluation results, (b)-(h) regional

evaluation results for the Southern Ocean, Arctic Ocean, North Pacific, South Pacific, North Atlantic, South Atlantic, and Indian Ocean, respectively. Due to temporal and spatial gaps in the other inversion datasets being compared, the standard deviation in the Taylor diagram is replaced by the ratio of the estimated standard deviation to the measured standard deviation. The legend in the upper right corner of the small figure shows the names of the inversion datasets and the total number of grid points used for evaluation in each region.

The comparative assessment of RMSE and R² indicators show that the sea surface pCO_2 inversion results based on TSSCXG-17 outperform other datasets (Table 4). Globally, the RMSE of our inversion was 24.45 µatm, ranking second among the seven datasets compared and close to the uncertainty reflected by the training error; the R² score was 0.73, better than the other six inversion datasets. The TSSCXG inversion results in the Southern Ocean had an RMSE of 16.80 µatm and an R² of 0.92, while in the Arctic Ocean, the RMSE was 16.17 µatm and the R² was 0.91, both significantly outperforming other SOCAT-based inversion models.

Dataset	**area**	**RMSE**	**R²**	**area**	**RMSE**	**R²**	**area**	**RMSE**	**R²**
TSSCXG-17	Global	24.45	**0.73**	Southern Ocean	**16.80**	**0.92**	Arctic Ocean	**16.17**	**0.91**
LSCE-FFNN		26.72	0.62		29.03	0.58		38.58	0.59
SOM-FFN		31.41	0.55		34.42	0.64		31.40	0.67
CSIR-ML6		29.03	0.59		31.04	0.66		33.09	0.65
NIES-NN		31.57	0.48		40.73	0.40		48.18	0.35
JMA-MLR		**23.18**	0.62		31.29	0.48		29.76	0.37
JENA-MLS		28.82	0.63		32.68	0.69		29.37	0.71

Table 4 Comparison of pCO_2 Observational Data from the LDEOv2019 Dataset and Sea Surface pCO_2 Estimates from Different Inversion Datasets. The table shows global average results and average evaluation results for the Southern Ocean and Arctic Ocean.

4.2 Evaluation with Time-Series and Autonomous Platform Data

In addition to the LDEO dataset, this study also used data from time series stations to validate the sea surface pCO2 inversion results. Time series stations can directly monitor sea surface partial pressure of CO2 and other carbonate system parameters. Some stations provide long-term time series observations, which can be used to evaluate the model's ability of reconstructing the temporal trends and variability of sea surface pCO2 (Bushinsky et al., 2019; Chai et al., 2020). This study selected sea surface CO2 partial pressure time series data from 35 observation stations across the world, each with long-term observations ranging from several months to several years. In calculation, we excluded observations marked as doubtful, abnormal, or faulty, retaining only those marked as good, and resampled the raw data to monthly time series for inversion result evaluation. These moored stations are distributed in the Pacific (27), Atlantic (6), Indian Ocean (1), and Southern Ocean (1). Based on the environment of their location, these stations were categorized into open ocean (18), coastal ocean (10), and coral reef ocean (7). For robust evaluation, we compared the inversion results at the nearest grid point for coastal and coral reef stations, and the average of the four nearest grid points for open ocean stations. Note that some station data were included in the SOCAT dataset and were used for model training, so evaluation errors may be lower for these data. Nonetheless, since all machine learning and deep learning models compared used SOCAT data for pCO_2, comparing these time series still reflects the relative strengths of different inversion datasets.

Table 5 shows the evaluation results of different inversion datasets against observational data in three types of marine regions. Regarding data coverage, most datasets provided inversion results for grid points at open ocean and coral reef stations, with most data missing occurring in coastal regions. The missing of estimates in coastal environments reflects the coverage limitation of previous data-based machine learning inversion datasets. After gap-filling the feature variables, TSSCXG provided inversion results for all 35 observation stations, meeting one of this study's goals: achieving global ocean coverage for sea surface pCO_2 and air-sea carbon flux inversion. In terms of accuracy, all datasets showed lower errors in open ocean regions and higher errors in coral reef and coastal regions (Table 5). Time series analysis revealed RMSE values of approximately 11-25 µatm in the open ocean, around 20-30 µatm in coral reef regions, and between 35-80 µatm in coastal areas. This discrepancy is likely due to the high spatio-temporal variability of the carbonate system in coral reef and coastal regions, where comprehensive observations of the carbonate system are still lacking. Consequently, the limited observational data and generalization capability of machine learning models currently hinder high-precision reconstruction. Among the inversion datasets, TSSCXG had the lowest RMSE and highest correlation and determination coefficients in open ocean regions, indicating superior inversion quality compared to other products. Considering inversion results with data reconstructed in coral reef and coastal regions, TSSCXG-17 performed best among all 7 data products in coastal regions and second best in coral reef regions, showing lower quality than JENA-MLS.

Dataset	Open Ocean (18)				Coral Reef (17)				Coastal (10)			
	RMSE	Corr	R^2	missing	RMSE	Corr	R^2	missing	RMSE	Corr	R^2	missing
CMEMS-FFNN	14.74	0.78	0.4	1	29.11	0.72	-0.52	**0**	63.09	0.61	0.07	4
CSIR-ML6	14.96	0.79	0.43	**0**	29.87	0.74	-0.84	**0**	66.81	0.59	0.01	1
JENA-MLS	11.76	0.86	0.59	**0**	24.90	0.81	-0.67	**0**	**38.30**	**0.83**	**0.56**	**0**
JMA-MLR	18.17	0.72	0.11	**0**	32.73	0.69	-1.01	**0**	89.74	0.20	-0.56	5
MPI-SOMFFN	15.37	0.76	0.37	**0**	30.89	0.62	-0.95	**0**	65.76	0.07	-0.36	6
NIES-FNN	23.54	0.61	-0.68	**0**	30.99	0.64	-1.13	**0**	75.70	0.48	-0.19	1
TSSCXG-17	**11.41**	**0.87**	**0.69**	**0**	**24.28**	0.80	**-0.05**	**0**	52.24	0.79	0.33	**0**

Table 5 Comparison of Sea Surface pCO_2 Time Series Data from Observational Stations and Sea Surface pCO_2 Estimates at Corresponding Grid Point from Different Inversion Datasets. The data in the table are the average evaluation results for all non-null grid points. If the inversion dataset has no inversion results at the grid point of the observation station, it is recorded as missing.

Further analysis of the temporal characteristics of time series stations data and the reconstruction capability of the TSSCXG-17 machine learning inversion model evaluated model performance at various observation stations. In the three types of regions, stations with long and continuous observation periods were selected, and their sea surface pCO_2 observations were compared with TSSCXG-17 model reconstructions (Figure 6). Both observed and reconstructed sea surface pCO_2

showed an upward trend over the years and seasonal variations with lower values in winter and higher values in summer (Figure 6). Among the 18 open ocean stations, 16 had reconstruction errors below 10% of the global average sea surface pCO_2, with 14 stations having errors below 5%. The largest open ocean inversion error was at the TAO165E station (0°, 165°E, equatorial western Pacific). Time series analysis indicated a strong consistency between the reconstructed results and observed trends, though there were some deviations in fitting winter low anomalies and summer highs (Figure 6).

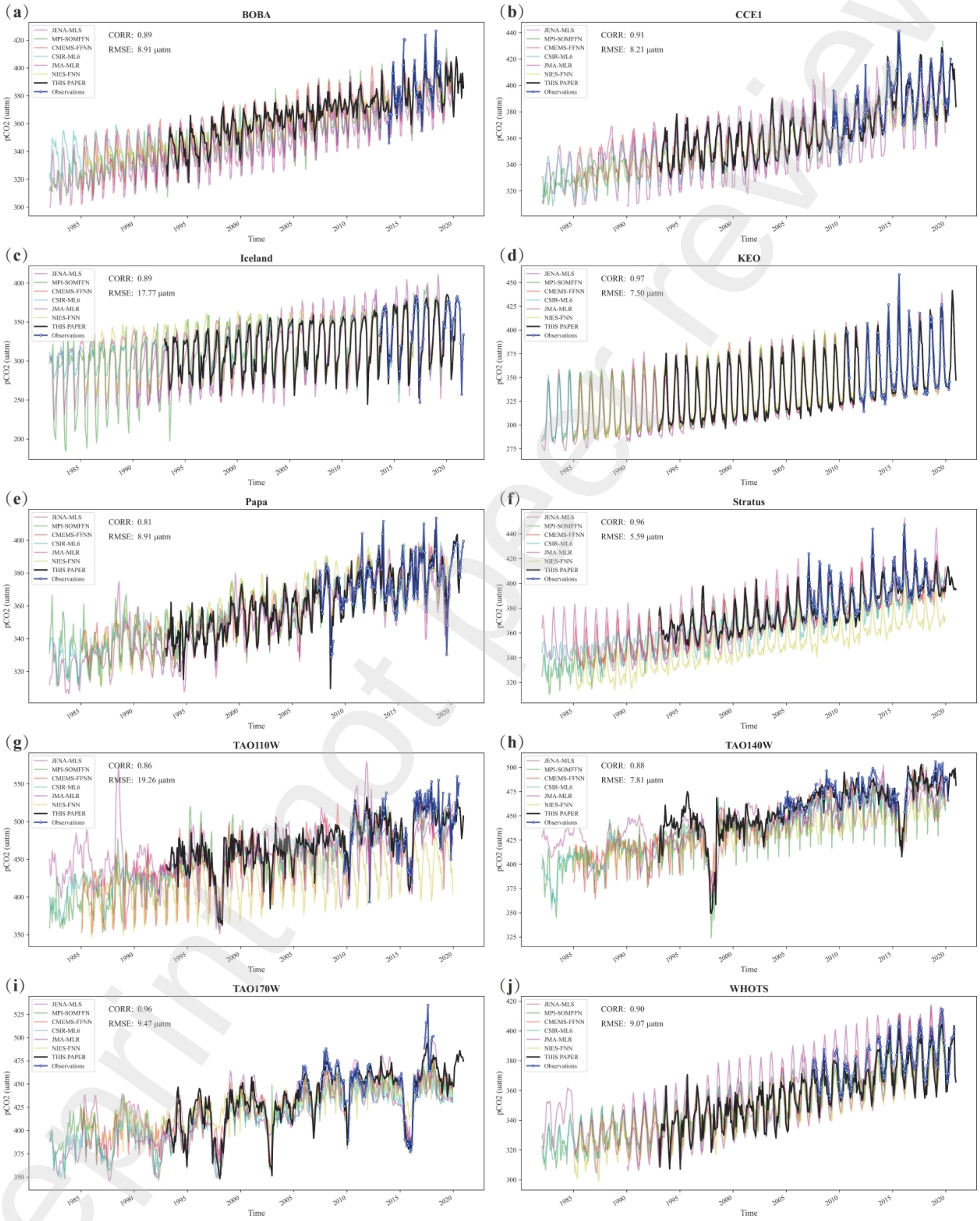


Figure 6 Examples of Sea Surface pCO_2 Observational Data and Inversion Data at Corresponding Grid Points for Open Ocean Observational Stations. The observation stations are (a) BOBOA (15°N, 90°E); (b) CCE1 (33.48°N, 122.51°W); (c) Iceland (68.0°N, 12.6°W); (d) KEO (32.28°N, 144.84°E); (e) Papa (50.13°N, 144.84°W); (f) Stratus (19.70°S, 85.60°W); (g) TAO110W (0°, 110°W); (h) TAO140W (0°, 140°W); (i) TAO170W (0°, 170°W); (j) WHOT (22.67°N, 157.98°W). The monthly observational data are shown as blue scatter points, the inversion results from this study are shown as black solid lines, and other inversion datasets are detailed in the legend.

Additionally, coastal regions are intersection points for terrestrial, atmospheric, and marine carbon cycles, with highly active biogeochemical cycles. The seasonal variability of sea surface CO_2 partial pressure and air-sea carbon flux in coastal regions far exceeds that in open oceans, with greater differences among different coastal stations. In coastal regions, TSSCXG-17 dataset can reconstruct direction of seasonal variation of coastal areas. However, due to coarse spatial resolution, all machine learning-based inversion datasets tend to underestimate the seasonal variability of sea surface CO_2 in coastal areas. Time series evaluations showed that ten coastal stations showed in Figure 7 had inversion results with an average correlation coefficient of 0.795. But the machine learning-based inversion datasets exhibit larger errors. Among similar datasets, the TSSCXG-17 dataset has the smallest error, though it is still higher than the Jena-MLS dataset, which is based on diagnostic inversion models. Among all stations used for validation, Coastal LA, Coastal MS, and First Landing OA stations had poor inversion results, with wrongly reproduced seasonal variation, evident low and high value overestimation, and abnormal high values. indicate that the lack of training data and insufficiently fine inversion resolution still hinders the model's accuracy in reconstructing coastal areas.

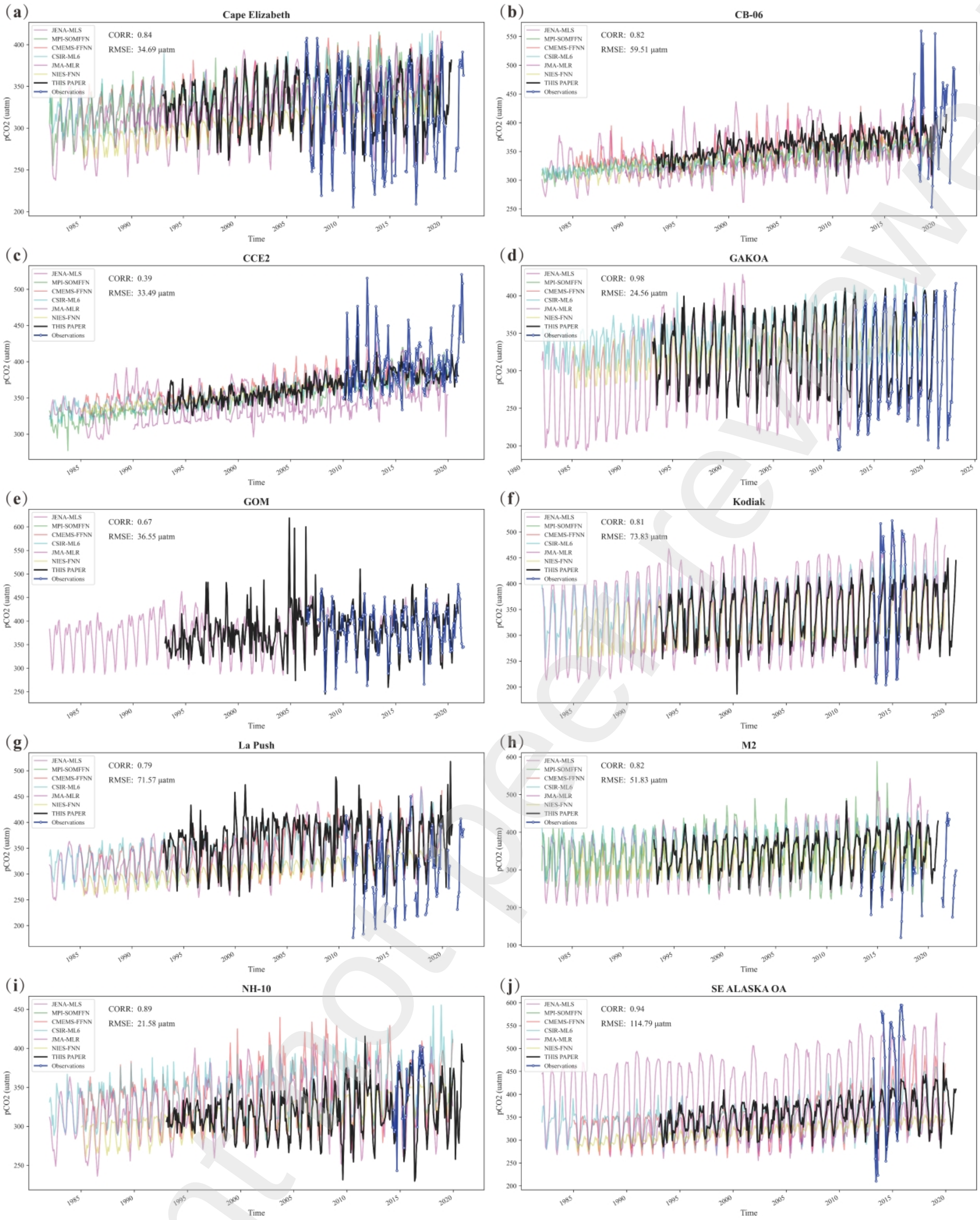


Figure 7 As in Figure 6, observational data for coastal observational stations. The observation stations are (a) Cape Elizabeth (47.35°N, 124.73°W); (b)CB-06(); (c)CCE2(); (d) GAKOA (59.91°N, 149.35°W); (e) GOM (43.02°N, 70.54°W); (f) Kodiak (57.70°N, -152.31°W); (g) La Push (47.97°N, 124.95°W); (h)M2(); (i)NH-10(); (j) SE ALASKA OA (56.26°N, 134.67°W).

Moreover, sea surface pCO_2 in coral reef regions is primarily controlled by their ecosystems and calcium carbonate production. Studies have found that the diurnal variability of sea surface pCO_2 in coral reefs can be up to 10 times that of open oceans. Time series analysis showed a significant interdecadal increase in coral reef sea surface pCO_2, with smaller seasonal variations (Figure 8). Comparisons of observed and reconstructed results indicated good reconstruction at some stations like BOBOA, Chuuk_K1, and Crescent Reef, but some stations showed small regression model errors with low determination coefficients, reflecting insufficient model fitting for coral reef regions. The machine learning model, trained mainly on open ocean samples, reconstructed the carbon dioxide partial pressure for coral reef regions, leading to decreased model performance in some areas.

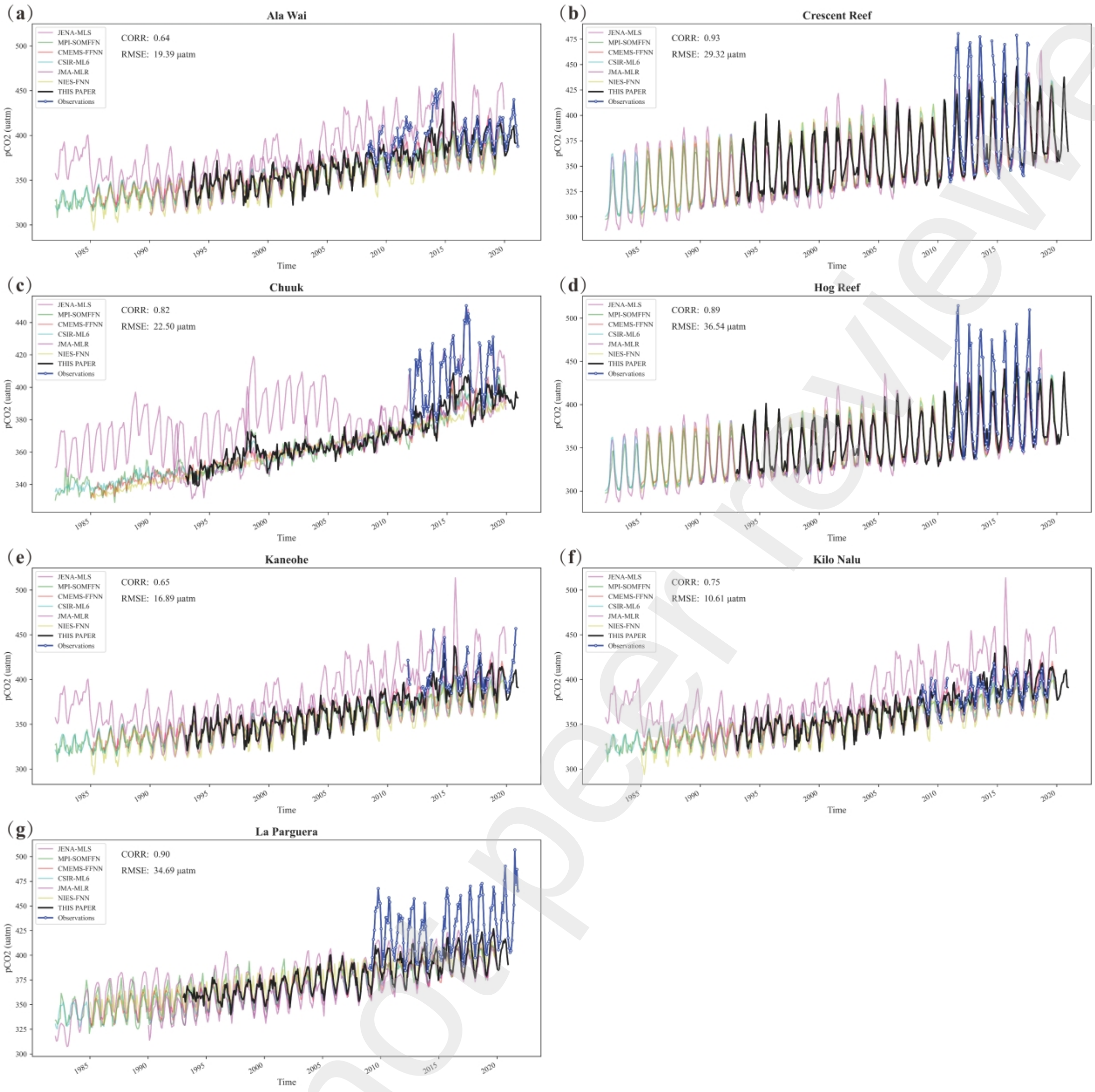


Figure 8 As in Figure 6, observational data for coral reef observational stations. The observation stations are (a) Ala Wai (21.28°N, 157.85°W); (b)Crescent Reef (32.40°N, 64.79°W); (c) Chuuk (7.46°N, 151.90°E); (d) Hog Reef (); (e) Kaneohe (21.48°N, 157.78°W); (f)Kilo Nalu(); (g) La Parguera ().

4.3 Feature Importance in Inversion Models

Although machine learning models cannot accurately establish dynamic equations between feature variables and target observations, we can understand the utilization of input features in the trained regression model using the characteristics of the XGBoost model by each feature’s importance. Figure 9 shows the relative importance of each feature variable in the XGBoost regression models trained for the 17 clusters configuration. In tree-based models like XGBoost, feature importance can be measured by "information gain," which represents the average gain brought by a feature across all tree splits. This is the average contribution of each feature in every tree, indicating the average impact of the feature on the optimization of the objective function during the node splitting process. In the application of machine learning models, feature importance is a key metric, where higher feature importance indicates a greater contribution of that feature to generating predictions. The heatmap in Figure 10 displays the relative importance of features within the 17 clusters. By definition, the sum of the relative importance of all features within the training model for each cluster is 1. The bar chart on the right side of Figure 10 represents the sum of the relative importance of the 12 feature variables across all clusters, reflecting the contribution of different feature variables to the reconstruction results of the global sea surface CO_2 partial pressure field.

Globally, the relative importance of input features varies across regions. Key variables such as mole fraction of CO_2 in dry air(xCO_2), mixed layer depth (MLD), sea surface temperature (SST), chlorophyl-a concentration (Chla), and sea surface salinity (SSS) are the dominant features in terms of information gain, while other feature variables have relatively smaller influence. Specifically, the influence of sea ice and eddy kinetic energy (EKE) is minimal and are significant only in specific regions. Although there is significant spatiotemporal variability in sea surface CO_2 partial pressure, it exhibits a long-term upward trend driven by the increase in atmospheric CO_2 concentration globally. Therefore, despite the primary influencing factors varying across different regions, xCO_2 is the feature with the highest relative importance in the regression model, reflecting that external forcing from natural and anthropogenic CO_2 emissions is the main factor influencing the evolution of sea surface CO_2 partial pressure. Next in importance are mixed layer depth and sea temperature. Mixed layer depth, representing ocean circulation and vertical mixing, provides the critical information in areas like the eastern equatorial Pacific (cluster 15), low-latitude Southern Pacific and Indian Oceans (cluster 5), and the mid-to-high latitude Atlantic Ocean (clusters 9, 12, 13). SST contributes over 30% of the information gain in the South Pacific Subtropical Gyre (cluster 14) and about 17% in the eastern equatorial Pacific (cluster 15). The equatorial Pacific releases a large amount of CO_2 to the atmosphere annually, significantly influenced by the interannual variability of the El Niño-Southern Oscillation (ENSO) phenomenon, contributing to much of the global ocean carbon sink's interannual variability (Chatterjee et al., 2017; Liu and Xie, 2017; Rödenbeck et al., 2022). Chlorophyll concentration is the most important feature variable in the Southern Ocean region (cluster 1), consistent with previous research results (Yang et al., 2024). This indicates that after feature imputation, the imputed chlorophyll concentration in this region contains information from other variables such as SST, salinity, and MLD. This explains the strong performance of the TSSCXG-17 machine learning model in the Southern Ocean. In summary, the ocean carbon cycle is influenced by atmospheric external forcing, ocean thermodynamics, ocean circulation, and biochemical processes, which together shape the complex evolution of global sea surface CO_2 partial pressure.

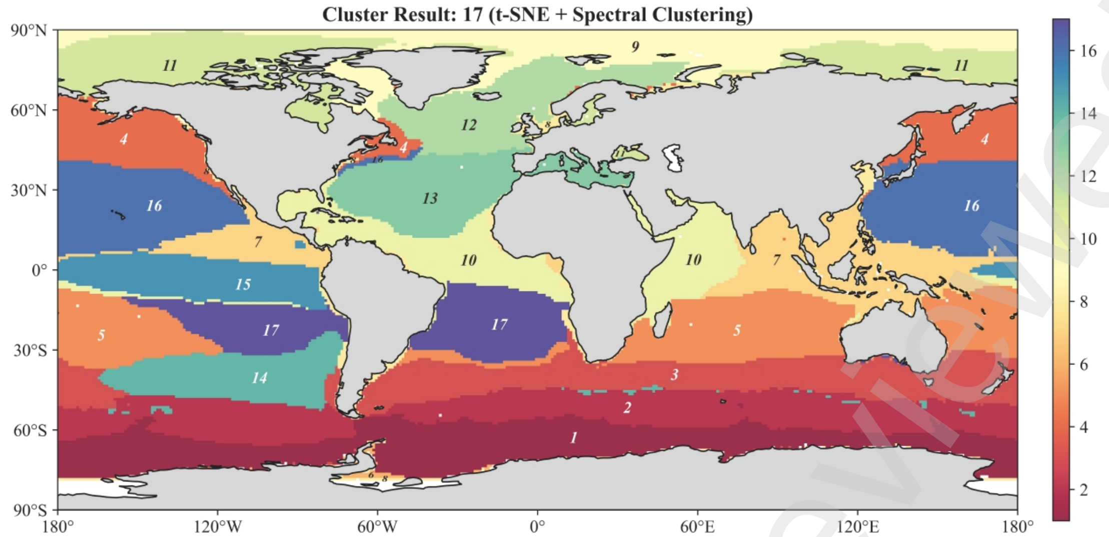


Figure 9 Ocean Clustering Results Using t-SNE Nonlinear Dimensionality Reduction and Spectral Clustering Algorithm with 17 Total Clusters

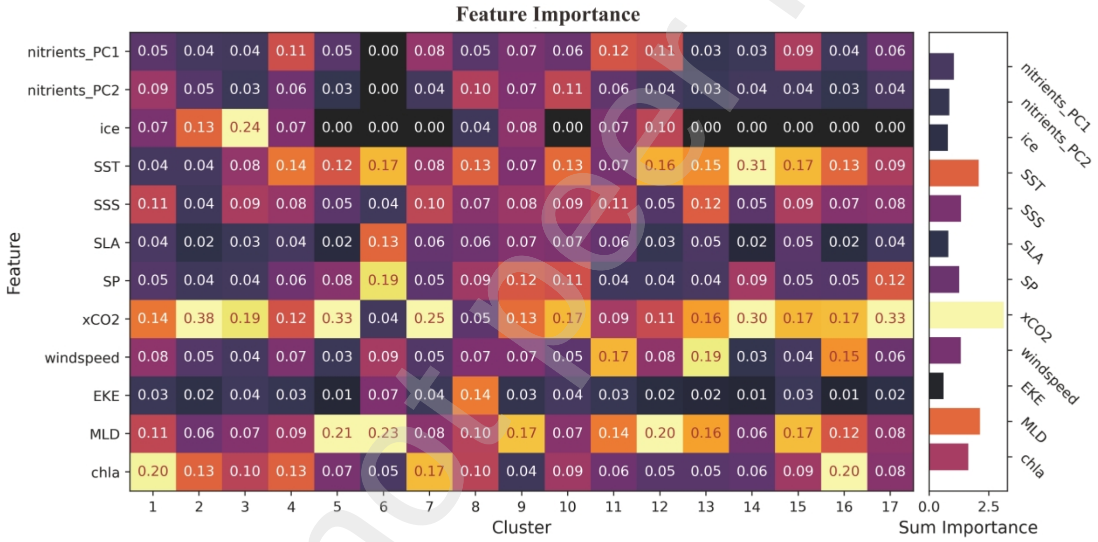


Figure 10 Feature Importance Obtained from the Trained Model. The relative importance of feature variables in the 17 clusters (left, heatmap) and the total relative importance of the 12 feature variables across all clusters (right, bar chart) are shown.

4.4 Results of Global Air-sea CO_2 Flux Inversion

Based on the formula described in Section 2.3, this study calculated the global air-sea carbon flux at a monthly 1°×1° spatial resolution from 1993 to 2020. Figure 11 shows the climatological mean spatial distribution, the uncertainty of the reconstruction results, the interannual trend, and the amplitude of seasonal variation of the air-sea CO_2 flux inversion within the reconstruction period. In Figure 11a, negative values indicate that the direction of the CO_2 flux is from the atmosphere to the ocean, and positive values indicate the opposite. As shown in Figure 11, there is a strong CO_2 outgassing from the ocean to the atmosphere in the equatorial Pacific Ocean during the period from 1993 to 2020, while the mid-latitude regions of both the Northern and Southern Hemispheres show strong ocean uptake, consistent with previous studies (Gruber et al., 2009; Takahashi et al., 2009). Additionally, the time series for most grid points in the global ocean show a significant increasing trend of ocean carbon uptake from 1993 to 2020 (Figure 11c). The centers of the enhanced carbon sink are concentrated in the mid-to-high latitude regions of both hemispheres. The Southern Ocean, acting as a weak carbon source, shows a decreasing trend in carbon flux to the atmosphere year by year (Figure 11c). The regions with higher uncertainty in carbon flux are concentrated in the high latitudes of the Northern Hemisphere (Figure 11b). The spatial distribution characteristics of uncertainty are similar to those in other studies, but the values are smaller, indicating that using only the standard deviation of the data does not fully reflect the total uncertainty in the interpolation and extrapolation process based on limited observational data (Rodgers et al., 2023).

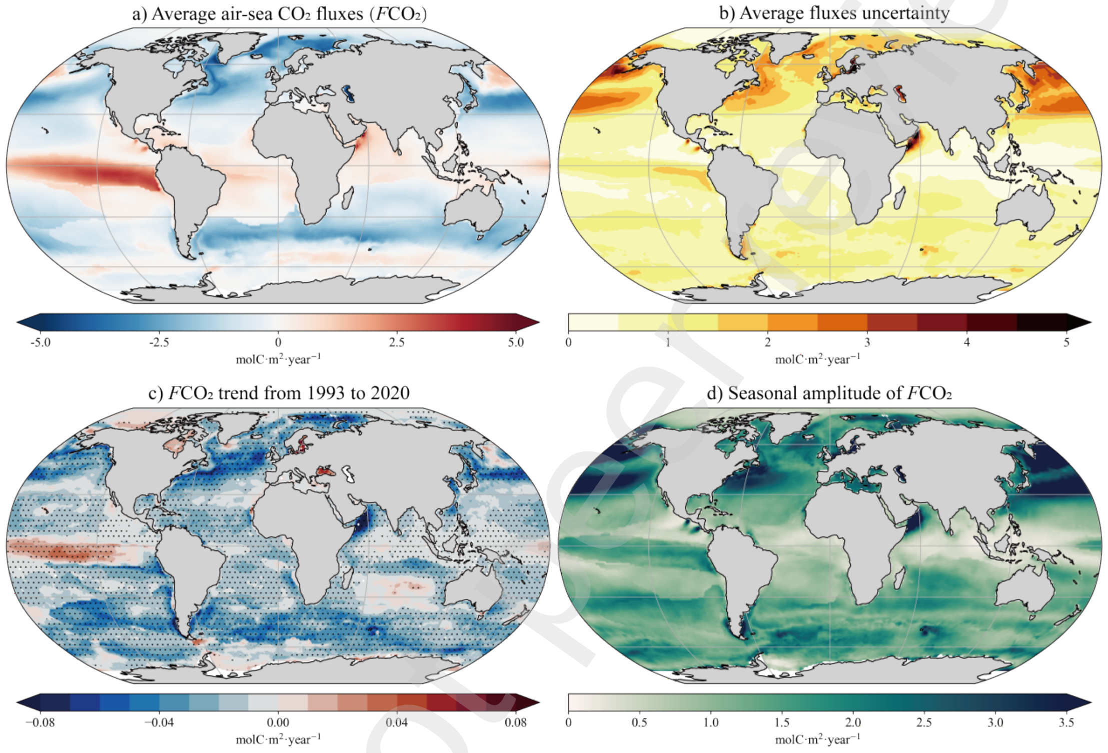

a) Average air-sea CO₂ fluxes (FCO₂)
b) Average fluxes uncertainty
-5.0
-2.5
0
2.5
5.0
molC·m²·year⁻¹
0
1
2
3
4
5
molC·m²·year⁻¹
c) FCO₂ trend from 1993 to 2020
d) Seasonal amplitude of FCO₂
−0.08
−0.04
0.00
0.04
0.08
molC·m²·year⁻¹
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
molC·m²·year⁻¹

Figure 11. Reconstruction results of global air-sea carbon flux from 1993 to 2020, including: (a) Average spatial distribution of global air-sea CO_2 flux; (b) Average spatial distribution of the uncertainty in air-sea carbon flux; (c) Trend in global air-sea carbon flux from 1993 to 2020, with dotted regions indicating significant linear trends ($p<0.05$); (d) Average seasonal variation intensity of global air-sea carbon flux.

By area-weighted summation of the monthly air-sea CO_2 flux gridded data, the time series of global integrated air-sea CO_2 flux can be obtained. Using the STL algorithm (Seasonal-Trend decomposition using LOESS), the TSSCXG-17 reconstructed global integrated CO_2 flux is decomposed into nonlinear interannual trends and seasonal variation components (Cleveland et al., 1990). The global carbon flux time series, interannual trends, and seasonal variability are shown in Figure 11. The TSSCXG reconstruction results indicate that, on average, the ocean carbon sink absorbed about -2.45 PgC·yr-1 from the atmosphere and terrestrial systems between 1993 and 2020 (Figure 12a). There are two distinct periods in the interannual variation trend of global carbon flux (Figure 12b). From the 1990s to the early 2000s, the total global ocean carbon sink showed a weakening trend, termed the stagnation of the ocean carbon sink. During this time, the global ocean carbon sink weakened at a rate of 0.048 PgC per year. However, from the early 2000s to the present, the global ocean carbon sink capacity rebounded rapidly, increasing annually at a rate of -0.105 PgC. This phenomenon of decreased ocean carbon absorption rates in the 1990s and increased rates from 2000 to 2020 has been reported in various studies on both global and regional scales (DeVries, 2022b; Dong et al., 2022; Gruber et al., 2023; Le Quéré et al., 2007; McKinley et al., 2020; Pérez et al., 2013; Ritter et al., 2017). Studies have pointed out that external factors, such as volcanic eruptions, caused a decrease in sea surface temperature and changes in ocean circulation, leading to the stagnation of the ocean carbon sink; subsequently, the substantial amount of CO_2 emissions from human activities became the primary reason for the rapid rebound of the carbon sink with negative decadal variations (McKinley et al., 2020).

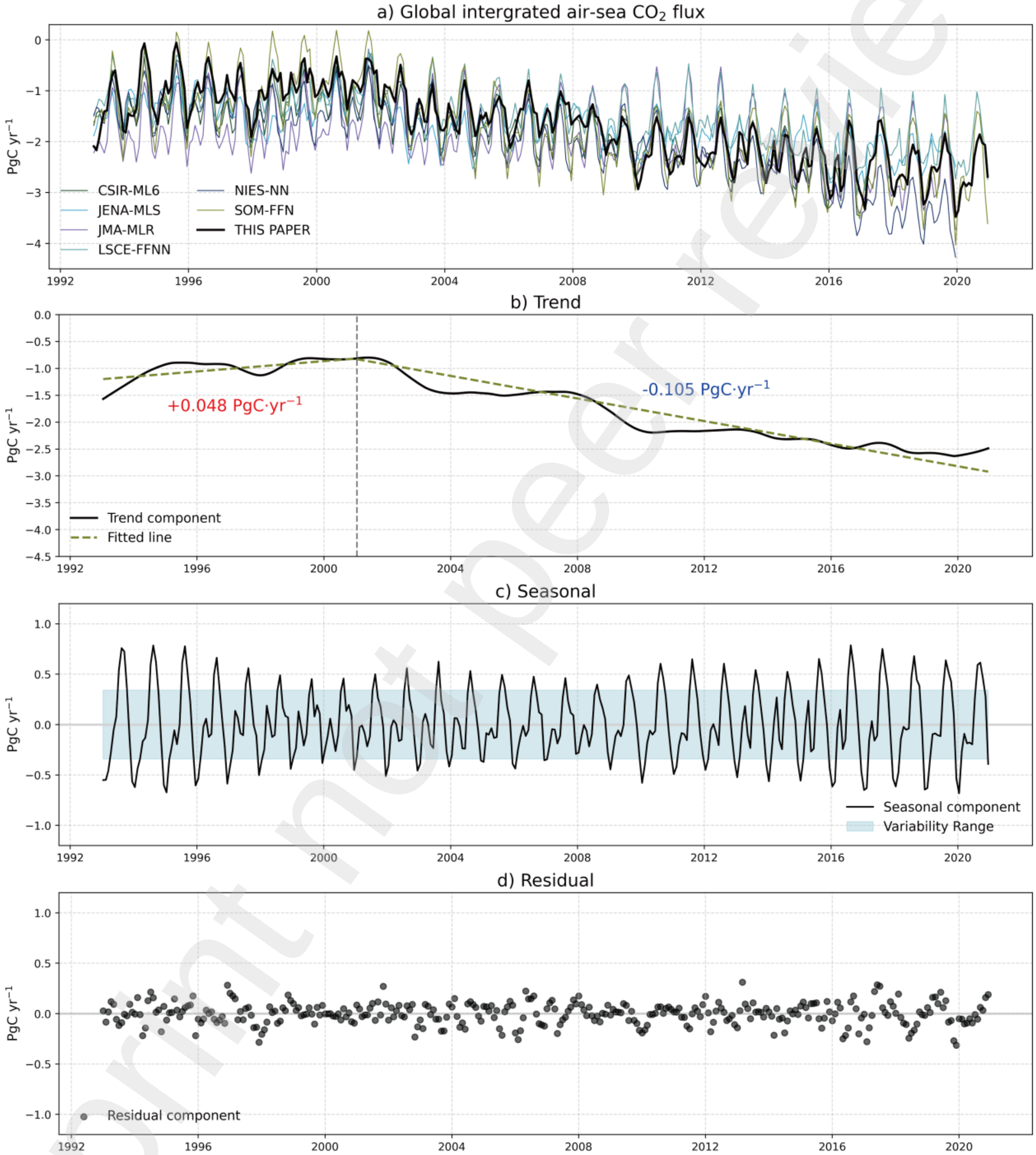


Figure 12. Global integrated air-sea CO_2 flux (a) and its interannual trend (b), seasonal variability (c), and residuals (d) decomposed by the STL method.

After removing the interannual variation trend, the global air-sea carbon flux exhibits a semi-bimodal seasonal cycle (Figure 12c). The average seasonal variation shows that the total ocean carbon sink reaches its peak in winter and then gradually decreases until the late spring. After a short-term increase in summer (April to May), the ocean's uptake of atmospheric CO_2 gradually decreases, reaching the lowest point of the total carbon sink in August, and then gradually recovers. This semi-bimodal seasonal variation pattern is mainly caused by the seasonal periodic cycle of difference in partial pressures of CO_2 in the sea surface and atmosphere. Near-surface atmospheric CO_2 content shows a seasonal oscillation, with higher levels in summer and lower levels in autumn. This is due to the Northern Hemisphere's extensive land area and biomass, which control the global seasonal fluctuation of atmospheric CO_2. Meanwhile, the sea surface partial pressure of CO_2 is influenced by both the Northern and Southern Hemispheres with different phase in thermal properties, showing a semi-bimodal trend. Therefore, the global air-sea carbon flux exhibits a semi-bimodal seasonal oscillation. Since the 2000s, the amplitude of the seasonal cycle has shown an increasing trend on a decadal scale (Landschützer et al., 2018). Comparing the TSSCXG-17 pCO_2 reconstruction results with other data-based reconstruction products reveals that the seasonal variation intensity of pCO_2 reconstructed by TSSCXG is slightly smaller than that of other inversion products, and is closer to the results of the seaflux ensemble result (Figure 13). For the equatorial regions and mid-to-high latitude regions of both hemispheres, the average seasonal variation of TSSCXG reconstruction results falls within the distribution range of the other six inversion products. The differences in seasonal variation amplitude are mainly reflected in the extra-equatorial mid-to-low latitude regions of both hemispheres (10°~40°, Northern and Southern Hemispheres), where the TSSCXG reconstruction results are smaller (Figure 13c, e).

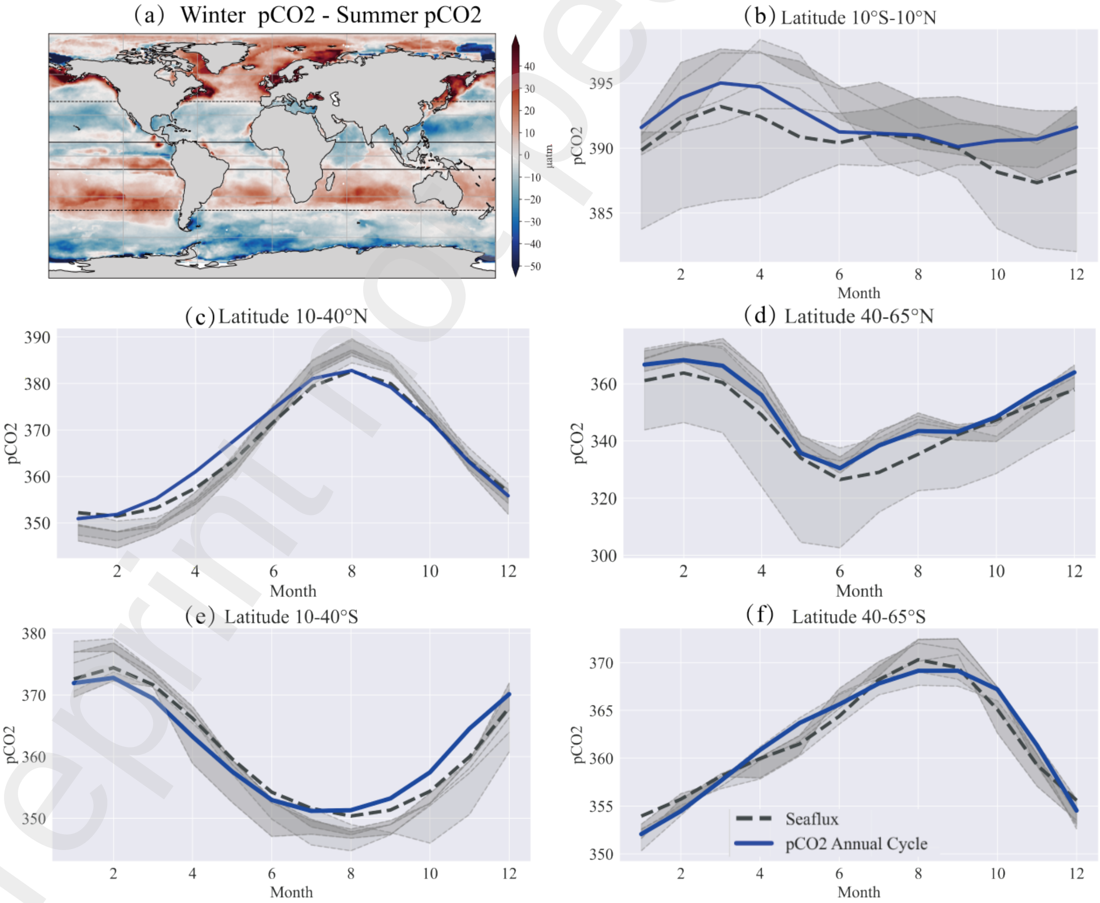


Figure 13. (a) Difference between the climatological average pCO_2 in Northern Hemisphere winter and Northern Hemisphere summer for each 1°×1° grid from 1993 to 2020. (b) 10°S-10°N; (c) 10-40°N; (d) 40-

65°N; (e) 10-40°S; (f) 40-65°S, each subplot shows the average seasonal cycle of pCO_2 within its respective latitude range from 1993 to 2020.

5 Discussion

This study introduces a novel two-stage air-sea carbon flux inversion method based on multi-source data fusion. In the first stage, we supplemented in-situ observational data, satellite remote sensing data, and reanalysis data with ocean model simulation data to create a comprehensive feature dataset using a machine learning model. This approach addressed gaps in ocean observational data. In the second stage, we developed the TSSCXG inversion algorithm, which combines t-SNE dimensionality reduction, spectral clustering, and XGBoost regression, to derive a global 1°×1° sea surface pCO_2 inversion dataset and corresponding air-sea carbon flux dataset from 1993 to 2020. This two-stage TSSCXG scheme overcomes major limitations of current spatiotemporal interpolation methods, such as incomplete temporal and spatial coverage and high inversion errors in polar and coastal regions.

Evaluation with independent flight measurement datasets and station time series observations indicates that our inversion model accurately reproduces observed spatial and temporal variations. The global inversion results are more precise than other data inversion products using similar or more complex algorithms. Compared to other pCO_2 reconstruction datasets, our results show a lower global grid average root mean square error (RMSE) and higher coefficient of determination (R2). Specifically, the TSSCXG reconstruction reduces the average grid RMSE by about 12 μatm in the Southern Ocean compared to the LSCE-FFNN dataset and by about 13.1 μatm in the Arctic Ocean compared to the JENA-MLS dataset. Consequently, the estimated global air-sea carbon flux from 1993 to 2020 using the TSSCXG method is -2.45 PgC·yr-1. Our results are consistent with previous studies, showing significant CO_2 emissions from the equatorial Pacific and strong absorption in the mid-latitude regions of both hemispheres (Gruber et al., 2009; Takahashi et al., 2009). The increasing trend in oceanic carbon uptake from 1993 to 2020, particularly in the mid to high latitude regions, corroborates findings from earlier research (DeVries, 2022b; Dong et al., 2022; Gruber et al., 2023). These trends support the hypothesis that external factors, such as volcanic eruptions and human activities, significantly influence global carbon cycles, leading to fluctuations in oceanic carbon uptake rates (McKinley et al., 2020).

Despite these advancements, several limitations remain. The use of model simulation data as a substitute for observational data introduces uncertainties that are difficult to fully quantify. Additionally, while robust, the TSSCXG algorithm may still suffer from biases inherent in the input datasets and modeling assumptions. Future research should focus on refining the algorithm, incorporating more diverse data sources, and extending the temporal coverage of datasets. Integrating advanced machine learning techniques and hybrid models could also enhance the accuracy and reliability of inversion results.

This study presents a novel approach to address the challenges associated with missing ocean observational data and offers a new perspective for future research on air-sea carbon flux inversion. By leveraging multi-source data fusion and advanced machine learning algorithms, we can better understand and predict the dynamics of the oceanic carbon cycle, contributing to more accurate assessments of global carbon budgets and informing climate policy decisions.

Data availability

The TSSCXG-17 dataset is openly available in Zenodo at https://doi.org/10.5281/zenodo.13823172.

Author contribution

Yongqiang Chen: Data curation, Methodology, Software, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Siyi Wang:** Writing - Review & Editing, Visualization. **Wenping He:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

Special issue statement

Acknowledgements

References

Azeem, M.A., Dey, P., Dev, S., 2023. A Multidimensionality Reduction Approach to Rainfall Prediction, in: 2023 Photonics & Electromagnetics Research Symposium (PIERS). Presented at the 2023 Photonics & Electromagnetics Research Symposium (PIERS), pp. 499–508. https://doi.org/10.1109/PIERS59004.2023.10221498

Bakker, D.C.E., Alin, S.R., Becker, M., Bittig, H.C., Castaño-Primo, R., Feely, R.A., Gkritzalis, T., Kadono, K., Kozyr, A., Lauvset, S.K., Metzl, N., Munro, D.R., Nakaoka, S.-I., Nojiri, Y., O'Brien, K.M., Olsen, A., Pfeil, B., Pierrot, D., Steinhoff, T., Sullivan, K.F., Sutton, A.J., Sweeney, C., Tilbrook, B., Wada, C., Wanninkhof, R., Willstrand Wranne, A., Akl, J., Apelthun, L.B., Bates, N., Beatty, C.M., Burger, E.F., Cai, W.-J., Cosca, C.E., Corredor, J.E., Cronin, M., Cross, J.N., De Carlo, E.H., DeGrandpre, M.D., Emerson, S.R., Enright, M.P., Enyo, K., Evans, W., Frangoulis, C., Fransson, A., García-Ibáñez, M.I., Gehrung, M., Giannoudi, L., Glockzin, M., Hales, B., Howden, S.D., Hunt, C.W., Ibánhez, J.S.P., Jones, S.D., Kamb, L., Körtzinger, A., Landa, C.S., Landschützer, P., Lefèvre, N., Lo Monaco, C., Macovei, V.A., Maenner Jones, S., Meinig, C., Millero, F.J., Monacci, N.M., Mordy, C., Morell, J.M., Murata, A., Musielewicz, S., Neill, C., Newberger, T., Nomura, D., Ohman, M., Ono, T., Passmore, A., Petersen, W., Petihakis, G., Perivoliotis, L., Plueddemann, A.J., Rehder, G., Reynaud, T., Rodriguez, C., Ross, A.C., Rutgersson, A., Sabine, C.L., Salisbury, J.E., Schlitzer, R., Send, U., Skjelvan, I., Stamataki, N., Sutherland, S.C., Tadokoro, K., Tanhua, T., Telszewski, M., Trull, T., Vandemark, D., Van Ooijen, E., Voynova, Y.G., Wang, H., Weller, R.A., Whitehead, C., Wilson, D.,

2022. Surface Ocean CO2 Atlas Database Version 2022 (SOCATv2022) (NCEI Accession 0253659). https://doi.org/10.25921/1H9F-NB73

Bakker, D.C.E., Pfeil, B., Landa, C.S., Metzl, N., O’Brien, K.M., Olsen, A., Smith, K., Cosca, C., Harasawa, S., Jones, S.D., Nakaoka, S., Nojiri, Y., Schuster, U., Steinhoff, T., Sweeney, C., Takahashi, T., Tilbrook, B., Wada, C., Wanninkhof, R., Alin, S.R., Balestrini, C.F., Barbero, L., Bates, N.R., Bianchi, A.A., Bonou, F., Boutin, J., Bozec, Y., Burger, E.F., Cai, W.-J., Castle, R.D., Chen, L., Chierici, M., Currie, K., Evans, W., Featherstone, C., Feely, R.A., Fransson, A., Goyet, C., Greenwood, N., Gregor, L., Hankin, S., Hardman-Mountford, N.J., Harlay, J., Hauck, J., Hoppema, M., Humphreys, M.P., Hunt, C.W., Huss, B., Ibánhez, J.S.P., Johannessen, T., Keeling, R., Kitidis, V., Körtzinger, A., Kozyr, A., Krasakopoulou, E., Kuwata, A., Landschützer, P., Lauvset, S.K., Lefèvre, N., Lo Monaco, C., Manke, A., Mathis, J.T., Merlivat, L., Millero, F.J., Monteiro, P.M.S., Munro, D.R., Murata, A., Newberger, T., Omar, A.M., Ono, T., Paterson, K., Pearce, D., Pierrot, D., Robbins, L.L., Saito, S., Salisbury, J., Schlitzer, R., Schneider, B., Schweitzer, R., Sieger, R., Skjelvan, I., Sullivan, K.F., Sutherland, S.C., Sutton, A.J., Tadokoro, K., Telszewski, M., Tuma, M., Van Heuven, S.M.A.C., Vandemark, D., Ward, B., Watson, A.J., Xu, S., 2016. A multi-decade record of high-quality CO_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

Balamurali, M., Silversides, K.L., Melkumyan, A., 2019. A comparison of t-SNE, SOM and SPADE for identifying material type domains in geological data. Computers & Geosciences 125, 78–89. https://doi.org/10.1016/j.cageo.2019.01.011

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

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G., 2013. API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. pp. 108–122.

Burger, F.A., John, J.G., Frölicher, T.L., 2020. Increase in ocean acidity variability and extremes under increasing atmospheric CO_2. Biogeosciences 17, 4633–4662. https://doi.org/10.5194/bg-17-4633-2020

Bushinsky, S.M., Takeshita, Y., Williams, N.L., 2019. Observing Changes in Ocean Carbonate Chemistry: Our Autonomous Future. Curr Clim Change Rep 5, 207–220. https://doi.org/10.1007/s40641-019-00129-8

Carroll, D., Menemenlis, D., Adkins, J.F., Bowman, K.W., Brix, H., Dutkiewicz, S., Fenty, I., Gierach, M.M., Hill, C., Jahn, O., Landschützer, P., Lauderdale, J.M., Liu, J., Manizza, M., Naviaux, J.D., Rödenbeck, C., Schimel, D.S., Van Der Stocken, T., Zhang, H., 2020. The ECCO-Darwin Data-Assimilative Global Ocean Biogeochemistry Model: Estimates of Seasonal to Multidecadal

Surface Ocean p CO $_2$ and Air-Sea CO $_2$ Flux. J Adv Model Earth Syst 12, e2019MS001888. https://doi.org/10.1029/2019MS001888

Chai, F., Johnson, K.S., Claustre, H., Xing, X., Wang, Y., Boss, E., Riser, S., Fennel, K., Schofield, O., Sutton, A., 2020. Monitoring ocean biogeochemistry with autonomous platforms. Nat Rev Earth Environ 1, 315–326. https://doi.org/10.1038/s43017-020-0053-y

Chatterjee, A., Gierach, M.M., Sutton, A.J., Feely, R.A., Crisp, D., Eldering, A., Gunson, M.R., O'Dell, C.W., Stephens, B.B., Schimel, D.S., 2017. Influence of El Niño on atmospheric CO2 over the tropical Pacific Ocean: Findings from NASA's OCO-2 mission. Science 358, eaam5776. https://doi.org/10.1126/science.aam5776

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: A seasonal-trend decomposition. J. Off. Stat 6, 3–73.

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 pCO$_2$ over the global ocean. Geoscientific Model Development 12, 2091–2105. https://doi.org/10.5194/gmd-12-2091-2019

DeVries, T., 2022a. The Ocean Carbon Cycle. Annual Review of Environment and Resources 47, 317–341. https://doi.org/10.1146/annurev-environ-120920-111307

DeVries, T., 2022b. Atmospheric CO $_2$ and Sea Surface Temperature Variability Cannot Explain Recent Decadal Variability of the Ocean CO $_2$ Sink. Geophysical Research Letters 49, e2021GL096018. https://doi.org/10.1029/2021GL096018

DeVries, T., Yamamoto, K., Wanninkhof, R., Gruber, N., Hauck, J., Müller, J.D., Bopp, L., Carroll, D., Carter, B., Chau, T., Doney, S.C., Gehlen, M., Gloege, L., Gregor, L., Henson, S., Kim, J.H., Iida, Y., Ilyina, T., Landschützer, P., Le Quéré, C., Munro, D., Nissen, C., Patara, L., Pérez, F.F., Resplandy, L., Rodgers, K.B., Schwinger, J., Séférian, R., Sicardi, V., Terhaar, J., Triñanes, J., Tsujino, H., Watson, A., Yasunaka, S., Zeng, J., 2023. Magnitude, Trends, and Variability of the Global Ocean Carbon Sink From 1985 to 2018. Global Biogeochemical Cycles 37, e2023GB007780. https://doi.org/10.1029/2023GB007780

Dong, Y., Bakker, D.C.E., Bell, T.G., Huang, B., Landschützer, P., Liss, P.S., Yang, M., 2022. Update on the Temperature Corrections of Global Air-Sea CO $_2$ Flux Estimates. Global Biogeochemical Cycles 36. https://doi.org/10.1029/2022GB007360

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., Taylor, K.E., 2016. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development 9, 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Fay, A.R., Gregor, L., Landschützer, P., McKinley, G.A., Gruber, N., Gehlen, M., Iida, Y., Laruelle, G.G., Rödenbeck, C., Roobaert, A., Zeng, J., 2021. SeaFlux: harmonization of air–sea CO$_2$ fluxes from surface pCO$_2$ data products using a standardized approach. Earth System Science Data 13, 4693–4710. https://doi.org/10.5194/essd-13-4693-2021

Fay, A.R., McKinley, G.A., Lovenduski, N.S., 2014. Southern Ocean carbon trends: Sensitivity to methods. Geophysical Research Letters 41, 6833–6840. https://doi.org/10.1002/2014GL061324

Friedlingstein, P., O'Sullivan, M., Jones, M.W., Andrew, R.M., Bakker, D.C.E., Hauck, J., Landschützer, P., Le Quéré, C., Luijkx, I.T., Peters, G.P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J.G., Ciais, P., Jackson, R.B., Alin, S.R., Anthoni, P., Barbero, L., Bates, N.R., Becker, M., Bellouin, N., Decharme, B., Bopp, L., Brasika, I.B.M., Cadule, P., Chamberlain, M.A., Chandra, N., Chau, T.-T.-T., Chevallier, F., Chini, L.P., Cronin, M., Dou, X., Enyo, K., Evans, W., Falk, S., Feely, R.A., Feng, L., Ford, D.J., Gasser, T., Ghattas, J., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Hefner, M., Heinke, J., Houghton, R.A., Hurtt, G.C., Iida, Y., Ilyina, T., Jacobson, A.R., Jain, A., Jarníková, T., Jersild, A., Jiang, F., Jin, Z., Joos, F., Kato, E., Keeling, R.F., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J.I., Körtzinger, A., Lan, X., Lefèvre, N., Li, H., Liu, J., Liu, Z., Ma, L., Marland, G., Mayot, N., McGuire, P.C., McKinley, G.A., Meyer, G., Morgan, E.J., Munro, D.R., Nakaoka, S.-I., Niwa, Y., O'Brien, K.M., Olsen, A., Omar, A.M., Ono, T., Paulsen, M., Pierrot, D., Pocock, K., Poulter, B., Powis, C.M., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T.M., Schwinger, J., Séférian, R., Smallman, T.L., Smith, S.M., Sospedra-Alfonso, R., Sun, Q., Sutton, A.J., Sweeney, C., Takao, S., Tans, P.P., Tian, H., Tilbrook, B., Tsujino, H., Tubiello, F., van der Werf, G.R., van Ooijen, E., Wanninkhof, R., Watanabe, M., Wimart-Rousseau, C., Yang, D., Yang, X., Yuan, W., Yue, X., Zaehle, S., Zeng, J., Zheng, B., 2023. Global Carbon Budget 2023. Earth System Science Data 15, 5301–5369. https://doi.org/10.5194/essd-15-5301-2023

Friedrich, T., Oschlies, A., 2009. Neural network-based estimates of North Atlantic surface pCO$_2$ from satellite data: A methodological study. J. Geophys. Res. 114, 2007JC004646. https://doi.org/10.1029/2007JC004646

Geng, G., Xiao, Q., Liu, S., Liu, X., Cheng, J., Zheng, Y., Xue, T., Tong, D., Zheng, B., Peng, Y., Huang, X., He, K., Zhang, Q., 2021. Tracking Air Pollution in China: Near Real-Time PM$_{2.5}$ Retrievals from Multisource Data Fusion. Environ. Sci. Technol. 55, 12106–12115. https://doi.org/10.1021/acs.est.1c01863

Gregor, L., 2023. SeaFlux v2023: harmonised sea-air CO2 fluxes from surface pCO2 data products using a standardised approach. https://doi.org/10.5281/ZENODO.4133802

Gregor, L., Humphreys, M., 2021. lukegre/SeaFlux: Updated continuous integration and docs. https://doi.org/10.5281/ZENODO.4659162

Gregor, L., Lebehot, A.D., Kok, S., Scheel Monteiro, P.M., 2019. A comparative assessment of the uncertainties of global surface ocean CO$_2$ estimates using a machine-learning ensemble (CSIR-ML6 version 2019a) – have we hit the wall? Geoscientific Model Development 12, 5113–5136. https://doi.org/10.5194/gmd-12-5113-2019

Gruber, N., Bakker, D.C.E., DeVries, T., Gregor, L., Hauck, J., Landschützer, P., McKinley, G.A., Müller, J.D., 2023. Trends and variability in the ocean carbon sink. Nat Rev Earth Environ 4, 119–134. https://doi.org/10.1038/s43017-022-00381-x

Gruber, N., Boyd, P.W., Frölicher, T.L., Vogt, M., 2021. Biogeochemical extremes and compound events in the ocean. Nature 600, 395–407. https://doi.org/10.1038/s41586-021-03981-7

Gruber, N., Gloor, M., Mikaloff Fletcher, S.E., Doney, S.C., Dutkiewicz, S., Follows, M.J., Gerber, M., Jacobson, A.R., Joos, F., Lindsay, K., Menemenlis, D., Mouchet, A., Müller, S.A., Sarmiento, J.L., Takahashi, T., 2009. Oceanic sources, sinks, and transport of atmospheric CO2. Global Biogeochemical Cycles 23. https://doi.org/10.1029/2008GB003349

Harris, C.R., Millman, K.J., Walt, S.J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M.H. van, Brett, M., Haldane, A., Río, J.F. del, Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hauck, J., Zeising, M., Le Quéré, C., Gruber, N., Bakker, D.C.E., Bopp, L., Chau, T.T.T., Gürses, Ö., Ilyina, T., Landschützer, P., Lenton, A., Resplandy, L., Rödenbeck, C., Schwinger, J., Séférian, R., 2020. Consistency and Challenges in the Ocean Carbon Sink Estimate for the Global Carbon Budget. Frontiers in Marine Science 7.

Hauri, C., Gruber, N., McDonnell, A.M.P., Vogt, M., 2013. The intensity, duration, and severity of low aragonite saturation state events on the California continental shelf. Geophysical Research Letters 40, 3424–3428. https://doi.org/10.1002/grl.50618

He, W., Xie, X., Mei, Y., Wan, S., Zhao, S., 2021. Decreasing predictability as a precursor indicator for abrupt climate change. Clim Dyn 56, 3899–3908. https://doi.org/10.1007/s00382-021-05676-1

Hoyer, S., Hamman, J., 2017. xarray: N-D labeled arrays and datasets in Python. Journal of Open Research Software 5. https://doi.org/10.5334/jors.148

Iida, Y., Takatani, Y., Kojima, A., Ishii, M., 2021. Global trends of ocean CO2 sink and ocean acidification: an observation-based reconstruction of surface ocean inorganic carbon variables. J Oceanogr 77, 323–358. https://doi.org/10.1007/s10872-020-00571-5

Intergovernmental Panel On Climate Change (Ipcc), 2022. The Ocean and Cryosphere in a Changing Climate: Special Report of the Intergovernmental Panel on Climate Change, 1st ed. Cambridge University Press. https://doi.org/10.1017/9781009157964

IOCCG, 2000. Remote sensing of ocean colour in coastal, and other optically-complex, waters. (Report). International Ocean Colour Coordinating Group (IOCCG). https://doi.org/10.25607/OBP-95

Jones, S.D., Le Quéré, C., Rödenbeck, C., Manning, A.C., Olsen, A., 2015. A statistical gap-filling method to interpolate global monthly surface ocean carbon dioxide data. J Adv Model Earth Syst 7, 1554–1575. https://doi.org/10.1002/2014MS000416

Jönsson, B.F., Follett, C.L., Bien, J., Dutkiewicz, S., Hyun, S., Kulk, G., Forget, G.L., Müller, C., Racault, M.-F., Hill, C.N., Jackson, T., Sathyendranath, S., 2023. Using Probability Density Functions to Evaluate Models (PDFEM, v1.0) to compare a biogeochemical model with satellite-derived chlorophyll. Geosci. Model Dev. 16, 4639–4657. https://doi.org/10.5194/gmd-16-4639-2023

Khatiwala, S., Tanhua, T., Mikaloff Fletcher, S., Gerber, M., Doney, S.C., Graven, H.D., Gruber, N., McKinley, G.A., Murata, A., Ríos, A.F., Sabine, C.L., 2013. Global ocean storage of anthropogenic carbon. Biogeosciences 10, 2169–2191. https://doi.org/10.5194/bg-10-2169-2013

Land, P.E., Shutler, J.D., Findlay, H.S., Girard-Ardhuin, F., Sabia, R., Reul, N., Piolle, J.-F., Chapron, B., Quilfen, Y., Salisbury, J., Vandemark, D., Bellerby, R., Bhadury, P., 2015. Salinity from Space

Unlocks Satellite-Based Assessment of Ocean Acidification. Environ. Sci. Technol. 49, 1987–1994. https://doi.org/10.1021/es504849s

Landschützer, P., Gruber, N., Bakker, D.C.E., 2016. Decadal variations and trends of the global ocean carbon sink. Global Biogeochemical Cycles 30, 1396–1417. https://doi.org/10.1002/2015GB005359

Landschützer, P., Gruber, N., Bakker, D.C.E., Schuster, U., Nakaoka, S., Payne, M.R., Sasse, T.P., Zeng, J., 2013. A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink. Biogeosciences 10, 7793–7815. https://doi.org/10.5194/bg-10-7793-2013

Landschützer, P., Gruber, N., Bakker, D.C.E., Stemmler, I., Six, K.D., 2018. Strengthening seasonal marine CO2 variations due to increasing atmospheric CO2. Nature Clim Change 8, 146–150. https://doi.org/10.1038/s41558-017-0057-x

Le Quéré, C., Rödenbeck, C., Buitenhuis, E.T., Conway, T.J., Langenfelds, R., Gomez, A., Labuschagne, C., Ramonet, M., Nakazawa, T., Metzl, N., Gillett, N., Heimann, M., 2007. Saturation of the Southern Ocean CO $_2$ Sink Due to Recent Climate Change. Science 316, 1735–1738. https://doi.org/10.1126/science.1136188

Liu, W.T., Xie, X., 2017. Space Observation of Carbon Dioxide Partial Pressure at Ocean Surface. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10, 5472–5484. https://doi.org/10.1109/JSTARS.2017.2766138

Liu, Q., He, W., Gu, B., Jiang, Y., 2017. Detecting abrupt dynamic change based on changes in the fractal properties of spatial images. Theor Appl Climatol 130, 435–442. https://doi.org/10.1007/s00704-016-1889-4

Liu, Q., He, W., Xie, X., Mei, Y., Sun, H., Boers, N., 2024. Early warning signal of abrupt change in sea level pressure based on changing spectral exponent. Chaos, Solitons & Fractals 187, 115350. https://doi.org/10.1016/j.chaos.2024.115350

Maaten, L. van der, Hinton, G., 2008. Visualizing Data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

McKinley, G.A., Fay, A.R., Eddebbar, Y.A., Gloege, L., Lovenduski, N.S., 2020. External Forcing Explains Recent Decadal Variability of the Ocean Carbon Sink. AGU Advances 1, e2019AV000149. https://doi.org/10.1029/2019AV000149

Mei, Y., He, W., Xie, X., Wan, S., Gu, B., 2024. Increasing Long-Term Memory as an Early Warning Signal for a Critical Transition. Journal of Climate 37, 487–504. https://doi.org/10.1175/JCLI-D-22-0263.1

Ono, T., Saino†, T., Kurita, N., Sasaki, K., 2004. Basin-scale extrapolation of shipboard pCO $_2$ data by using satellite SST and Chl *a*. International Journal of Remote Sensing 25, 3803–3815. https://doi.org/10.1080/01431160310001657515

Pérez, F.F., Mercier, H., Vázquez-Rodríguez, M., Lherminier, P., Velo, A., Pardo, P.C., Rosón, G., Ríos, A.F., 2013. Atlantic Ocean CO2 uptake reduced by weakening of the meridional overturning circulation. Nature Geosci 6, 146–152. https://doi.org/10.1038/ngeo1680

Resplandy, L., Keeling, R.F., Rödenbeck, C., Stephens, B.B., Khatiwala, S., Rodgers, K.B., Long, M.C., Bopp, L., Tans, P.P., 2018. Revision of global carbon fluxes based on a reassessment of oceanic and riverine carbon transport. Nature Geosci 11, 504–509. https://doi.org/10.1038/s41561-018-0151-3

Ritter, R., Landschützer, P., Gruber, N., Fay, A.R., Iida, Y., Jones, S., Nakaoka, S., Park, G.-H., Peylin, P., Rödenbeck, C., Rodgers, K.B., Shutler, J.D., Zeng, J., 2017. Observation-Based Trends of the Southern Ocean Carbon Sink. Geophysical Research Letters 44, 12,339-12,348. https://doi.org/10.1002/2017GL074837

Rödenbeck, C., Bakker, D.C.E., Gruber, N., Iida, Y., Jacobson, A.R., Jones, S., Landschützer, P., Metzl, N., Nakaoka, S., Olsen, A., Park, G.-H., Peylin, P., Rodgers, K.B., Sasse, T.P., Schuster, U., Shutler, J.D., Valsala, V., Wanninkhof, R., Zeng, J., 2015. Data-based estimates of the ocean carbon sink variability – first results of the Surface Ocean pCO_2 Mapping intercomparison (SOCOM). Biogeosciences 12, 7251–7278. https://doi.org/10.5194/bg-12-7251-2015

Rödenbeck, C., DeVries, T., Hauck, J., Le Quéré, C., Keeling, R.F., 2022. Data-based estimates of interannual sea–air CO_2 flux variations 1957–2020 and their relation to environmental drivers. Biogeosciences 19, 2627–2652. https://doi.org/10.5194/bg-19-2627-2022

Rödenbeck, C., Keeling, R.F., Bakker, D.C.E., Metzl, N., Olsen, A., Sabine, C., Heimann, M., 2013. Global surface-ocean p^{CO_2} and sea–air CO_2 flux variability from an observation-driven ocean mixed-layer scheme. Ocean Science 9, 193–216. https://doi.org/10.5194/os-9-193-2013

Rodgers, K.B., Schwinger, J., Fassbender, A.J., Landschützer, P., Yamaguchi, R., Frenzel, H., Stein, K., Müller, J.D., Goris, N., Sharma, S., Bushinsky, S., Chau, T., Gehlen, M., Gallego, M.A., Gloege, L., Gregor, L., Gruber, N., Hauck, J., Iida, Y., Ishii, M., Keppler, L., Kim, J., Schlunegger, S., Tjiputra, J., Toyama, K., Vaittinada Ayar, P., Velo, A., 2023. Seasonal Variability of the Surface Ocean Carbon Cycle: A Synthesis. Global Biogeochemical Cycles 37, e2023GB007798. https://doi.org/10.1029/2023GB007798

Rohr, T., Richardson, A.J., Lenton, A., Chamberlain, M.A., Shadwick, E.H., 2023. Zooplankton grazing is the largest source of uncertainty for marine carbon cycling in CMIP6 models. Commun Earth Environ 4, 1–22. https://doi.org/10.1038/s43247-023-00871-w

Sabine, C.L., Feely, R.A., Gruber, N., Key, R.M., Lee, K., Bullister, J.L., Wanninkhof, R., Wong, C.S., Wallace, D.W.R., Tilbrook, B., Millero, F.J., Peng, T.-H., Kozyr, A., Ono, T., Rios, A.F., 2004. The Oceanic Sink for Anthropogenic CO2. Science 305, 367–371. https://doi.org/10.1126/science.1097403

Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., Izquierdo-Verdiguier, E., Muñoz-Marí, J., Mosavi, A., Camps-Valls, G., 2020. Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. Information Fusion 63, 256–272. https://doi.org/10.1016/j.inffus.2020.07.004

Sauzède, R., Johnson, J.E., Claustre, H., Camps-Valls, G., Ruescas, A.B., 2020. ESTIMATION OF OCEANIC PARTICULATE ORGANIC CARBON WITH MACHINE LEARNING. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. V-2–2020, 949–956. https://doi.org/10.5194/isprs-annals-V-2-2020-949-2020

Séférian, R., Berthet, S., Yool, A., Palmiéri, J., Bopp, L., Tagliabue, A., Kwiatkowski, L., Aumont, O., Christian, J., Dunne, J., Gehlen, M., Ilyina, T., John, J.G., Li, H., Long, M.C., Luo, J.Y., Nakano, H., Romanou, A., Schwinger, J., Stock, C., Santana-Falcón, Y., Takano, Y., Tjiputra, J., Tsujino, H., Watanabe, M., Wu, T., Wu, F., Yamamoto, A., 2020. Tracking Improvement in Simulated Marine

Biogeochemistry Between CMIP5 and CMIP6. Curr Clim Change Rep 6, 95–119. https://doi.org/10.1007/s40641-020-00160-0

Shutler, J.D., Gruber, N., Findlay, H.S., Land, P.E., Gregor, L., Holding, T., Sims, R., Green, H., Piolle, J.-F., Chapron, B., Sathyendranath, S., Rousseaux, C.S., Donlon, C., Cooley, S., Turner, J., Valauri-Orton, A., Lowder, K., Widdicombe, S., Newton, J., Sabia, R., Rio, M.-H., Gaultier, L., 2024. The increasing importance of satellite observations to assess the ocean carbon sink and ocean acidification. Earth-Science Reviews 104682. https://doi.org/10.1016/j.earscirev.2024.104682

Shutler, J.D., Wanninkhof, R., Nightingale, P.D., Woolf, D.K., Bakker, D.C., Watson, A., Ashton, I., Holding, T., Chapron, B., Quilfen, Y., Fairall, C., Schuster, U., Nakajima, M., Donlon, C.J., 2020. Satellites will address critical science priorities for quantifying ocean carbon. Frontiers in Ecol & Environ 18, 27–35. https://doi.org/10.1002/fee.2129

Siddique, T., Mahmud, M., Keesee, A., Ngwira, C., Connor, H., 2022. A Survey of Uncertainty Quantification in Machine Learning for Space Weather Prediction. Geosciences 12, 27. https://doi.org/10.3390/geosciences12010027

Solidoro, C., Bandelj, V., Barbieri, P., Cossarini, G., Fonda Umani, S., 2007. Understanding dynamic of biogeochemical properties in the northern Adriatic Sea by using self-organizing maps and k-means clustering. Journal of Geophysical Research: Oceans 112. https://doi.org/10.1029/2006JC003553

Sonnewald, M., Dutkiewicz, S., Hill, C., Forget, G., 2020. Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. Science Advances 6, eaay4740. https://doi.org/10.1126/sciadv.aay4740

Sonnewald, M., Wunsch, C., Heimbach, P., 2019. Unsupervised Learning Reveals Geography of Global Ocean Dynamical Regions. Earth and Space Science 6, 784–794. https://doi.org/10.1029/2018EA000519

Takahashi, T., Sutherland, S.C., Kozyr, A., 2017. LDEO Database (Version 2019): Global Ocean Surface Water Partial Pressure of CO2 Database: Measurements Performed During 1957-2019 (NCEI Accession 0160492). https://doi.org/10.3334/CDIAC/OTG.NDP088(V2015)

Takahashi, T., Sutherland, S.C., Wanninkhof, R., Sweeney, C., Feely, R.A., Chipman, D.W., Hales, B., Friederich, G., Chavez, F., Sabine, C., Watson, A., Bakker, D.C.E., Schuster, U., Metzl, N., Yoshikawa-Inoue, H., Ishii, M., Midorikawa, T., Nojiri, Y., Körtzinger, A., Steinhoff, T., Hoppema, M., Olafsson, J., Arnarson, T.S., Tilbrook, B., Johannessen, T., Olsen, A., Bellerby, R., Wong, C.S., Delille, B., Bates, N.R., de Baar, H.J.W., 2009. Climatological mean and decadal change in surface ocean pCO2, and net sea–air CO2 flux over the global oceans. Deep Sea Research Part II: Topical Studies in Oceanography, Surface Ocean CO2 Variability and Vulnerabilities 56, 554–577. https://doi.org/10.1016/j.dsr2.2008.12.009

Vafaei, B., Ezam, M., Saghaei, A., Bidokhti, A.A., 2022. Automatic identification and tracking of meso-scale eddies in the Persian Gulf using the pattern mining approach. Int. J. Environ. Sci. Technol. 19, 6011–6022. https://doi.org/10.1007/s13762-021-03779-0

Verdy, A., Mazloff, M.R., 2017. A data assimilating model for estimating S outhern O cean biogeochemistry. JGR Oceans 122, 6968–6988. https://doi.org/10.1002/2016JC012650

Wang, G., Dai, M., Shen, S.S.P., Bai, Y., Xu, Y., 2014. Quantifying uncertainty sources in the gridded data of sea surface CO $_2$ partial pressure. JGR Oceans 119, 5181–5189. https://doi.org/10.1002/2013JC009577

Wang, Shuai, Wang, P., Qi, Q., Wang, Siyu, Meng, X., Kan, H., Zhu, S., Zhang, H., 2023. Improved estimation of particulate matter in China based on multisource data fusion. Science of The Total Environment 869, 161552. https://doi.org/10.1016/j.scitotenv.2023.161552

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

Wanninkhof, R., Park, G.-H., Takahashi, T., Sweeney, C., Feely, R., Nojiri, Y., Gruber, N., Doney, S.C., McKinley, G.A., Lenton, A., Le Quéré, C., Heinze, C., Schwinger, J., Graven, H., Khatiwala, S., 2013. Global ocean carbon uptake: magnitude, variability and trends. Biogeosciences 10, 1983–2000. https://doi.org/10.5194/bg-10-1983-2013

Woolf, D.K., Land, P.E., Shutler, J.D., Goddijn-Murphy, L.M., Donlon, C.J., 2016. On the calculation of air-sea fluxes of CO $_2$ in the presence of temperature and salinity gradients. JGR Oceans 121, 1229–1248. https://doi.org/10.1002/2015JC011427

Yang, X., Wynn-Edwards, C.A., Strutton, P.G., Shadwick, E.H., 2024. Drivers of Air-Sea CO2 Flux in the Subantarctic Zone Revealed by Time Series Observations. Global Biogeochemical Cycles 38, e2023GB007766. https://doi.org/10.1029/2023GB007766

Zeng, J., Nojiri, Y., Landschützer, P., Telszewski, M., Nakaoka, S., 2014. A Global Surface Ocean fCO2 Climatology Based on a Feed-Forward Neural Network. Journal of Atmospheric and Oceanic Technology 31, 1838–1849. https://doi.org/10.1175/JTECH-D-13-00137.1