Reconstruction and Spatiotemporal Analysis of Global Surface Ocean pCO_2 Considering Sea Area Characteristics

Huisheng Wu, Yunlong Ji, Lejie Wang, Xiaoke Liu, Wenliang Zhou, Long Cui, Yang Wang, Min Liu and Zhuang Li

College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, Shandong 266580, China;

Correspondence to: Huisheng Wu (wuhuisheng@upc.edu.cn)

Abstract:

The partial pressure of carbon dioxide (pCO_2) on the surface of the ocean is crucial for quantifying and evaluating the ocean carbon budget. Insufficient consideration of the effects at the sea area scale makes it difficult to comprehensively evaluate the spatiotemporal distribution characteristics and variation patterns of pCO_2. This study constructed a pCO_2 evaluation dataset based on LDEO measurement data and multi-source data. After conducting correlation testing on a global, far sea, and near sea scale, an ocean surface pCO_2 evaluation model was constructed using multiple linear regression, convolutional neural network, gated recurrent unit, long short-term memory network, generalized additive model, extreme gradient boosting, least squares boosting, and random forest. Performance evaluation indicates that the random-forest model consistently achieves the best accuracy across all spatial scales, yielding a global RMSE of 6.123 μatm and an R^2 of 0.986. In the open ocean, RMSE decreases to 4.699 μatm and R^2 rises to 0.988, whereas in coastal waters RMSE increases to 8.044 μatm and R^2 declines to 0.972. Based on this, the annual sea surface pCO_2 distribution of 0.25°× 0.25°from 2000 to 2019 was reconstructed. The reconstructed field shows a typical equatorial high/polar low pattern, as well as an overall upward trend consistent with independent observations, with acceleration particularly evident in specific regions of subtropical coastal oceans.

Keywords:

Surface ocean pCO_2; Global Oceans; Machine learning; Spatiotemporal changes; Random forest

Synopsis：

This study reconstructs global ocean surface pCO_2 (2000–2019) using multi-source data and machine learning, identifying RF as the optimal model and revealing equatorial-high/polar-low patterns with rising trends.

1. Introduction

The partial pressure of carbon dioxide on the surface of the ocean (pCO_2) is an important indicator for measuring the exchange of CO_2 between the ocean and the atmosphere, and can evaluate the contribution of the ocean's carbon absorption and storage capacity to the global carbon cycle(Falkowski et al., 2000; Jain, 2022) .

Numerous scholars have conducted research on pCO_2 estimation and distribution reconstruction by combining satellite remote sensing data and machine learning algorithms. In the study of sea surface pCO_2 in local sea areas, Telszewski et al. reconstructed the distribution of pCO_2 in the North Atlantic using self-organizing neural

networks (Telszewski et al., 2009); Landschützer et al. reconstructed the distribution map of Atlantic sea surface pCO_2 using self-organizing map feedforward neural network method (Landschützer et al., 2013). Chierici et al. evaluated the feasibility of jointly estimating sea surface pCO_2 in Antarctica and the Pacific region using ship borne measured data and remote sensing data (Chierici et al., 2011). Nakaoka et al. established a nonlinear relationship between sea surface pCO_2 and multiple parameters based on self-organizing neural networks, and reconstructed the spatiotemporal variation of sea surface pCO_2 in the North Pacific (Nakaoka et al., 2013). Marrec et al. used multiple linear regression to estimate the sea surface pCO_2 in the waters of the Northwest European continental shelf (Wang et al., 2021). Gregor et al. proposed methods such as support vector regression and random forest regression to reconstruct the Southern Ocean surface pCO_2 (Gregor et al., 2017); Wang et al. reconstructed the distribution of pCO_2 on the surface of the Southern Ocean using correlation analysis and feed forward neural networks (Marrec et al., 2015). Lohrenz et al. reconstructed the sea surface pCO_2 in the northern Gulf of Mexico using regression tree algorithm (Lohrenz et al., 2021); Chen et al. compared the performance of various methods in estimating surface pCO_2 in the Gulf of Mexico (Chen et al., 2019); Fu et al. applied cubist models to estimate pCO_2 on the surface of the Gulf of Mexico (Fu et al., 2020). Zhang et al. constructed a sea surface pCO_2 regression model for the Baltic Sea region (Zhang et al., 2021). In the study of global ocean surface pCO_2, Landschützer et al. expanded the research scope to the global level, reconstructed the pCO_2 distribution map from 1998 to 2011, and further extended it to 1982 to 2011 (Landschützer et al., 2014; Landschützer et al., 2016). Gregor et al. reconstructed the pCO_2 distribution using various nonlinear regression methods (Gregor et al., 2019). Zhong et al. used generalized regression neural network and stepwise regression algorithm to construct the pCO_2 distribution map (Guorong et al., 2020), and combined stepwise regression algorithm and feed forward neural network, constructed a 1°× 1°pCO_2 distribution map from 1992 to 2019 according to the 11 biogeochemical provinces defined by the self-organizing map method (Zhong et al., 2020).

By summarizing previous research, the key limitations of current sea surface pCO_2 are:

(1) Insufficient Consideration of Spatial Heterogeneity

Most existing studies either focus on a single local sea area (e.g., the North Atlantic, Gulf of Mexico, Baltic Sea) or adopt a unified global modeling framework, neglecting the significant differences in environmental conditions, driving factors, and pCO_2 variation characteristics between far sea areas and near sea areas.

To address this issue, our study constructs a multi-scale analysis framework covering the global ocean, far sea areas (water depth > 200 meters), and near sea areas (water depth ≤ 200 meters). The research areas are divided into far sea areas and near sea areas based on water depth, and scale-specific pCO_2 evaluation models are established. For the environmentally stable far sea areas, we emphasize capturing long-term temporal dependencies and signals of large-scale hydrological and biological processes. For near sea areas affected by various complex factors, we incorporate region-specific driving factors and optimize the model structure to adapt to high variability. This targeted approach effectively improves the fitting accuracy and adaptability of the models in different sea area types.

(2) Inadequate Adaptability Between Models and Driving Factors

Existing studies mostly adopt fixed model structures or globally unified combinations of driving factors, failing to fully consider the requirements of environmental complexity differences in different sea areas for model adaptability. Additionally, the selection of driving factors lacks targeting, making it difficult for the models to accurately capture the core impact mechanisms of pCO_2 in different regions.

We resolve this limitation through the comprehensive optimization of models and driving factors: we compared eight machine learning models and identified the Random Forest (RF) model as the optimal model across all scales. Its advantage in capturing complex nonlinear relationships enables it to adapt to the environmental characteristics of different sea areas. Meanwhile, based on Spearman correlation analysis and the SHAP (SHapley Additive

exPlanations) method, we screened key driving factors for each scale (e.g., Total alkalinity in sea water (talk) serves as the secondary key factor at the global scale, while the contribution rate of mole concentration of dissolved molecular oxygen in sea water (O_2) significantly increases in near sea areas), ensuring the rationality and targeting of driving factor selection.

(3) Low Reconstruction Resolution

Some existing studies lack the overall processing of spatiotemporal differences in multi-source data, resulting in low spatial resolution of pCO_2 reconstruction products (mostly 1°×1° or coarser), which makes it difficult to accurately reflect the spatiotemporal variation characteristics of pCO_2 within small scales.

We address this limitation through high-resolution and high-precision reconstruction strategies: by processing multi-source data (including strict data matching, outlier handling, and data balancing strategies), we reconstructed the annual pCO_2 distribution with a high resolution of 0.25°×0.25° from 2000 to 2019. The results demonstrate that the accuracy of pCO_2 reconstruction is significantly improved compared with existing studies.

2. Methodology

2.1 Research Area

The global ocean, excluding the perennial ice-covered waters in the core area of the Arctic Ocean and the permanently frozen areas around the Antarctic continent, has a total area of 336 million square kilometers, accounting for approximately 92.8% of the global ocean surface area. This research focuses on the 0–10-meter water layer in the ocean surface, which is a critical interface for air sea exchange. Due to the complex types of water bodies, sea surface pCO_2 is influenced by various factors. The global ocean was divided into research area scales based on water depth, identifying the areas beyond the continental shelf (water depth > 200 meters) as far sea areas and the areas within the range (water depth ≤ 200 meters) as near sea areas.

2.2 Data sources

2.2.1 Actual measurement data

The measured data of pCO_2 is sourced from Global Surface pCO_2 (LDEO) Database V2019 (OCADS - Global Surface pCO_2 (LDEO) Database (noaa. gov)). This dataset covers 14.2 million measured data from 1957 to 2019 using the equalizer CO_2 analyzer system in the global ocean. The dataset provides various types of sea surface pCO_2 measured data. This study selected ocean surface pCO_2 values measured at actual temperatures from 2000 to 2019, which can truly reflect the pCO_2 level at the time of measurement.

2.2.2 Other data

A total of 25 potential influencing factors were selected for the study (Table 1), and their abbreviations are used for convenience. These data are divided into three types of sources: in-situ observations, satellite observations, and numerical models, with good spatiotemporal resolution and coverage, providing reliable data sources for research.

Table 1. Specific information about influencing factors (sort based on its resolution and name)

Variable name	Abbreviation	Spatial resolution	Temporal resolution	Data type	DOI
Mass concentration of chlorophyll a in sea water	chl	0.036	Daily	Satellite observations	https://doi.org/10.48670/moi-00281
Volume attenuation coefficient of	kd490	0.036	Daily	Satellite	https://doi.org/10.48670/moi-002

downwelling radiative flux in sea water				observations	81
Ocean mixed layer thickness defined by sigma theta	mlotst*	0.083	Daily	Numerical models	https://doi.org/10.48670/moi-00021
Sea water salinity	so	0.083	Daily	Numerical models	https://doi.org/10.48670/moi-00021
Sea water potential temperature	thetao	0.083	Daily	Numerical models	https://doi.org/10.48670/moi-00021
Eastward sea water velocity	uo	0.083	Daily	Numerical models	https://doi.org/10.48670/moi-00021
Northward sea water velocity	vo	0.083	Daily	Numerical models	https://doi.org/10.48670/moi-00021
Sea surface height above geoid	zos	0.083	Daily	Numerical models	https://doi.org/10.48670/moi-00021
Sea surface density	dos	0.125	Daily	In-situ observations Satellite observations	https://doi.org/10.48670/moi-00051
Sea surface salinity	sos	0.125	Daily	In-situ observations Satellite observations	https://doi.org/10.48670/moi-00051
Mole concentration of nitrate in sea water	no_3	0.25	Daily	Numerical models	https://doi.org/10.48670/moi-00019
Mole concentration of dissolved molecular oxygen in sea water	o_2	0.25	Daily	Numerical models	https://doi.org/10.48670/moi-00019
Mole concentration of phosphate in sea water	po_4	0.25	Daily	Numerical models	https://doi.org/10.48670/moi-00019
Mole concentration of silicate in sea water	si	0.25	Daily	Numerical models	https://doi.org/10.48670/moi-00019
Surface geostrophic eastward sea water velocity	ugos	0.25	Daily	Numerical models In-situ observations Satellite observations	https://doi.org/10.48670/mds-00327
Surface geostrophic northward sea water velocity	vgos	0.25	Daily	Numerical models In-situ observations Satellite observations	https://doi.org/10.48670/mds-00327
Ocean mixed layer thickness	mlotst	0.25	Weekly	In-situ observations	https://doi.org/10.48670/moi-00052

Sea water temperature	to	0.25	Weekly	Satellite observations In-situ observations	https://doi.org/10.48670/moi-00052
Eastward wind	uwind	0.25	Monthly	Satellite observations	https://doi.org/10.48670/moi-00181
Northward wind	vwind	0.25	Monthly	Satellite observations	https://doi.org/10.48670/moi-00181
Aragonite saturation state in sea water	ar	1	Monthly	In-situ observations	https://doi.org/10.48670/moi-00047
Calcite saturation state in sea water	ca	1	Monthly	In-situ observations	https://doi.org/10.48670/moi-00047
Sea water ph reported on total scale	ph	1	Monthly	In-situ observations	https://doi.org/10.48670/moi-00047
Total alkalinity in sea water	talk	1	Monthly	In-situ observations	https://doi.org/10.48670/moi-00047
Dissolved inorganic carbon in sea water	tco_2	1	Monthly	In-situ observations	https://doi.org/10.48670/moi-00047

2.3 Data Processing

2.3.1 Data Matching

To reduce the impact of spatial and temporal resolution differences in multi-source data, we adopted a dual matching strategy to process pCO_2 measured data and potential influencing factors. In the temporal dimension, influencing variables were first aligned with the in-situ pCO_2 observations; temporal gaps were subsequently infilled via nearest-time interpolation to ensure chronological consistency. In the spatial dimension, data points were aligned through precise geographic coordinate matching algorithms, and nearest neighbor interpolation was used to supplement missing points to improve spatial accuracy. After matching, each point contains the measured value of pCO_2, environmental variables, and corresponding spatiotemporal information (year, month, lat, lon).

2.3.2 Analysis of Outliers

The study conducted quality control on the matched data by removing missing values generated during the matching process. According to data statistics and previous research experience (Wu et al., 2024), measured data below 200 μatm and above 600 μatm are classified as outliers. The spatial distribution of outliers is mainly concentrated in coastal areas, reflecting the variability of land sea interaction effects. Outliers are valuable sample data for the study of pCO_2. Through comparative analysis of each route, it was found that many outliers matched the route, and it was determined that their outliers were caused by environmental changes rather than measurement errors. Therefore, valid outliers were retained and only obvious measurement error data were removed. For other environmental variable values, abnormal data was identified and removed based on the 3σcriterion ($\mu \pm 3\sigma$).

2.3.3 Data Balancing

The processed global ocean data was divided into far sea and near sea datasets (Figure 1a, b, c). Statistical analysis shows that the spatial and temporal distribution of data is uneven. Therefore, a 0.25°× 0.25°grid was used for spatial binning, and time binning was performed monthly to construct a spatiotemporal joint binning unit. The

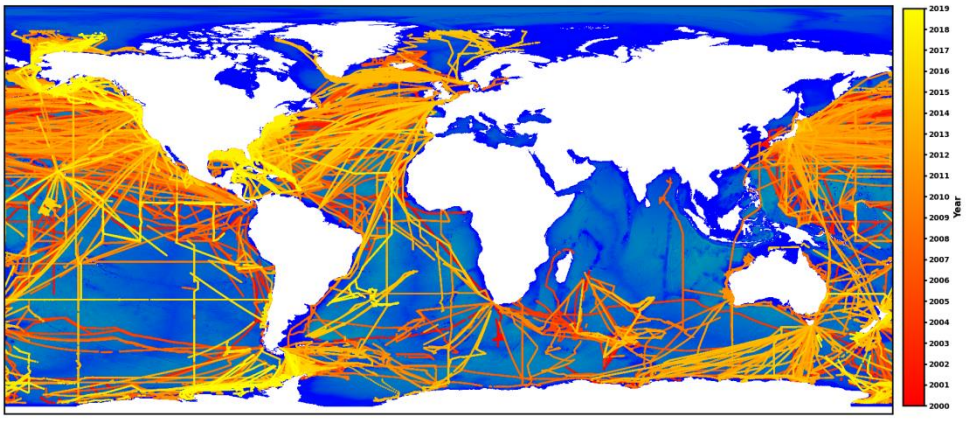
granularity setting of this box not only meets the research accuracy requirements, but also maintains compatibility with the spatiotemporal resolution of multi-source data.

Take the arithmetic mean of the data within each unit as the representative value, with the spatial position represented by the grid center point, and the time calculated as the weighted average based on the distribution of data points (Formula 1). This method effectively balances the data distribution while ensuring accuracy.

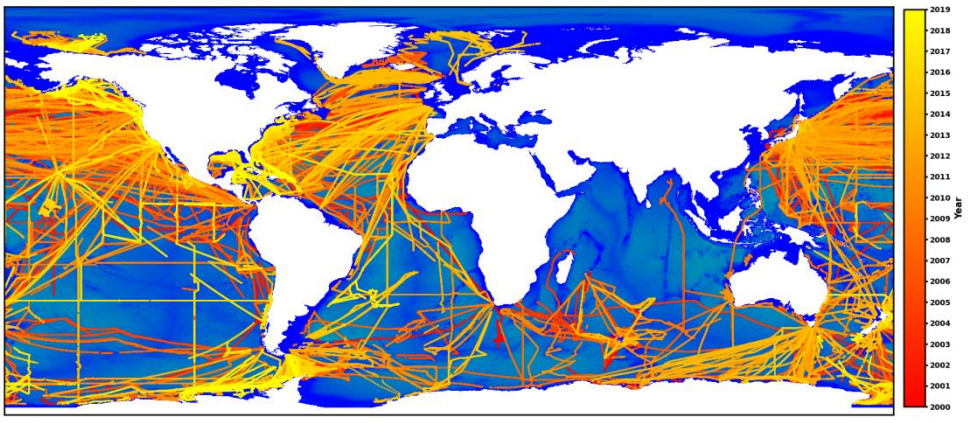
$$t_{avg} = \frac{\sum_{i=1}^{n} w_i t_i}{\sum_{i=1}^{n} w_i} \quad (1)$$

$$w_i = \Delta t_i \quad (2)$$

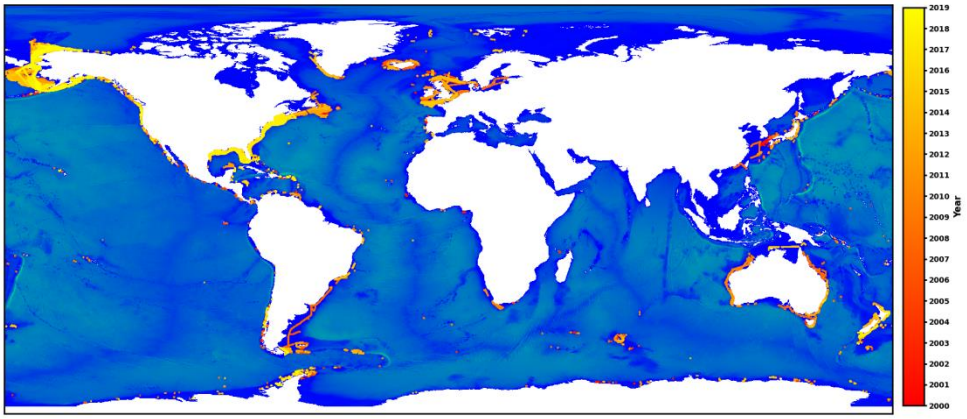
In the formula, t_{avg} is the weighted average time of the spatiotemporal box, n is the total amount of data in the spatiotemporal box, w_i is the weight of the i-th data point, t_i is the time of the i-th data point, and Δt_i is the sampling time interval between the i-th data point and the previous point. After data balancing processing, the dataset for this study was finally constructed, laying a solid data foundation for the construction of multi-scale models.



(a)



(b)



(c)

Figure 1. The spatiotemporal distribution of datasets at different scales. (a)Global spatial distribution of ocean data. (b)Spatial distribution of data in far sea areas. (c)Spatial distribution of data in near sea areas.

2.4 Spearman correlation analysis of pCO_2 drivers

The potential influencing factors involved do not fully follow a normal distribution, and there is a non-linear relationship between pCO_2. Therefore, selecting appropriate correlation indicators is particularly crucial. The Spearman correlation coefficient can effectively reveal the correlation between data (Formula 3).

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \quad (3)$$

In the formula, ρ represents the correlation coefficient, D represents the level difference of the variable, and n represents the sample size of the variable. The range of values for ρ is between -1 and 1, where -1 indicates a complete negative correlation between the influencing factors and pCO_2, 1 indicates a complete positive correlation, and 0 indicates no correlation.

2.5 Model selection

To evaluate the modeling ability of different algorithms for pCO_2, we constructed eight comparative models at different research regions, including multiple linear regression (MLR),convolutional neural network (CNN), gated recurrent unit (GRU), long short term memory (LSTM),generalized additive models (GAM), extreme gradient boosting (XGBoost), least squares boosting (LSBoost), and random forest (RF). MLR serves as a baseline that linearly links temperature, salinity and nutrients to sea-surface pCO_2. CNN extracts spatial features via convolution and pooling layers to produce fine-scale pCO_2 distributions, while GRU and LSTM, with their update-reset gates and memory cells, capture long-term temporal dependencies of oceanic periodic changes on pCO_2 for historical-to-future prediction. GAM relaxes the linearity assumption by modeling each predictor's additive nonlinear effect on pCO_2. XGBoost and LSBoost iteratively optimize tree ensembles through gradient boosting or weighted residuals to uncover complex nonlinear relationships between high-dimensional features and pCO_2. Finally, RF constructs and averages many decision trees on random feature subsets, delivering robust pCO_2 estimates for large-scale ocean datasets.

2.6 Performance evaluation

The datasets at different research regions were randomly divided into training, validation, and testing sets in an 8:1:1 ratio. Five statistical methods, Mean Absolute Error (MAE, μatm) – the average absolute difference between predicted and in-situ pCO_2, indicating overall bias; Mean Absolute Percentage Error (MAPE, %)–the relative error scaled by the observed pCO_2, enabling comparison across regions with contrasting background concentrations; Mean Squared Error (MSE, μatm^2) – the squared deviations averaged over all samples, emphasizing larger pCO_2 discrepancies; Root Mean Squared Error (RMSE, μatm) – the square root of MSE, providing a metric in the original pCO_2 units that is sensitive to outliers; Coefficient of Determination (R^2) – the proportion of pCO_2 variance explained by the model, with values approaching unity signifying high predictive skill.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \quad (4)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \quad (5)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \quad (7)$$

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

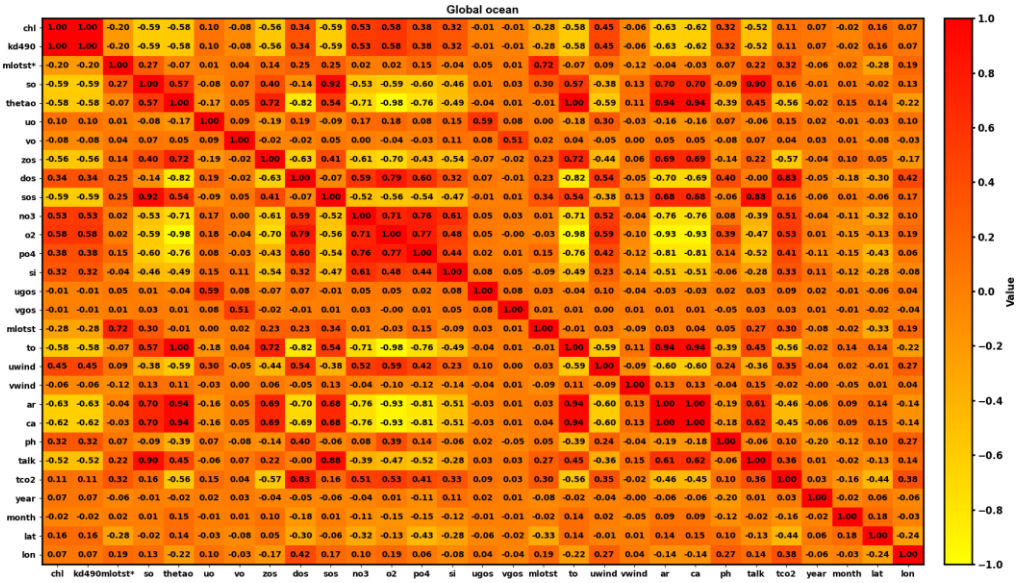
In the formula, n is the number of pCO_2 observations; y_i denotes the in-situ measured pCO_2 (μatm) for the i-th sample, $\hat{y}_i$ is the corresponding model-estimated pCO_2, $\bar{y}_i$ represents the mean of all measured pCO_2 values.

3. Results and discussion

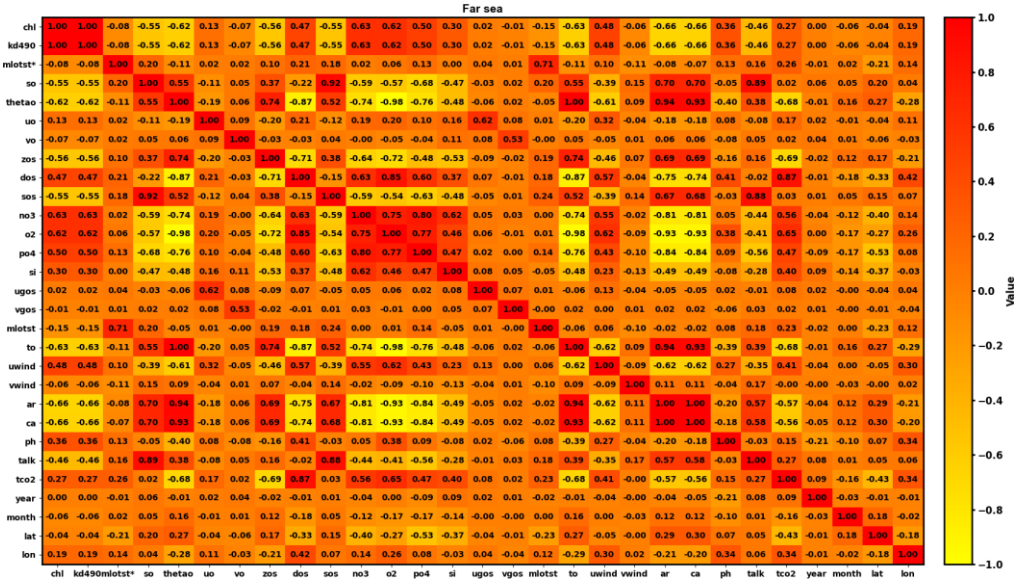
3.1 Correlation detection

3.1.1 Interaction detection

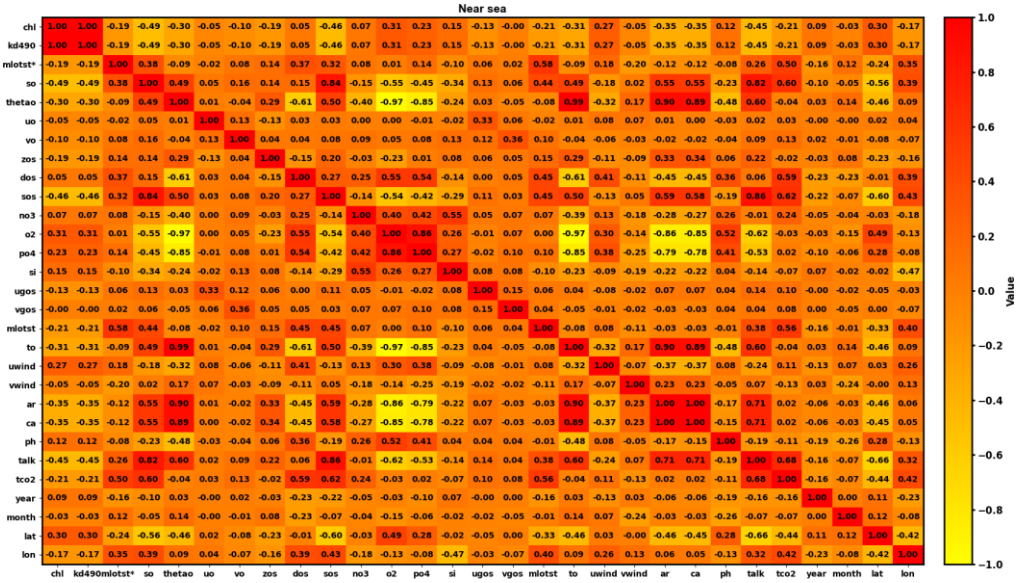
Interactive detection of variables was conducted in global oceans, far sea areas, and near sea areas (Figure 2). The concentration of chlorophyll and the volume attenuation coefficient of downwelling radiative flux have a ρ-value of 1 at all research area scales, indicating collinearity in numerical values. However, they respectively reflect marine biological activity and optical properties, providing comprehensive information for fitting surface pCO_2. The ρ value between the aragonite saturation state in sea water and aragonite in seawater is also 1, and they are positively correlated with the same magnitude of change. This usually stems from chemical equilibrium processes in seawater, where the dissolution and precipitation processes are influenced by similar physical and chemical conditions. The correlation between sea water potential temperature and sea water temperature is extremely high, but their physical meanings are different. The former reflects the equivalent temperature after considering pressure, while the latter reflects the actual temperature. Both can comprehensively capture temperature characteristics and improve the accuracy of surface pCO_2 evaluation.


Global ocean
chl kd490 mlotst* so thetao uo vo zos dos sos no3 o2 po4 si ugos vgos mlotst to uwind vwind ar ca ph talk tco2 year month lat lon
Value
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0

（a）


Far sea
chl kd490 mlotst* so thetao uo vo zos dos sos no3 o2 po4 si ugos vgos mlotst to uwind vwind ar ca ph talk tco2 year month lat lon
Value
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0

（b）


Near sea
chl kd490 mlotst* so thetao uo vo zos dos sos no3 o2 po4 si ugos vgos mlotst to uwind vwind ar ca ph talk tco2 year month lat lon
Value
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0

（c）

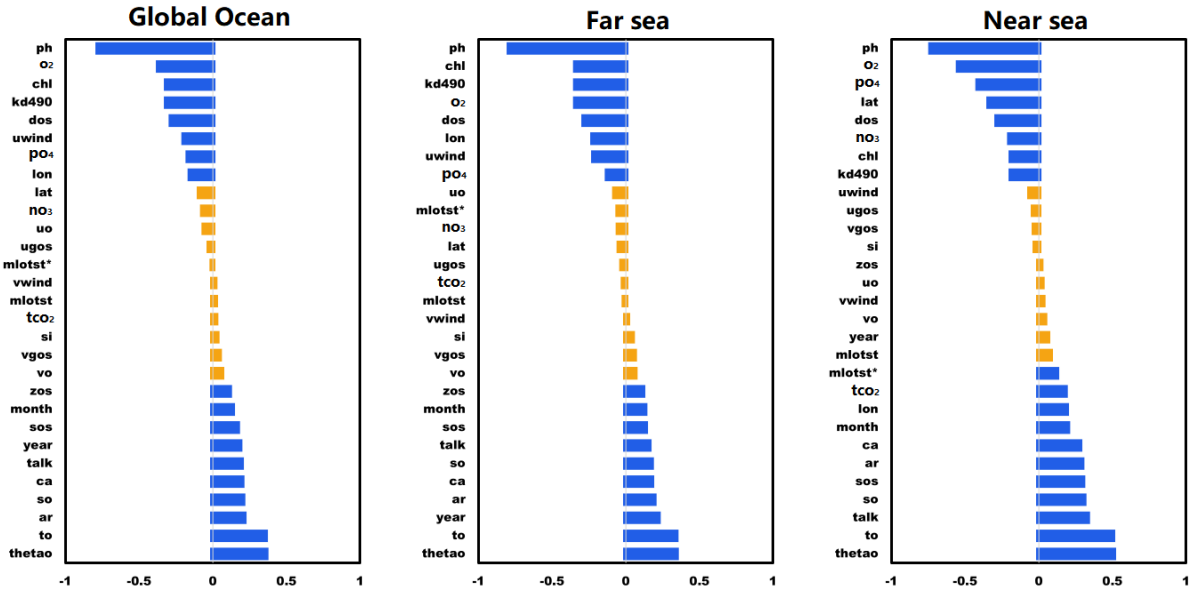
Figure 2. Results of interaction detection between variables at different research area scales. (a)Global Ocean Interaction Detection Results. (b)Interaction detection results in far sea areas. (c)Interactive detection results in near sea areas.

3.1.2 Single factor detection

The correlation between surface pCO_2 and various influencing factors (Figure 3) was analyzed. The results indicate that at different regional scales, there is a significant negative correlation between pCO_2 and ph, meaning that the stronger the acidity of seawater, the higher the surface pCO_2; the stronger the alkalinity, the lower the surface pCO_2. At the same time, surface pCO_2 is significantly positively correlated with temperature. In far sea areas, the negative correlation between pCO_2 and chlorophyll concentration and diffuse reflectance attenuation coefficient is more significant, indicating that it has higher stability and balance in regulating pCO_2. In contrast, the above correlation in near sea areas is weaker due to land-based pollution, human activities, and environmental changes, but the negative correlation between pCO_2 and seawater acidity is stronger. When selecting variables, the study included factors with a p-value greater than 0.1 or less than -0.1 in the screening range to ensure the validity of the results and improve model performance (Table 2). Additionally, SHAP method was used to quantitatively evaluate the contributions of various influencing factors to surface pCO_2 (Ge, Patino, Todisco, & Evans, 2022). There were differences in the contributions of influencing factors at different scales. The ph is the core driving factor at all scales, but its contribution intensity follows a distribution pattern of "far sea areas > global oceans > near sea areas"; The contribution of other factors shows significant regional heterogeneity, such as talk being the second key factor at the global ocean scale, while the contribution rate of o_2 in near sea areas has significantly increased, making ar a region specific factor.

Table 2. Selection results of influencing factors at different area scales

Research scale	Influence factor
Global Ocean	ph、o_2、chl、kd490、dos、uwind、po_4、lon、zos、month、sos、year、talk、ca、so、ar、to、thetao
Far sea	ph、chl、kd490、o_2、dos、lon、uwind、po_4、zos、month、sos、talk、so、ca、ar、year、to、thetao
Near sea	ph、o_2、po_4、lat、dos、no_3、chl、kd490、mlotst*、tco_2、lon、 month、ca、ar、sos、so、talk、to、thetao



（a） （b） （c）

Figure 3. Single factor detection results at different research area scales. (a)**Global ocean single factor detection results. (b)Far sea single factor detection results. (c)Near sea single factor detection results**

3.2 Model construction and evaluation

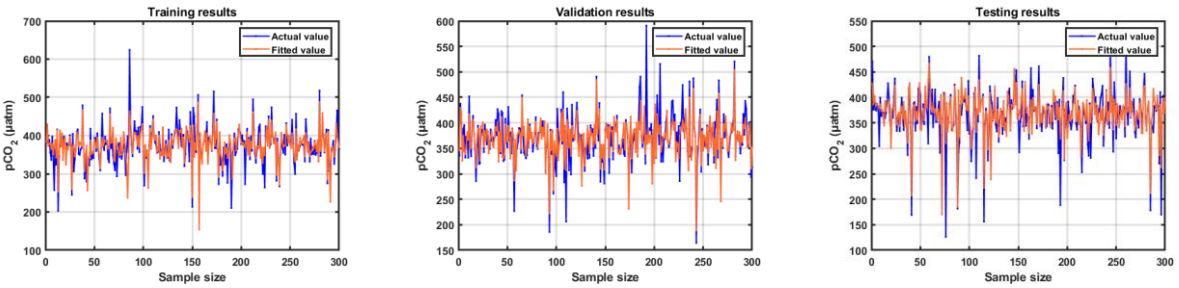
3.2.1 Construction and evaluation of global ocean surface pCO_2 model

Based on the correlation analysis results of the above factors, this study selected key driving factors to construct and evaluate a global sea surface pCO_2 reconstruction model. Owing to the large amount of data, we randomly selected some data from all the fitting results to show the observation performance. Different models exhibit significant performance differences in evaluating surface pCO_2 at the global ocean scale (Figure 4). Specifically, there is a significant gap between the model values of MLR, CNN, and GRU and the true values, especially in the low value (<300 μatm) and high value (>500 μatm) ranges where the fitting effect is poor (Table 3). The deviation is due to the model's insufficient ability to capture nonlinear relationships in complex marine environments, limitations in handling extreme values, and the model's own structure is not sufficient to adapt to complex data features. The LSTM and GAM models have relatively large errors and poor performance, indicating deficiencies in capturing the characteristics of surface pCO_2 changes. When extreme fluctuations occur in surface pCO_2, the fitting ability significantly decreases. The comprehensive performance of XGBoost and LSBoost has significantly improved, with MAE reduced to 15 μatm~18 μatm, RMSE reduced to 25 μatm~30 μatm, and R^2 exceeding 0.7. The effective explanation of multivariate nonlinear relationships and the application of model ensemble strategies have improved the accuracy of the two models within the normal range (300 μatm~500 μatm), but the extreme values processing still needs to be improved. The performance of RF is the best among all models, with MAE reduced to below 4 μatm, RMSE reduced to around 6 μatm, and R^2 reaching above 0.9. It not only achieves accurate fitting in the range of 300 μatm~500 μatm values, but also in the low and high value ranges. The good adaptability of RF to high-dimensional data and a large number of samples makes it perform well in fitting tasks in complex marine environments.

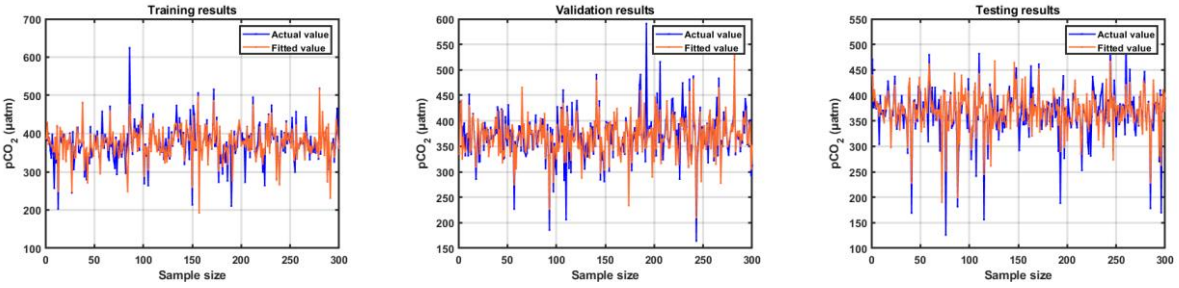
Table 3. Performance parameters of different models in the global ocean

	Model	MAE/μatm	MAPE	MSE/μatm²	RMSE/μatm	R²
Training	RF	3.895	0.011	46.162	6.794	0.983

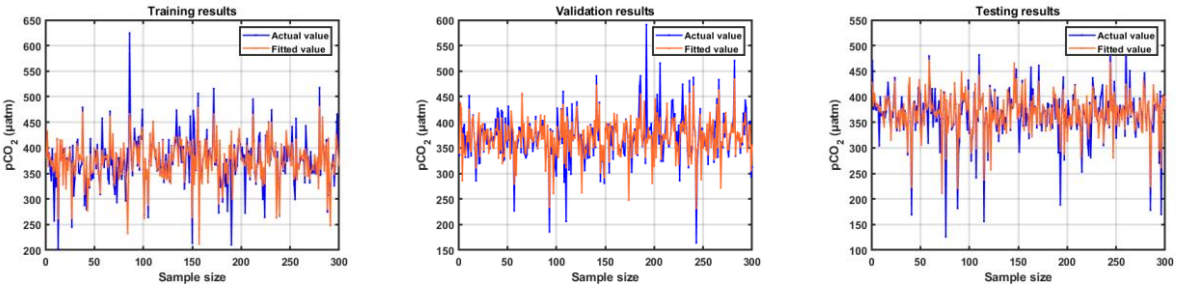
	LSBoost	15.626	0.045	664.186	25.772	0.783
	XGBoost	17.262	0.051	908.879	30.148	0.703
	GAM	19.903	0.058	1398.695	37.399	0.543
	LSTM	18.664	0.055	1430.072	37.816	0.533
	GRU	19.053	0.056	1480.157	38.473	0.516
	CNN	19.903	0.058	1484.621	38.531	0.515
	MLR	19.952	0.058	1615.155	40.189	0.472
Validation	RF	3.902	0.011	46.099	6.790	0.983
	LSBoost	15.604	0.045	661.203	25.714	0.788
	XGBoost	17.255	0.051	910.387	30.173	0.708
	GAM	19.905	0.058	1429.372	37.807	0.541
	LSTM	18.675	0.055	1463.378	38.254	0.529
	GRU	19.059	0.056	1515.286	38.927	0.513
	CNN	19.901	0.058	1520.882	38.999	0.511
	MLR	19.969	0.058	1656.093	40.695	0.468
Testing	RF	3.697	0.010	37.485	6.123	0.986
	LSBoost	15.602	0.045	660.401	25.698	0.785
	XGBoost	17.284	0.051	914.165	30.235	0.703
	GAM	19.916	0.058	1399.851	37.415	0.545
	LSTM	18.690	0.055	1431.489	37.835	0.535
	GRU	19.079	0.056	1483.044	38.510	0.518
	CNN	19.927	0.058	1488.331	38.579	0.516
	MLR	19.982	0.058	1621.378	40.266	0.473



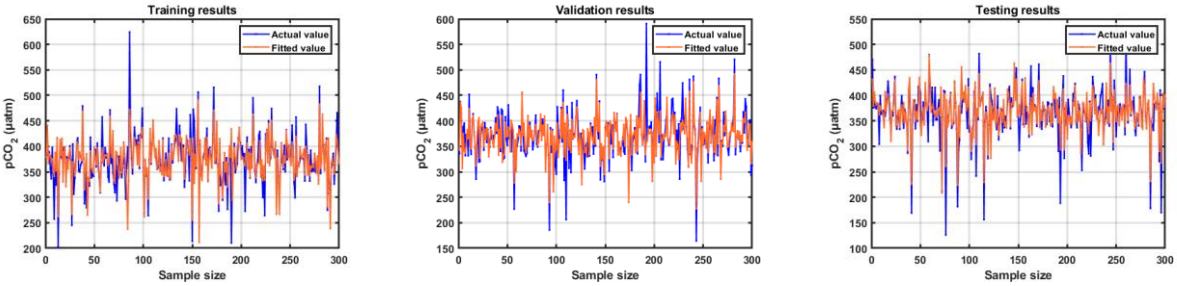
（a）MLR



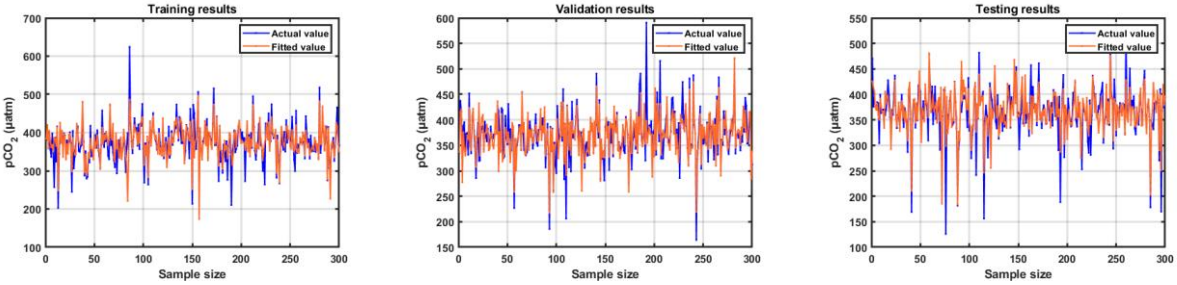
（b）CNN



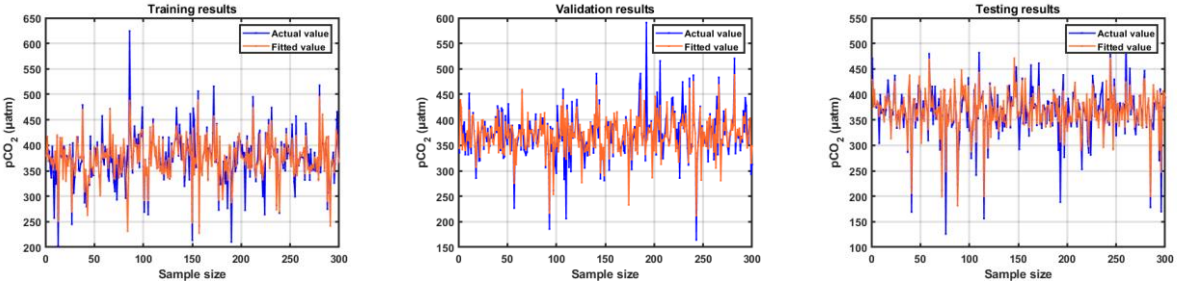
（c）GRU



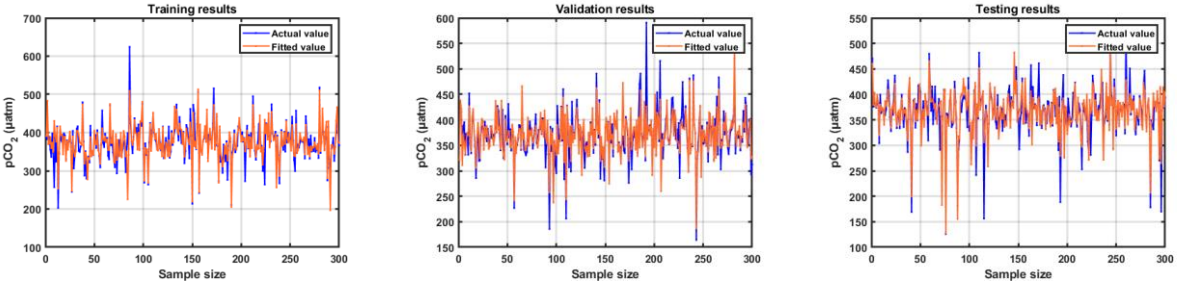
（d）LSTM



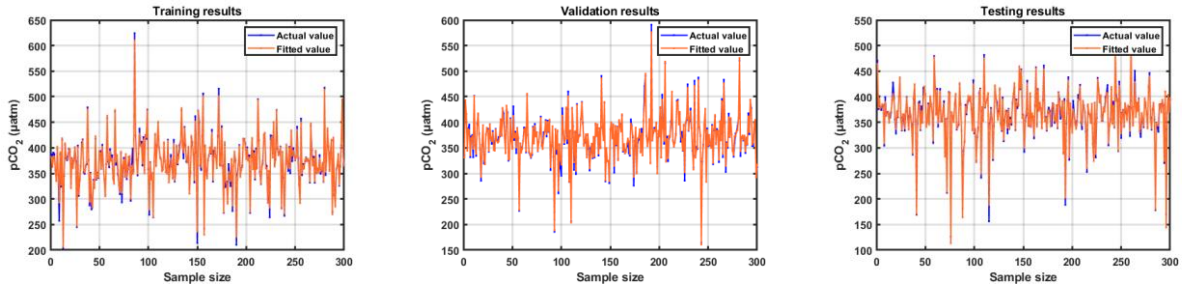
（e）GAM



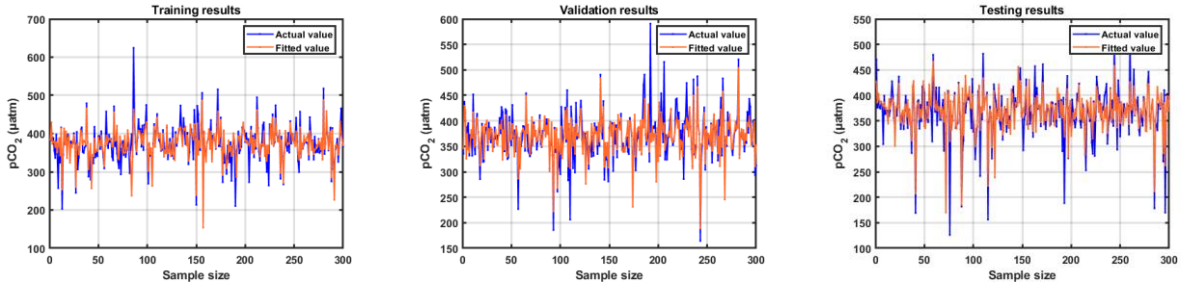
（f）XGBoost



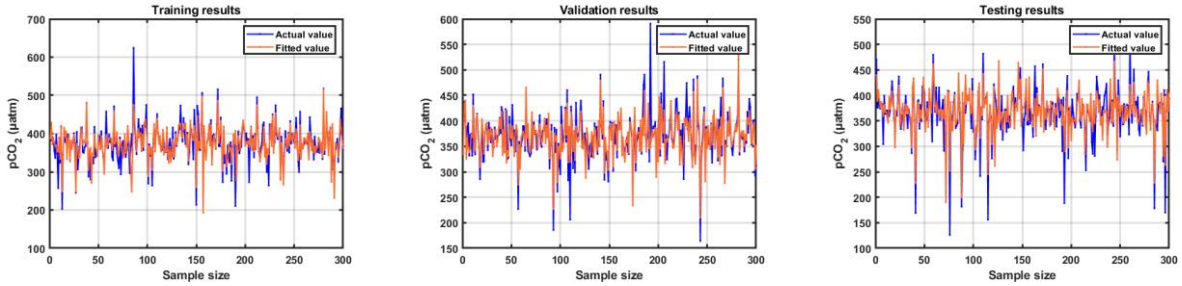
（g）LSBoost



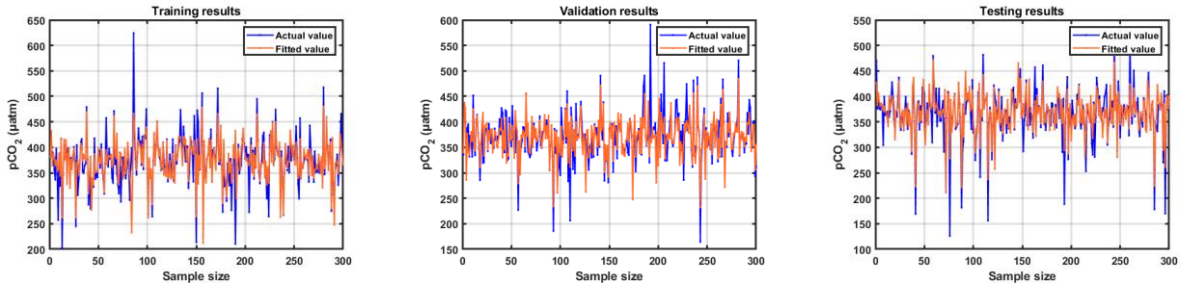
（h）RF



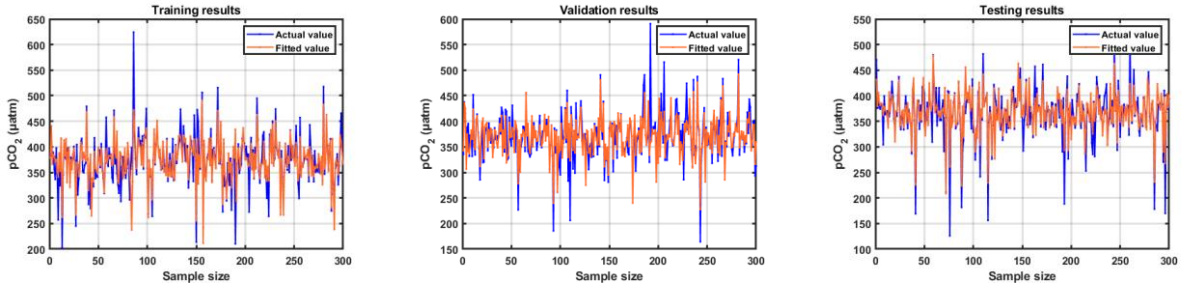
（a）MLR



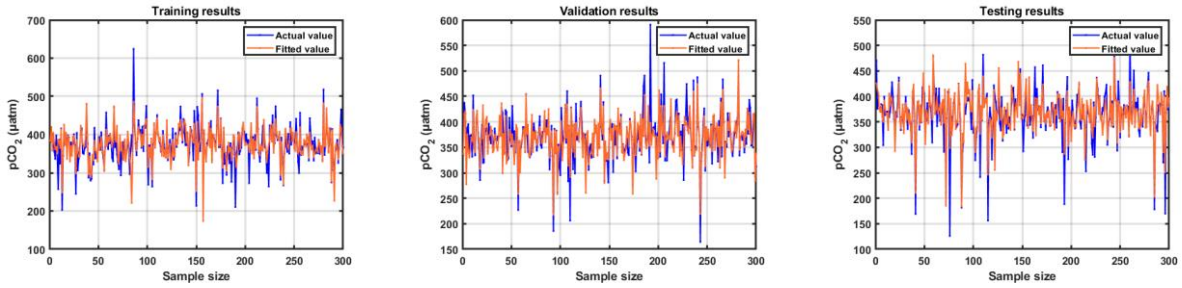
（b）CNN



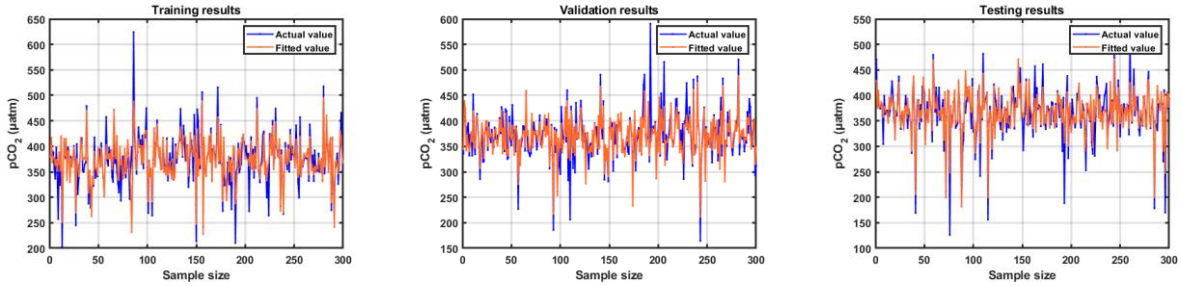
（c）GRU



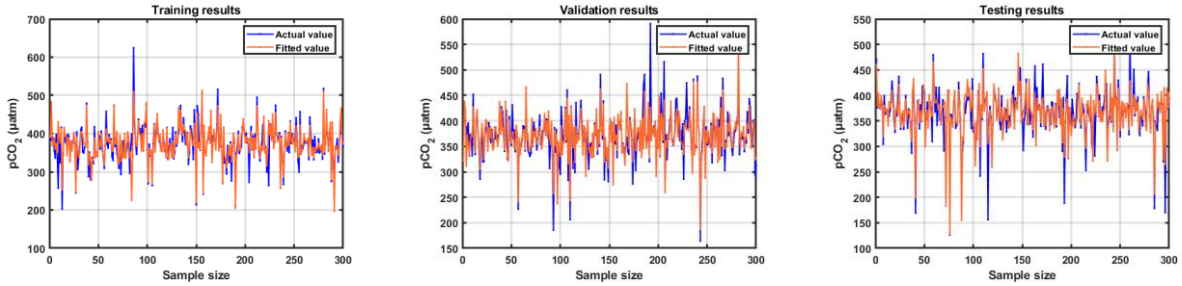
（d）LSTM



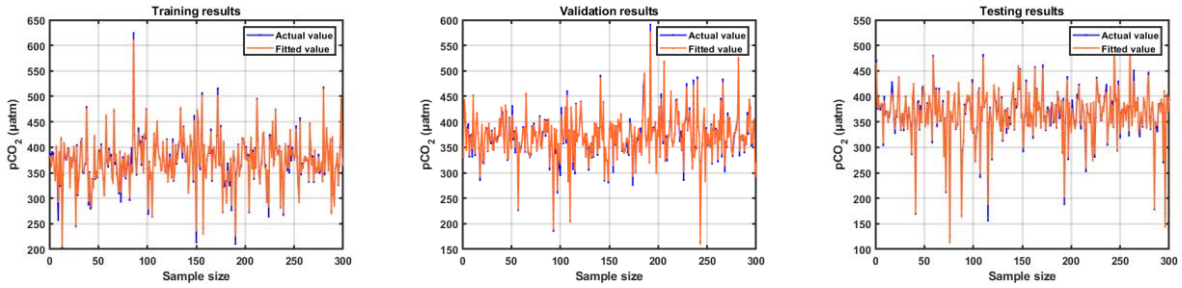
（e）GAM



（f）XGBoost



（g）LSBoost



（h）RF

Figure 4. Model performance at the global ocean

3.2.2 Construction and evaluation of surface pCO_2 model in far sea areas

The far sea environment is relatively stable, and the model performance has been improved (Table 4). The bias of MLR, CNN, and GRU models has been reduced, with MAE ranging from 14 μatm to 15 μatm, RMSE above 26 μatm, and R^2 remaining around 0.6. The MAE of LSTM and GAM is around 14 μatm; RMSE is above 25 μatm, and R^2 is around 0.64. The performance of the two models has improved compared to extreme value ranges, thanks to the ability of LSTM to process time series data and capture the dynamic characteristics of surface pCO_2 over time, and GAM fitted the relationship between surface pCO_2 and influencing factors by constructing a nonlinear additive model. XGBoost and LSBoost perform even better in far sea areas, especially with high fitting accuracy in the range of 300 μatm~500 μatm, MAE around 11 μatm~13 μatm, RMSE reduced to below 23 μatm, and R^2 increased to around 0.8. The model performance of RF in far sea areas is also optimal, relying on strong

generalization ability and feature selection mechanisms to effectively address the variability factors in marine environments.

Table 4. Performance parameters of different models in the far sea areas

	Model	MAE/μatm	MAPE	MSE/μatm²	RMSE/μatm	R²
Training	RF	3.068	0.009	27.456	5.240	0.985
	LSBoost	11.509	0.033	337.852	18.381	0.813
	XGBoost	13.191	0.038	500.054	22.362	0.723
	GAM	14.066	0.040	623.501	24.970	0.654
	LSTM	14.160	0.041	647.853	25.453	0.641
	GRU	14.377	0.041	665.920	25.805	0.631
	CNN	14.882	0.043	681.120	26.098	0.623
	MLR	15.274	0.044	737.902	27.164	0.591
Validation	RF	3.061	0.009	27.110	5.207	0.985
	LSBoost	11.532	0.032	338.102	18.388	0.814
	XGBoost	13.243	0.038	511.318	22.612	0.719
	GAM	14.143	0.040	644.144	25.380	0.646
	LSTM	14.219	0.040	667.947	25.845	0.632
	GRU	14.441	0.041	686.351	26.198	0.622
	CNN	14.929	0.042	701.278	26.482	0.614
	MLR	15.336	0.043	758.818	27.547	0.582
Testing	RF	2.900	0.008	22.082	4.699	0.988
	LSBoost	11.521	0.032	339.772	18.433	0.813
	XGBoost	13.223	0.038	508.771	22.556	0.720
	GAM	14.104	0.040	638.362	25.266	0.649
	LSTM	14.201	0.040	663.510	25.759	0.635
	GRU	14.423	0.041	681.866	26.113	0.625
	CNN	14.914	0.042	696.718	26.395	0.617
	MLR	15.316	0.043	754.142	27.462	0.585

3.2.3 Construction and evaluation of surface pCO_2 model in near sea areas

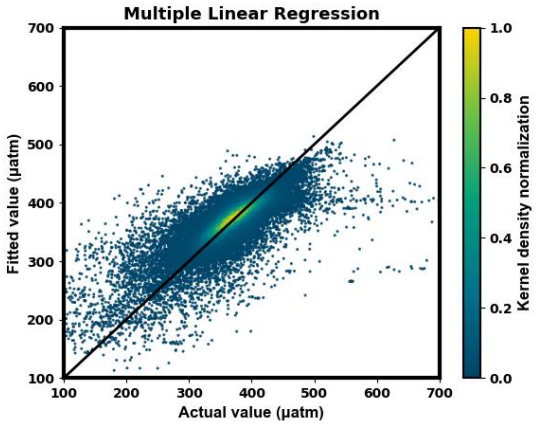
Due to various complex factors, the spatiotemporal distribution of surface pCO_2 in the near sea area exhibits high variability, resulting in a decrease in the performance of the constructed surface pCO_2 models. Table 5 results show that MLR, CNN, and GRU have limitations in handling complex nonlinear relationships. In the low and high value ranges, the MAE of the three models reaches over 34μatm, RMSE reaches over 62 μatm, and R^2 is below 0.5. LSTM constructs a nonlinear additive model through its gating mechanism and GAM, which improves the fitting ability to a certain extent. The MAE of the model is in the range of 33 μatm~34 μatm; the RMSE is in the range of 56 μatm~58 μatm, and the R^2 remains in the range of 0.55~0.60, but there is still deviation in the extreme numerical range. XGBoost and LSBoost improved the accuracy of fitting extreme values by constructing multiple weak learners to combine the fitting results. The MAE of both models decreased to around 23 μatm~27 μatm, the RMSE remained around 35 μatm~42 μatm, and the R^2increased to the range of 0.75~0.85. RF constructed multiple decision trees and integrated the fitting results to adapt to the variability and variability of the near sea environment, demonstrating robust fitting performance. Its MAE was below 5 μatm; RMSE was about 8 μatm, and R^2 remained above 0.95, significantly outperforming other models.

Table 5. Performance parameters of different models in the near sea areas

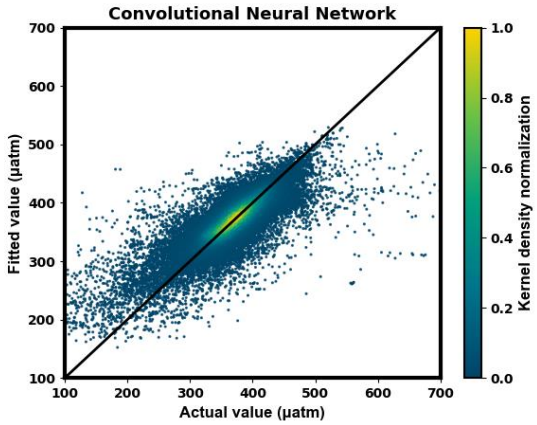
	Model	MAE/μatm	MAPE	MSE/μatm²	RMSE/μatm	R^2
Training	RF	5.396	0.016	98.332	9.916	0.977
	LSBoost	23.673	0.071	1267.869	35.607	0.833
	XGBoost	27.298	0.083	1783.422	42.231	0.765
	GAM	34.088	0.102	3058.776	55.306	0.597
	LSTM	32.738	0.100	3273.977	57.219	0.569
	GRU	34.022	0.103	3754.637	61.275	0.505
	CNN	36.309	0.110	3989.599	63.163	0.474
	MLR	36.264	0.109	4426.775	66.534	0.417
Validation	RF	5.346	0.016	93.028	9.645	0.978
	LSBoost	23.604	0.071	1263.495	35.546	0.832
	XGBoost	27.234	0.083	1766.706	42.032	0.765
	GAM	34.040	0.102	3033.228	55.075	0.596
	LSTM	32.686	0.100	3259.080	57.088	0.566
	GRU	33.987	0.103	3727.152	61.050	0.504
	CNN	36.239	0.110	3955.729	62.895	0.474
	MLR	36.188	0.109	4387.955	66.242	0.416
Testing	RF	4.756	0.014	64.708	8.044	0.972
	LSBoost	23.564	0.071	1244.921	35.283	0.839
	XGBoost	27.299	0.083	1788.363	42.289	0.769
	GAM	34.204	0.102	3134.086	55.983	0.595
	LSTM	32.911	0.100	3394.342	58.261	0.562
	GRU	34.236	0.103	3904.309	62.485	0.496
	CNN	36.465	0.110	4132.316	64.283	0.466
	MLR	36.405	0.109	4594.537	67.783	0.406

3.3 Independent validation of the model

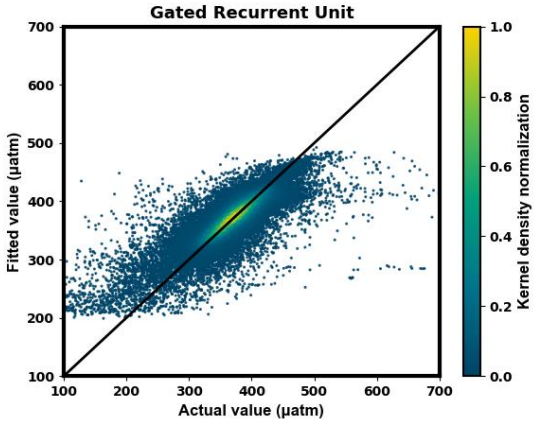
The surface pCO_2 models were independently validated at different regional scales, inputting data independent of the model construction, comparing the accuracy of the fitted values with the true values, and evaluating the applicability and accuracy of the model in complex marine environments. The scatter plot with true values as the x-axis and fitted values as the y-axis was drawn, with colors representing kernel density to reflect the distribution trend of points. At the global ocean scale (Figure 5), the scatter distribution of MLR, CNN, GRU, LSTM, and GAM shows a large elliptical shape, and the fitted values deviate significantly from the true values, especially around the extreme value of pCO_2 on the sea surface. The scatter distributions of XGBoost and LSBoost have shrunk. The RF model has the best fitting performance, with a clear convergence of the scatter distribution, concentrated on Y=X line, and can effectively avoid errors in the extreme value region, indicating that its fitted value is consistent with the true value and has good stability.


Multiple Linear Regression
Fitted value (μatm)
Actual value (μatm)
Kernel density normalization
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0

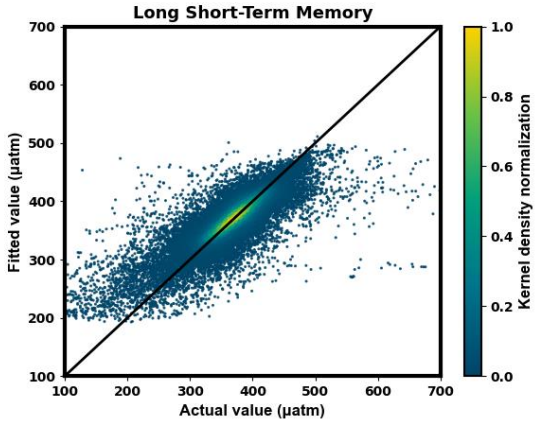
（a）


Convolutional Neural Network
Fitted value (μatm)
Actual value (μatm)
Kernel density normalization
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0

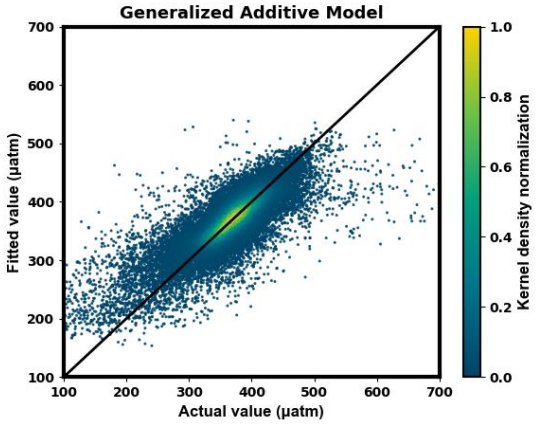
（b）


Gated Recurrent Unit
Fitted value (μatm)
Actual value (μatm)
Kernel density normalization
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0

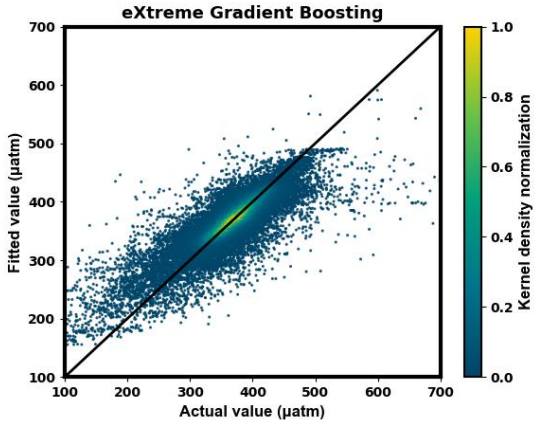
（c）


Long Short-Term Memory
Fitted value (μatm)
Actual value (μatm)
Kernel density normalization
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0

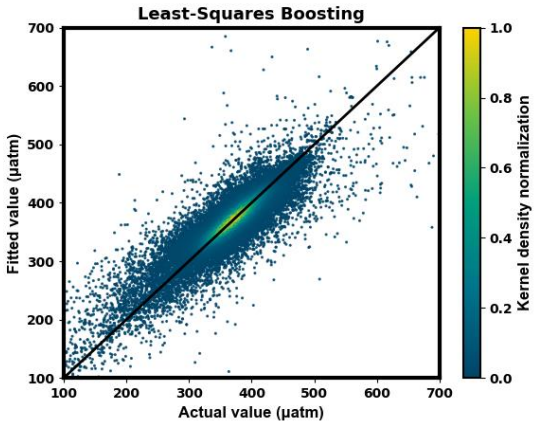
（d）


Generalized Additive Model
Fitted value (μatm)
Actual value (μatm)
Kernel density normalization
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0

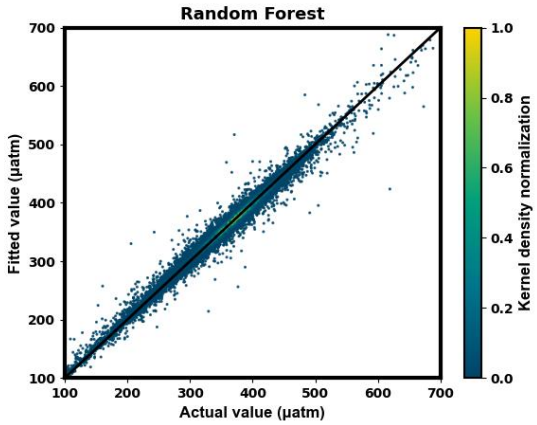
（e）


eXtreme Gradient Boosting
Fitted value (μatm)
Actual value (μatm)
Kernel density normalization
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0

（f）



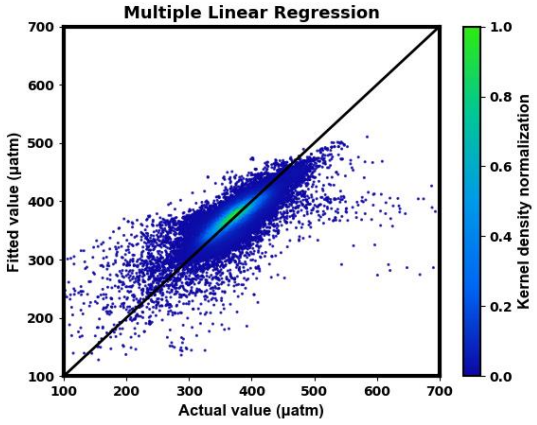
（g）



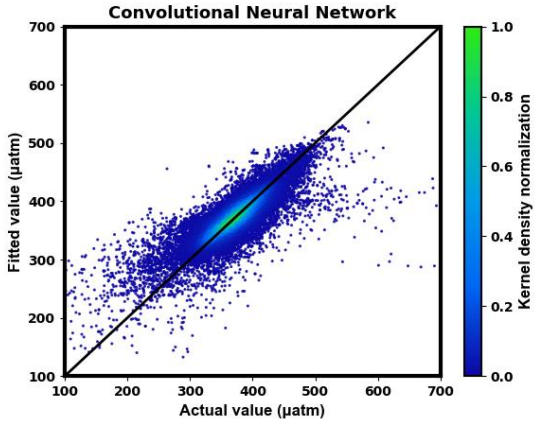
（h）

Figure 5. Independent verification performance of the models in the global ocean, right axis: Normalized probability density of model residuals. ((a) MLR, (b) CNN, (c) GRU, (d) LSTM, (e) GAM, (f) XGBoost, (g) LSBoost, (h)RF)

In far sea areas (Figure 6), the scatter points of MLR, CNN, GRU, LSTM, GAM, and XGBoost models exhibit elliptical distribution and diverge at both ends, indicating their limitations in dealing with extreme fluctuations of surface pCO_2. The scatter distribution ellipse of the LSBoost model significantly shrinks, and the divergence situation converges at extreme values, improving the fitting accuracy. The scatter distribution of the RF model is a flat ellipse, with the minimum difference between the fitted value and the true value, effectively reducing extreme errors.



（a）



（b）

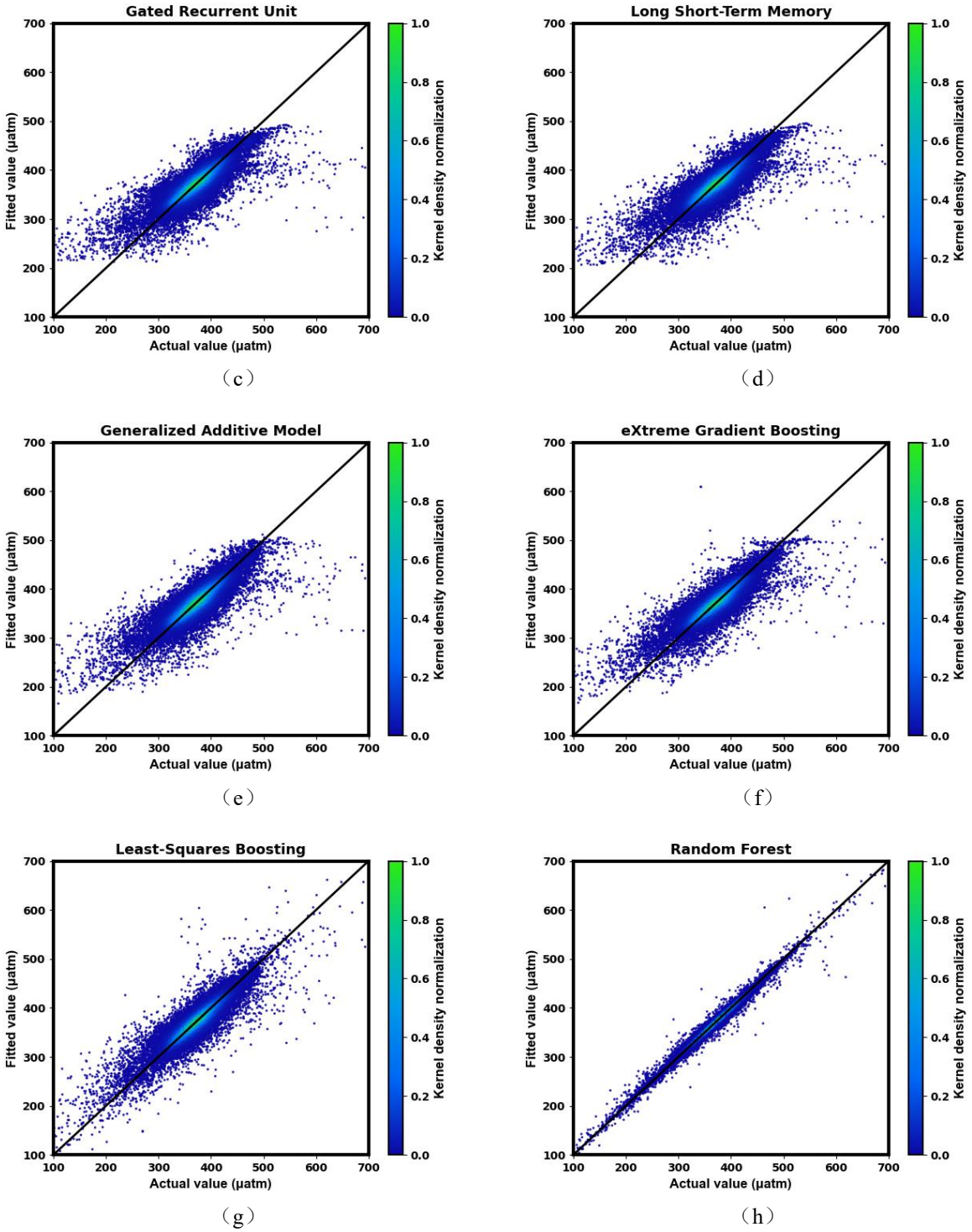
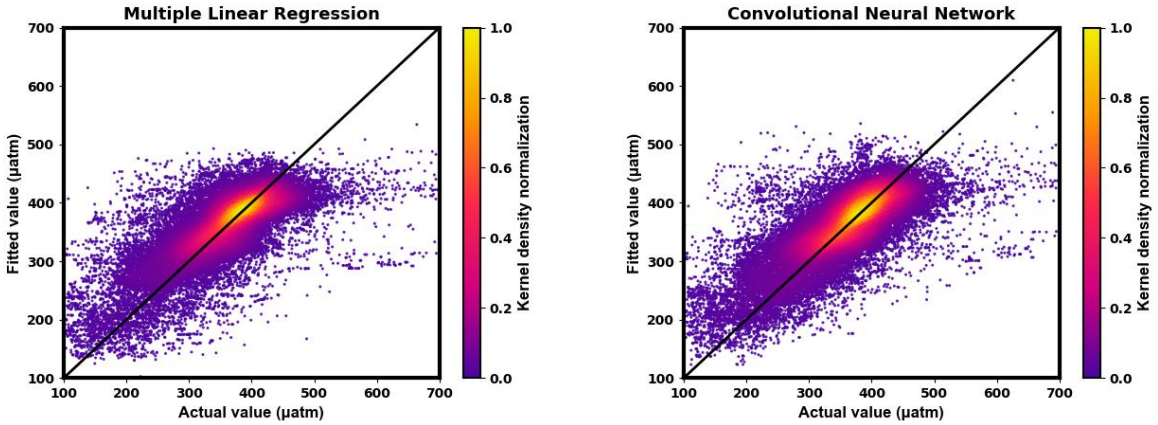


Figure 6. Independent verification performance of the models in the far sea areas, right axis: Normalized probability density of model residuals. ((a) MLR, (b) CNN, (c) GRU, (d) LSTM, (e) GAM, (f) XGBoost，(g) LSBoost, (h)RF)

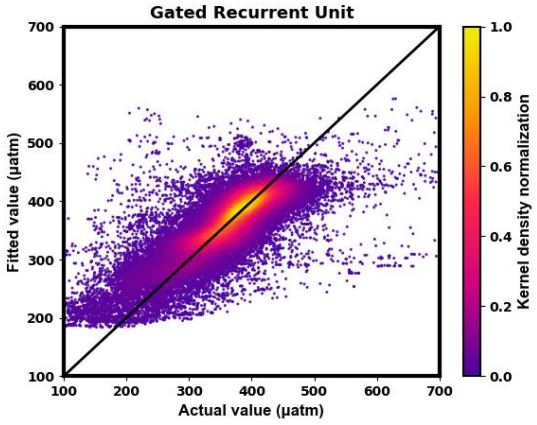
In the independent validation of models in near sea areas, each model showed different performances (Figure 7). The scatter of MLR, CNN, GRU, and LSTM shows an irregular distribution, with significant differences between the fitted values and the true values, and severe divergence in high-value areas. This is due to the high variability in near sea areas, which makes it difficult for the model to cope with. The scatter distribution of GAM and XGBoost has begun to show an elliptical shape, which has certain adaptability to complex environments. The scatter

distribution of LSBoost shows a clear elliptical shape, which improves the fitting stability. The RF model shows significant improvement in performance, with overall convergence of scatter distribution and no significant divergence in both low and high value oceans. It can effectively reduce extreme errors and reconstruct surface pCO_2 with high accuracy in complex near sea environments.

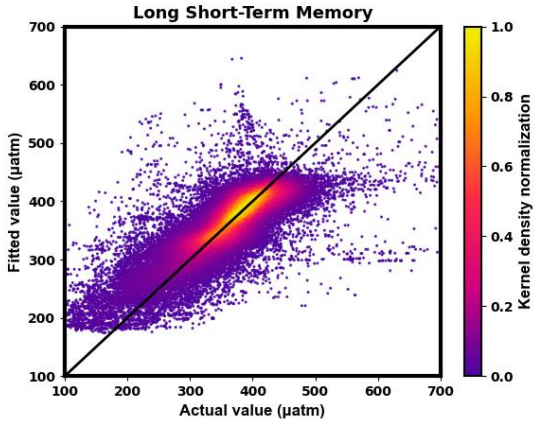


（a）

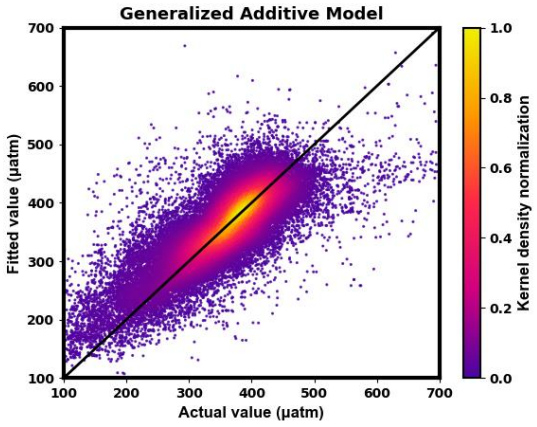
（b）



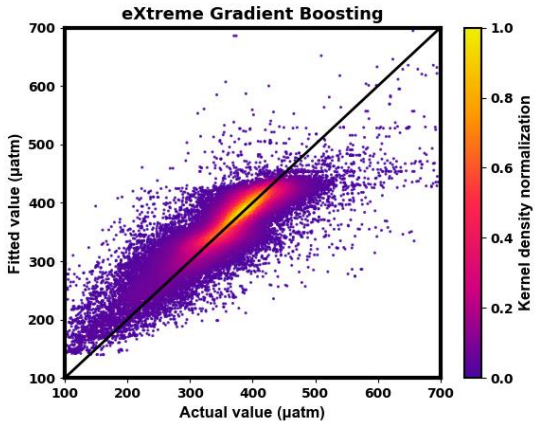
（c）



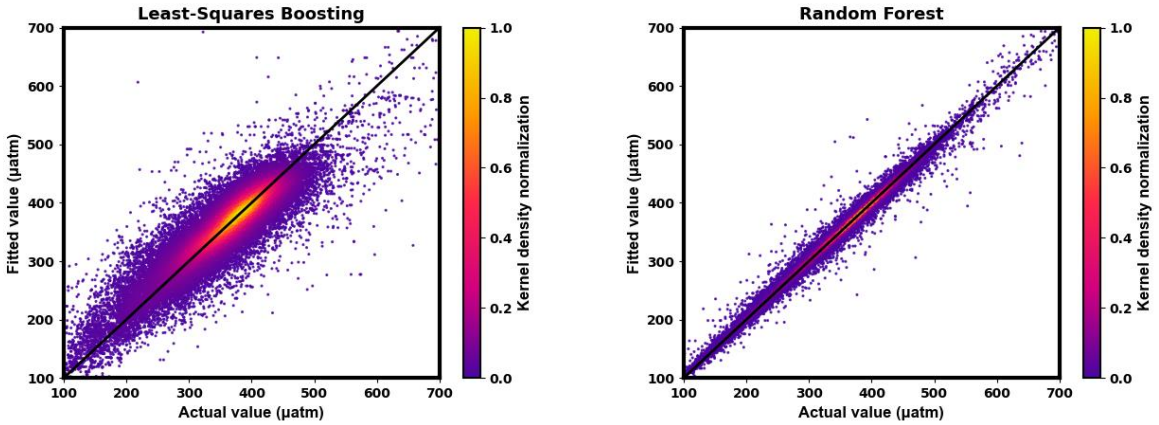
（d）



（e）



（f）



（g） （h）

Figure 7. Independent verification performance of the models in the near sea areas, right axis: Normalized probability density of model residuals. ((a) MLR, (b) CNN, (c) GRU, (d) LSTM, (e) GAM, (f) XGBoost，(g) LSBoost, (h)RF)

3.4 Reconstruction of surface pCO_2

The multi-source data was input into the constructed RF model at different area scales, with extracting the variable values of influencing factors from the multi-source data grid by grid to fit the surface pCO_2 values of the corresponding grid. If there are missing values in a certain grid in the multi-source data, the corresponding surface pCO_2 value at that location will be output as a blank value, ensuring that the reconstructed results are completely based on the original data. The blank values are mainly due to the systematic exclusion of land pixels and the limitations of data acquisition in high latitude sea areas: the former is excluded because it does not participate in ocean processes, while the latter is due to the lack of satellite data for key parameters caused by sea ice coverage or insufficient light, resulting in the inability to reconstruct the values in the region. The final generation of the surface pCO_2 distribution map for the year 2000~2019 at 0.25°× 0.25°is based on the original data.

The reconstruction results of surface pCO_2 at the global ocean scale are consistent with the distribution characteristics of LDEO actual observation data, confirming that the RF model can effectively capture the spatial distribution pattern of global ocean surface pCO_2. Through the reconstruction results (Figure 8), it was found that the spatial distribution of surface pCO_2 exhibits a clear latitude dependence, with a distribution pattern of "high at the equator and low at the poles". The independent observation data based on the route was compared with the reconstruction results obtained at the closest collection time. The global ocean surface pCO_2 reconstruction result showed MAE of 11.067 μatm, MAPE of 0.037, MSE of 396.060 μatm², RMSE of 19.901 μatm, and R^2 of 0.816. This indicates that the deviation between the reconstructed results and the actual observed data is small, and can accurately reflect the average distribution characteristics of surface pCO_2.

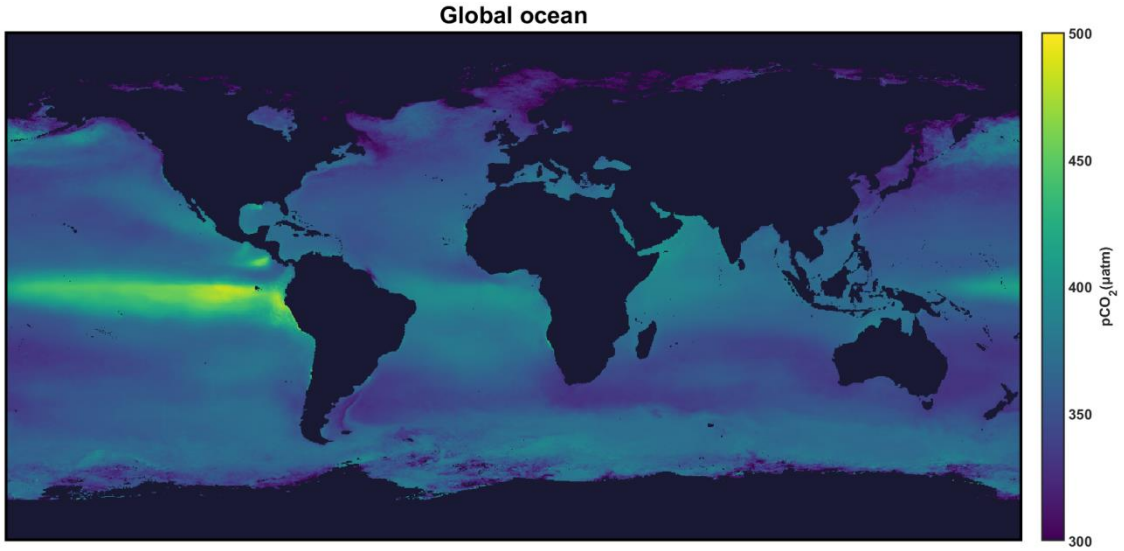
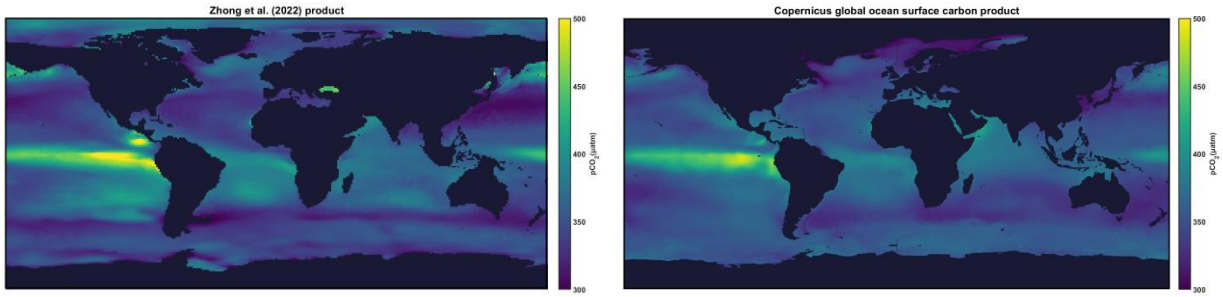


Figure 8. Surface ocean pCO_2 products in the global ocean



（a）Zhong et al.(2022) product

（b）Copernicus global ocean surface carbon product

Figure 9. Comparison of surface ocean pCO_2 products from different studies

Compared with other existing studies on the reconstruction of surface pCO_2 (Figure 9), these methods are highly consistent with our results in the reconstructed spatial model pattern（Zhong et al., 2022; Chau et al., 2021; Chau et al., 2022）. Although different studies have used different data sources, models, or methods, similar conclusions can be drawn when describing the overall distribution characteristics of pCO_2 on the global ocean surface, which to some extent verifies the reliability and accuracy of the reconstructed results. This study uses high-resolution data and RF models to make the reconstruction results more detailed, especially in the high latitude marginal sea areas of the North and South Poles.

The reconstruction results of the far sea region showed that the surface pCO_2 in the equatorial low latitude region was higher, while the surface pCO_2 in the polar high latitude region was lower (Figure 10). We evaluated the difference in fitting accuracy between the far sea regional model and the global ocean model in the far sea areas, by comparing independent observation data based on flight routes with the reconstructed results of the two models. The results showed that the MAE of the far-sea model was 9.060 μatm, the MAPE was 0.027, the MSE was 269.511 μatm^2, the RMSE was 16.417μatm, and R^2 was 0.826; the MAE of the global model was 9.125 μatm, the MAPE was 0.027, the MSE was 275.582 μatm^2, the RMSE was 16.601 μatm, and R^2 was 0.822.The reconstruction accuracy of the far sea area model has slightly improved compared to the global ocean model in the far sea area (Figure 11), indicating that the optimization of the far sea area model in local areas has improved the

reconstruction accuracy. However, the global ocean model can still provide accurate surface pCO_2 fitting in the far sea area by adapting to the overall ocean environment.

To verify the accuracy of the time series reconstruction of the model, a comparative analysis was conducted on the temporal changes between the observation data of the Hawaii Ocean Time series (HOT) and the reconstruction results of the global ocean and far sea areas (Figure 12). The results showed that the temporal trends of both scales were consistent with the actual measurement data of the Hawaii observation station. Research has shown that the model performs well in fitting the dynamic changes of time series and can accurately reflect the temporal evolution of surface pCO_2.

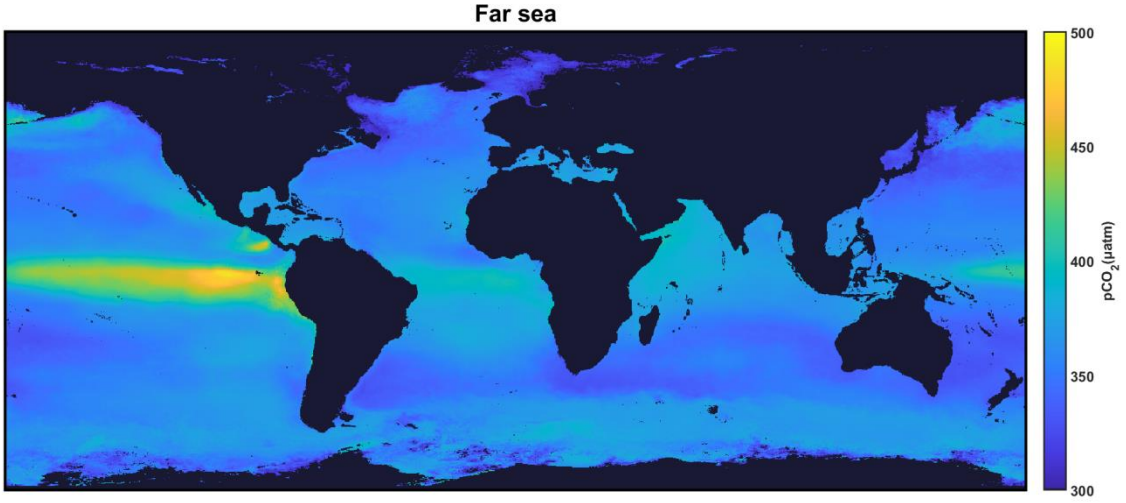


Figure 10. Surface ocean pCO_2 products in the far sea areas

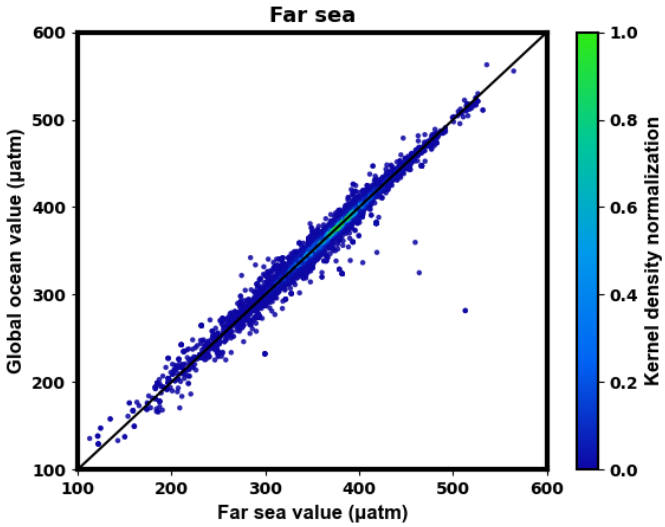


Figure 11. Comparison of reconstruction accuracy in the far sea areas using different scale models, right axis: Normalized probability density of model residuals.

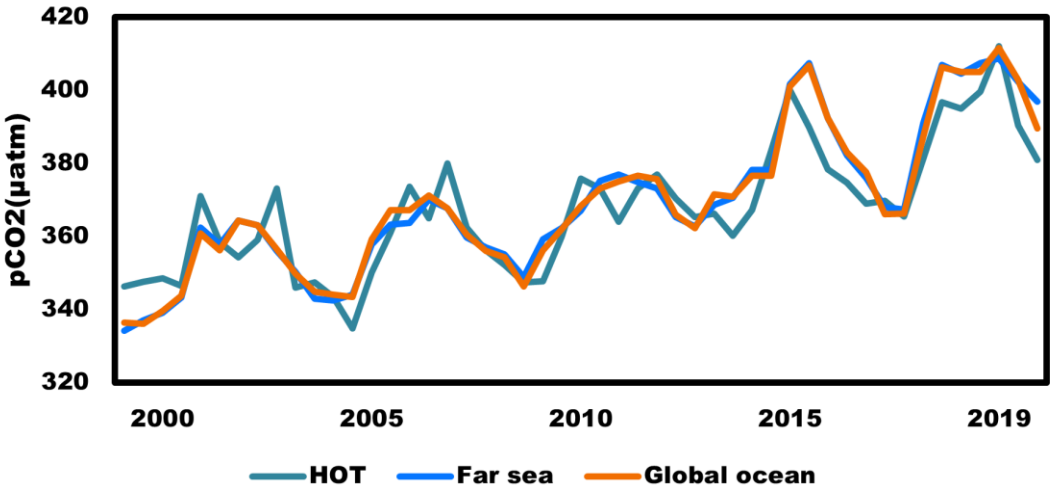


Figure 12. Independent verification based on time-series observation stations

The reconstruction results of surface pCO_2 in the near sea area showed (Figure 13) that the surface pCO_2 values in the low latitude near sea areas on both sides of the equator were higher, which was closely related to factors such as high seawater temperature and vigorous evaporation. The seawater temperature in high latitude oceans is lower, causing changes in ocean circulation and mixing processes, and the overall trend of surface pCO_2 is decreasing. A comparison was made between the fitting accuracy of the near sea area model and the global ocean model in the near sea region. The results showed that the MAE of the near-shore model was 20.145 μatm, the MAPE was 0.065, the MSE was 983.726 μatm², the RMSE was 31.364 μatm, and R^2 was 0.797; the MAE of the global model was 20.324 μatm, the MAPE was 0.065, the MSE was 999.147 μatm², the RMSE was 31.609 μatm, and R^2 was 0.794. The reconstruction effect of the near sea area model has been improved compared to the reconstruction results of the global ocean model in the near sea area (Figure 14), indicating that the use of RF can model the complex marine environment in the near sea area and accurately reflect the distribution characteristics of surface pCO_2 in the region.

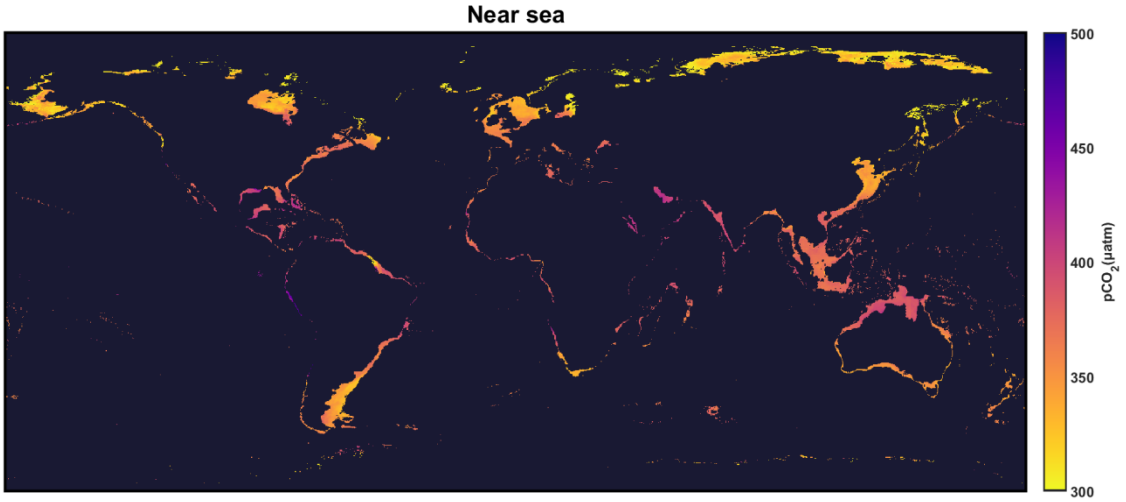


Figure 13. Surface ocean pCO_2 products in the near sea areas

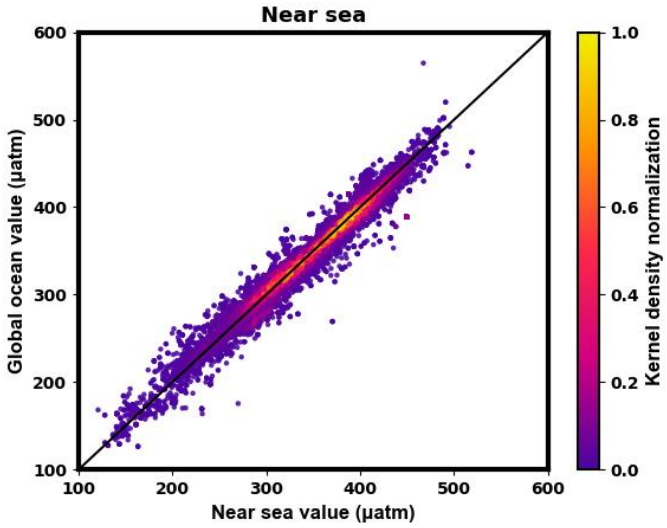
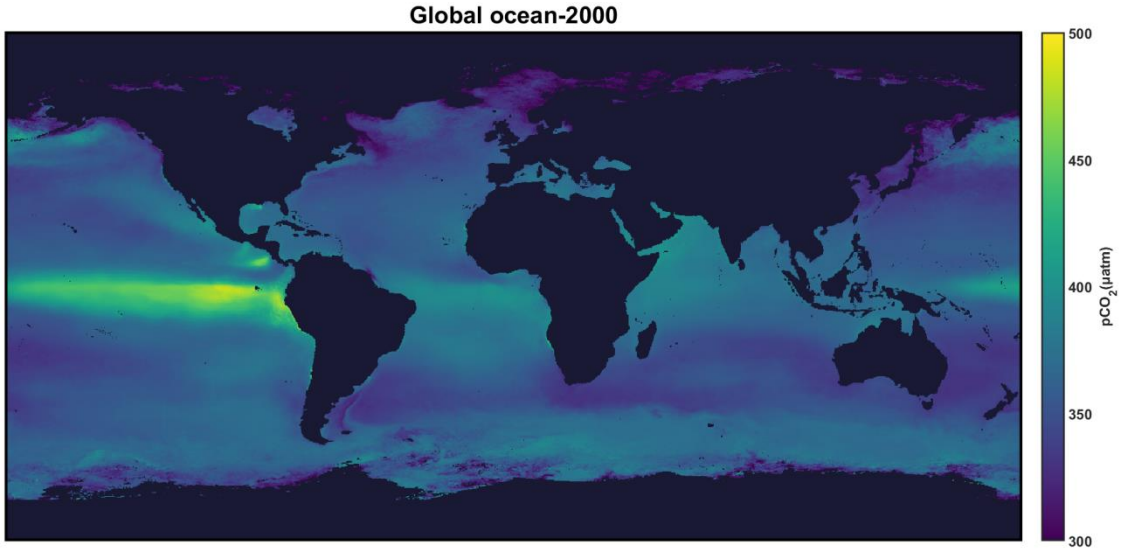
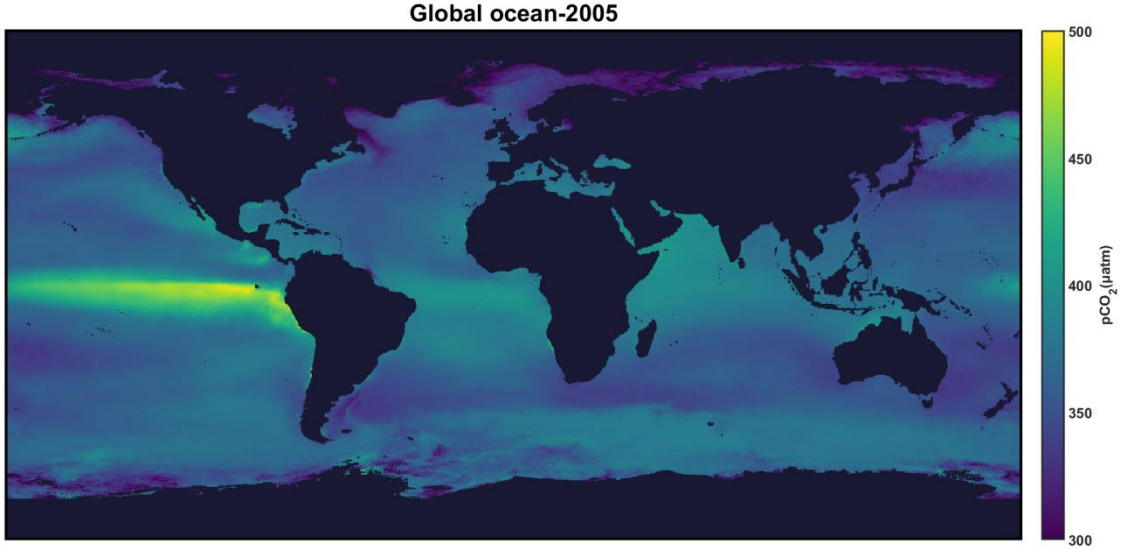


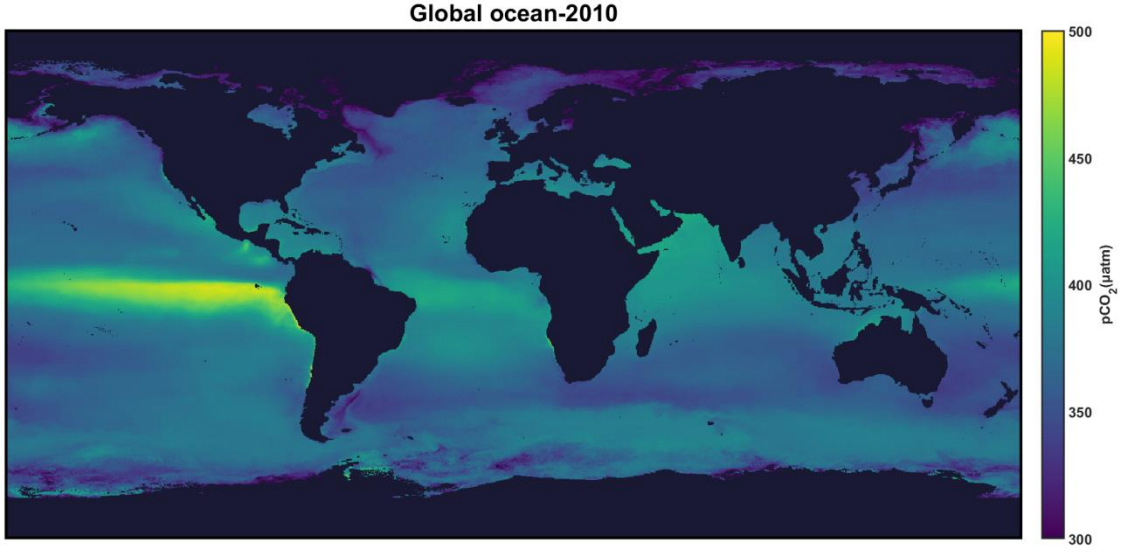
Figure 14. Comparison of reconstruction accuracy in the near sea areas using different scale models, right axis: Normalized probability density of model residuals.

3.5 Spatiotemporal analysis of surface pCO_2

At the global oceanic scale (Figure 15), the equatorial region experiences strong solar radiation and high temperatures, resulting in relatively low solubility of CO_2. Additionally, the presence of upwelling brings deep seawater rich in CO_2 to the surface, leading to an increase in surface pCO_2 concentration. Due to the low temperature environment in polar oceans, the solubility of CO_2 in seawater significantly increases. The sea ice coverage and strong wind fields in polar waters promote gas exchange between the atmosphere and the ocean, resulting in relatively low concentrations of pCO_2 on the sea surface. The surface pCO_2 in the Antarctic region is generally higher than that in the Arctic region, because the circulation system transports a large amount of seawater with high surface pCO_2 from low latitudes to high latitudes. At the same time, the melting and formation of sea ice also have an important impact on the distribution of surface pCO_2. Due to the wider coverage of sea ice, the Arctic region is less affected by the North Atlantic warm current, and its surface pCO_2 concentration is lower compared to the Antarctic region. In terms of time, the global ocean surface pCO_2 shows a trend of increasing year by year, which is related to global warming. The rising sea temperature in mid latitude waters leads to a decrease in CO_2 solubility and promotes an increase in surface pCO_2 concentration.


Global ocean-2000
500
450
400
350
300
pCO2 (µatm)


Global ocean-2005
500
450
400
350
300
pCO2 (µatm)


Global ocean-2010
500
450
400
350
300
pCO2 (µatm)

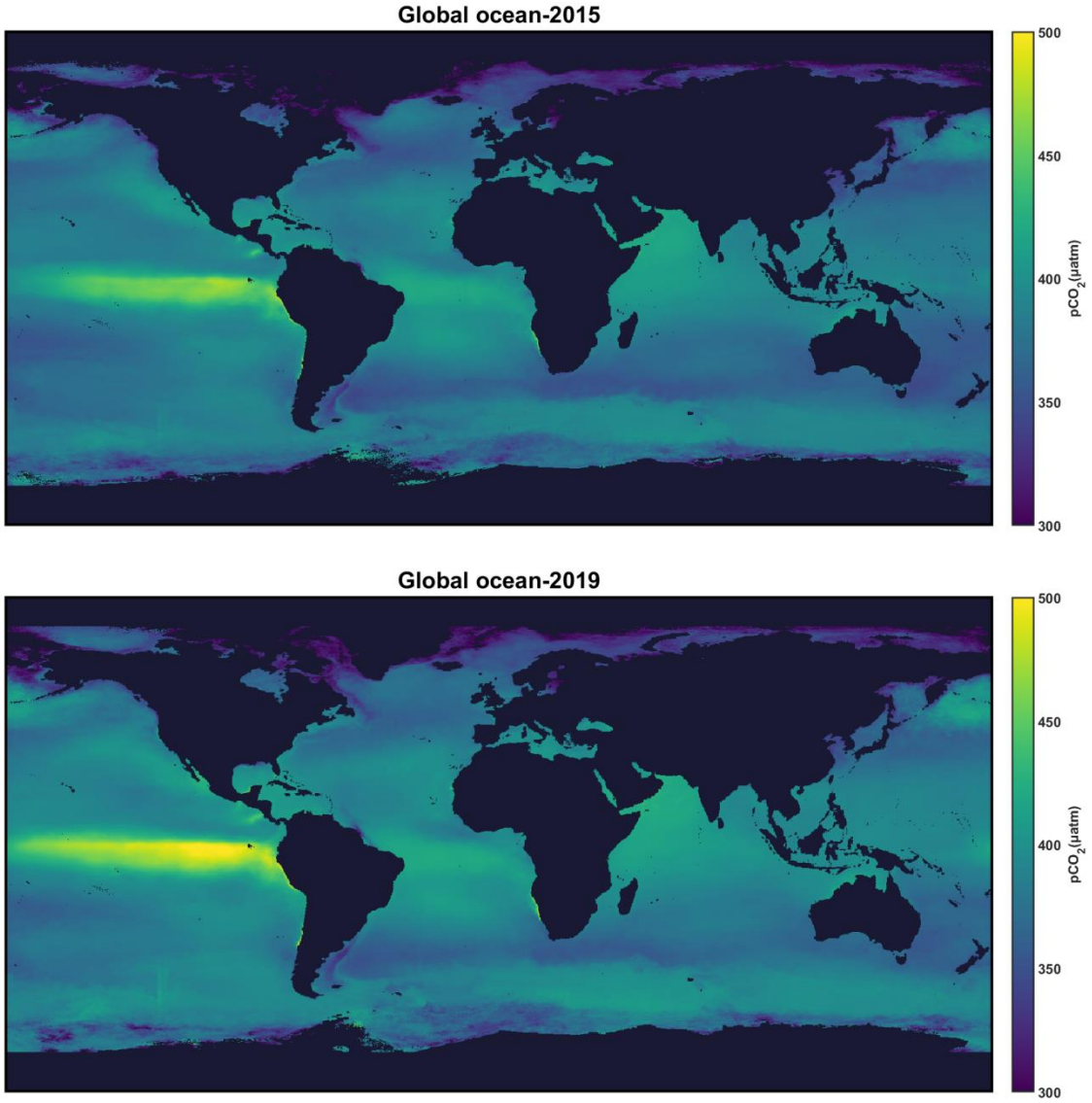
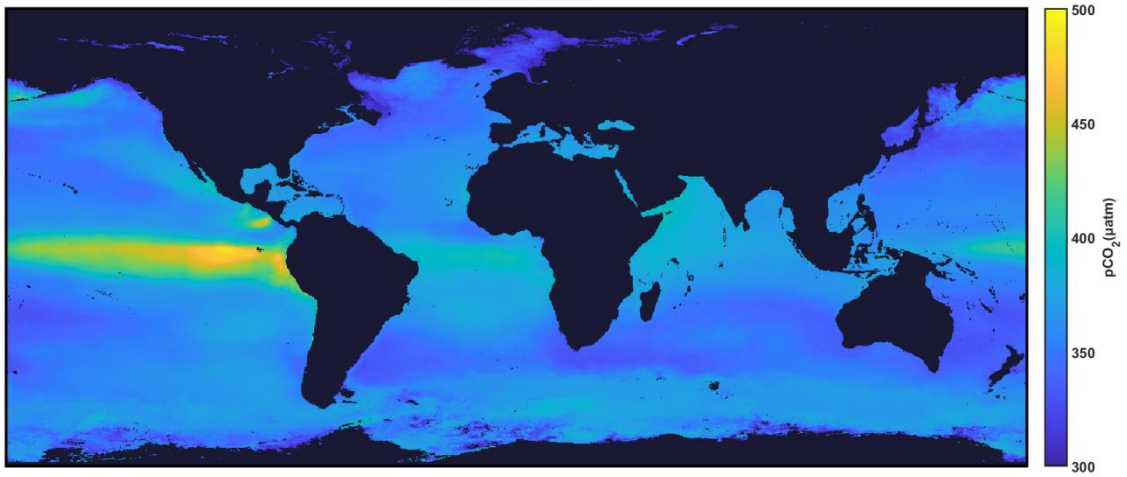


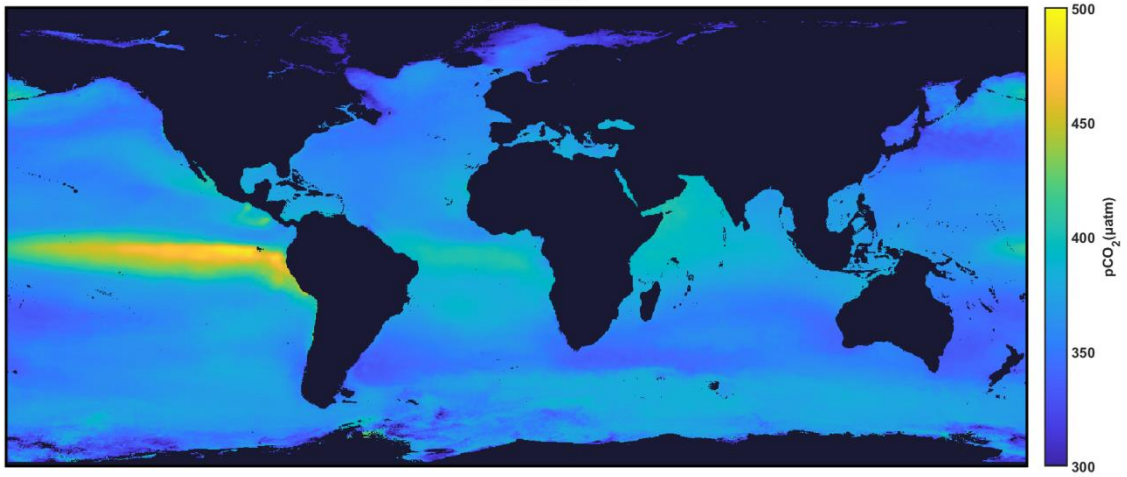
Figure 15. Annual spatiotemporal variations of surface ocean pCO_2 in the globle ocean

In the far sea areas (Figure 16), the surface pCO_2 is higher in the low latitude areas near the equator, particularly in the eastern equatorial Pacific. Mainly due to the upwelling of seawater in the region, which brings cold water rich in CO_2 from deep layers to the surface of the ocean, resulting in an increase in pCO_2 concentration on the sea surface. In the mid to high latitudes of the far sea region, the surface pCO_2 shows a low characteristic, which is due to the ocean circulation pattern promoting the mixing of surface seawater and deep seawater, resulting in relatively low surface pCO_2 concentration. The low temperature and strong biological pumping effect enhance the absorption of atmospheric CO_2 by the ocean, leading to a low surface pCO_2 concentration. In terms of time, the surface pCO_2 shows a trend of increasing year by year, especially after 2015. This is closely related to global climate change, changes in ocean circulation patterns, and the impact of human activities.

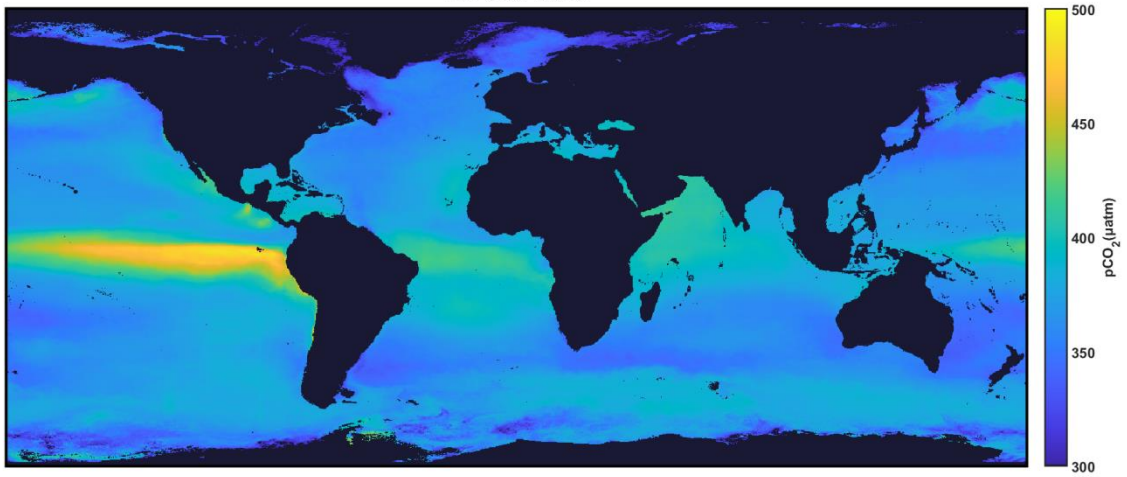
Far sea-2000



Far sea-2005



Far sea-2010



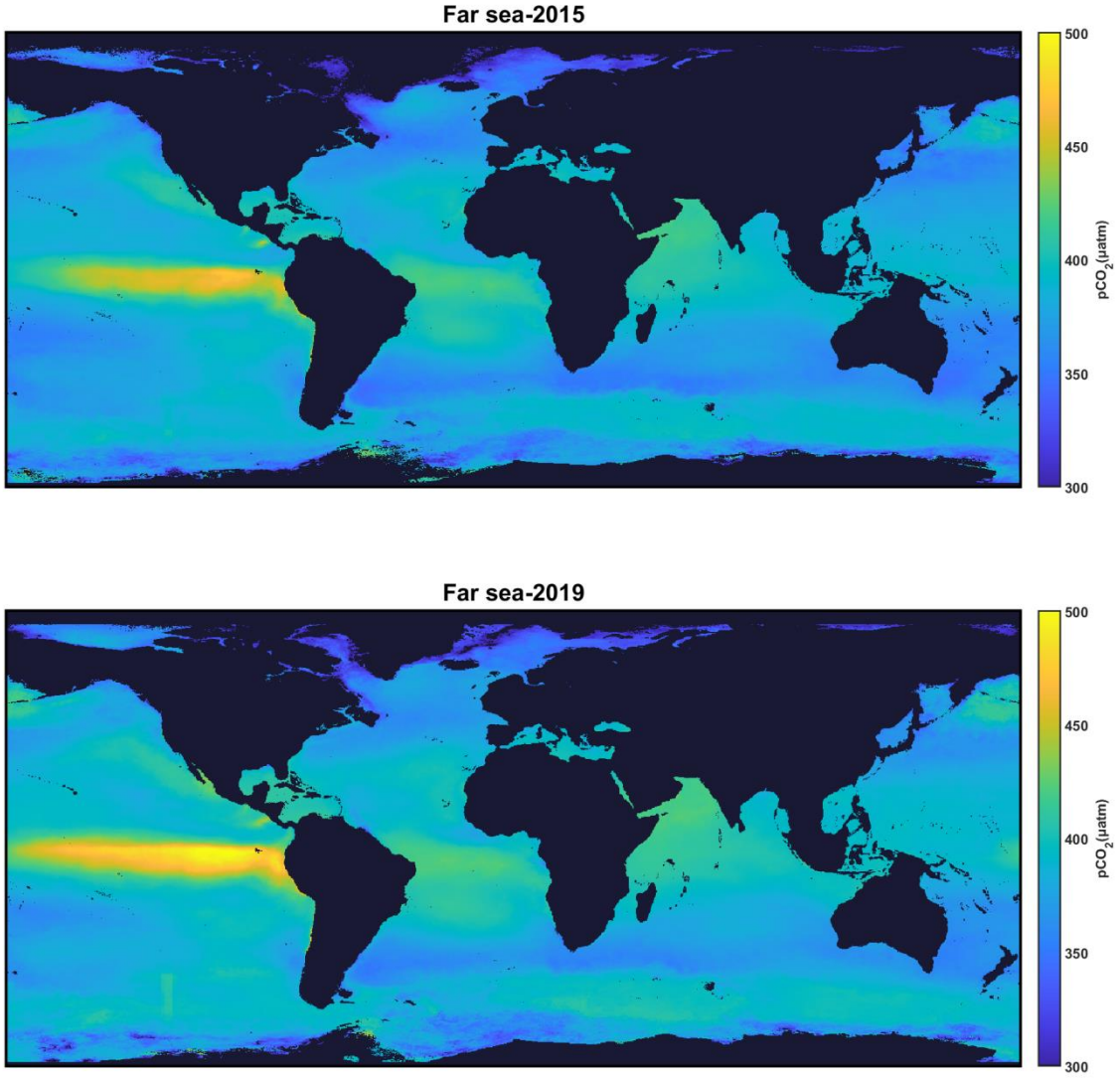
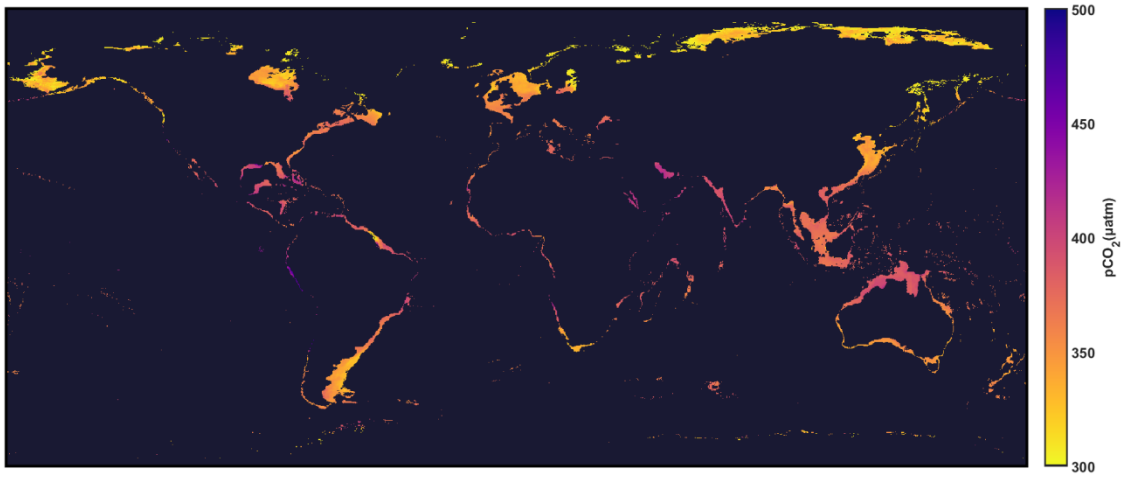


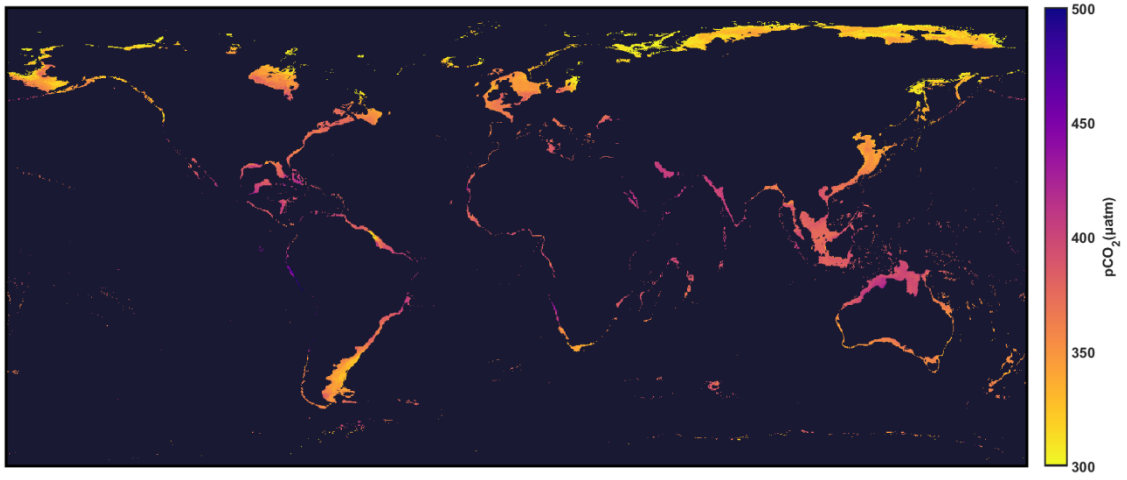
Figure 16. Annual spatiotemporal variations of surface ocean pCO_2 in the far sea areas

The exchange of CO_2 between seawater and atmosphere is frequent, and the surface pCO_2 value is relatively high. In mid to high latitude oceans, low-temperature seawater, polar cold water sinking, and deep seawater upwelling result in relatively low concentrations of pCO_2. The reconstruction results of surface pCO_2 in the near sea area (Figure 17) show that the equatorial region has strong solar radiation, high temperature seawater, and the influence of tropical cyclones and trade winds.The distribution characteristics of surface pCO_2 are significant along the eastern coast of Asia in the mid latitude region of the Northern Hemisphere. The surface pCO_2 in the Yellow Sea and Bohai Sea oceans is significantly lower than that in the coastal areas of eastern North America, which is related to the East Asian monsoon circulation and complex marine ecosystems. The surface pCO_2 in the border waters between Southeast Asia, the Indian Peninsula, North America, and South America is relatively high. Due to the influence of monsoon climate and tropical cyclones, high sea temperatures, as well as marine pollution caused by human activities, have collectively led to an increase in surface pCO_2. Temporally, the surface pCO_2 in near sea areas has been increasing year by year. Due to the increase in temperature in low latitude sea areas, the solubility of CO_2 in seawater decreases, and the upward trend of surface pCO_2 is more pronounced.

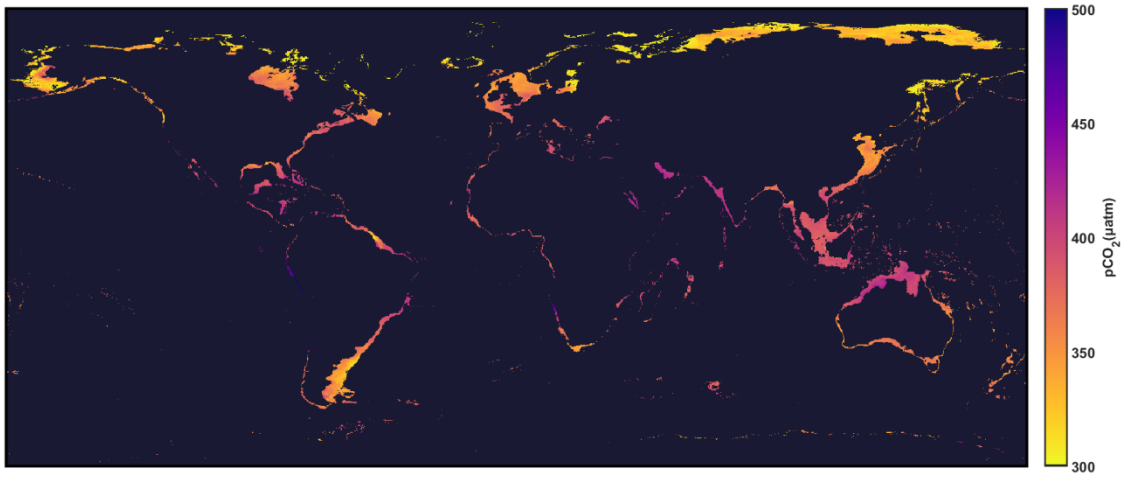
Near sea-2000



Near sea-2005



Near sea-2010



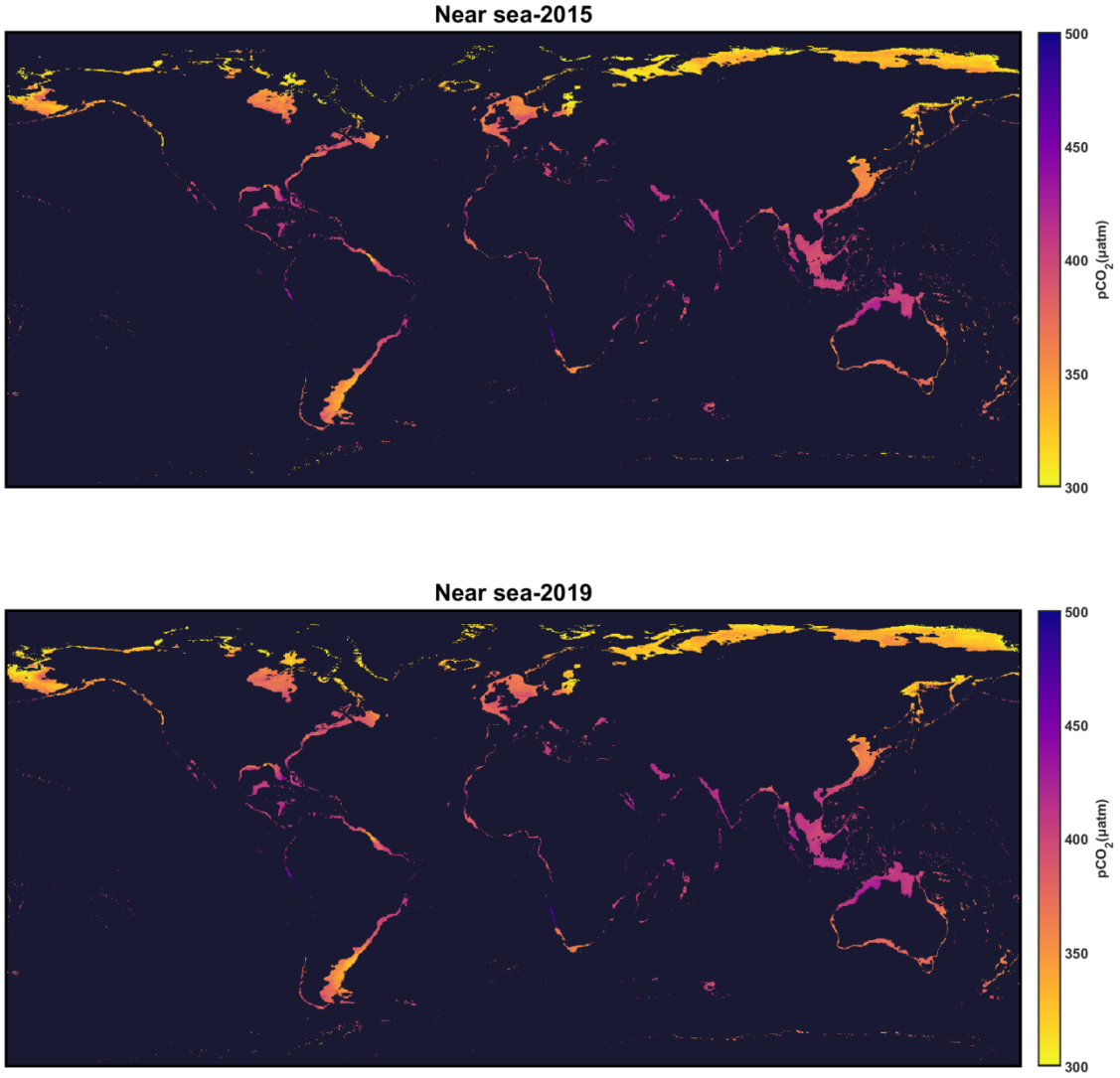


Figure 17. Annual spatiotemporal variations of surface ocean pCO_2 in the near sea areas

4. Conclusion

This study is based on a multi-scale analysis framework of the global ocean, far sea areas, and near sea areas. Using LDEO measured data combined with multi-source data, multiple machine learning models were used to construct and reconstruct the annual surface pCO_2 distribution of 0.25°× 0.25°from 2000 to 2019, revealing its spatiotemporal variation patterns and driving mechanisms. The research results indicate that the Random Forest (RF) model exhibits optimal performance at different scales and can effectively capture the spatiotemporal distribution characteristics of surface pCO_2. The distribution pattern of surface pCO_2 shows a pattern of "high at the equator and low at the poles" in space, and an increasing trend year by year in time. Different oceans exhibit different characteristics of changes due to the combined effects of natural factors and human activities. The acidity and alkalinity of seawater are the main driving factors for changes in surface pCO_2, and the contributions of other influencing factors vary at different scales.

Although this study has achieved certain results, the complexity of ocean carbon sinks still needs further exploration. Future research can focus on optimizing models, developing hybrid models, and combining advanced algorithms with ocean mechanism models; At the same time, we will strengthen interdisciplinary studies such as

oceanography, ecology, and climatology to comprehensively reveal the process of ocean carbon cycling and provide scientific basis for addressing climate change.

Author contribution

Conceptualization, **[H.W.]**; methodology, **[H.W.]** and **[Y.J.]**; software, **[X.L.]** and **[Y.J.]**; validation, **[W .Z.]**, **[L.C.]**, and **[L.W.]**; formal analysis, **[Y.J.]**; investigation, **[W .Z.]**, **[L.W.]** and **[L.C.]**; resources, **[X.L.]** and **[Y.J.]**; data curation, **[X.L.]** and **[Y.J.]**; writing—original draft preparation, **[Y.J.]**, **[Y.W.]** and **[M.L.]**; writing—review and editing, **[H.W.]** and **[Z.L.]**; visualization, **[X.L.]** and **[L.C.]**; supervision **[H.W.]**; project administration, **[H.W.]**; funding acquisition, **[H.W.]** All authors have read and agreed to the published version of the manuscript.

Competing interests

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

Financial support

This research was funded by Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People' s Republic of China, grant numbers G202211, and the Ministry of Education Industry-University Collaborative Education Project, grant numbers 220504039151258, and the Fundamental Research Funds for the Central Universities, grant numbers 18CX02064A.

Data availability

Data will be made available on request.

Reference

Falkowski, P., Scholes, R. J., Boyle, E., Canadell, J., Canfield, D., Elser, J., . . . Linder, S. J. s. (2000). The global carbon cycle: a test of our knowledge of earth as a system. *290*(5490), 291-296.

Ge, W., Patino, J., Todisco, M., & Evans, N. (2022). *Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations.* Paper presented at the ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP).

Jain, A. K. (2022). Global carbon budget 2022. *Earth System Science Data, 14*(11), 4811-4900.

Telszewski, M., Chazottes, A., Schuster, U., Watson, A., Moulin, C., Bakker, D., . . . Lüger, H. (2009). Estimating the monthly pCO 2 distribution in the North Atlantic using a self-organizing neural network. *Biogeosciences, 6*(8), 1405-1421.

Chierici M, Signorini S R, Mattsdotter-Björk M, et al. Surface water fCO2 algorithms for the high-latitude

Pacific sector of the Southern Ocean[J]. Remote Sensing of Environment, 2012, 119: 184-196.DOI: 10.1016/j.rse.2011.12.020

Nakaoka S, Telszewski M, Nojiri Y, et al. Estimating temporal and spatial variation of ocean surface pCO 2 in the North Pacific using a self-organizing map neural network technique[J]. Biogeosciences, 2013, 10(9): 6093-6106.DOI: 10.5194/bg-10-6093-2013

Landschützer P, Gruber N, Bakker D C E, et al. A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink[J]. Biogeosciences, 2013, 10(11): 7793-7815.DOI: 10.5194/bg-10-7793-2013

Marrec P, Cariou T, Macé E, et al. Dynamics of air-sea CO 2 fluxes in the North-West European Shelf based on Voluntary Observing Ship (VOS) and satellite observations[J]. Biogeosciences Discussions, 2015, 12(7).DOI: 10.5194/bgd-12-5641-2015

Gregor L, Kok S, Monteiro P. Empirical methods for the estimation of Southern Ocean CO 2: support vector and random forest regression[J]. Biogeosciences, 2017, 14(23): 5551-5569. DOI: 10.5194/bg-14-5551-2017

Lohrenz S E, Cai W J, Chakraborty S, et al. Satellite estimation of coastal pCO2 and air-sea flux of carbon dioxide in the northern Gulf of Mexico[J]. Remote Sensing of Environment, 2018, 207: 71-83.DOI: 10.1016/j.rse.2017.12.039

Chen S, Hu C, Barnes B B, et al. A machine learning approach to estimate surface ocean pCO2 from satellite measurements[J]. Remote Sensing of Environment, 2019, 228: 203-226.DOI: 10.1016/j.rse.2019.04.019

Fu Z, Hu L, Chen Z, et al. Estimating spatial and temporal variation in ocean surface pCO2 in the Gulf of Mexico using remote sensing and machine learning techniques[J]. Science of the Total Environment, 2020, 745: 140965.DOI: 10.1016/j.scitotenv.2020.140965

Zhang S, Rutgersson A, Philipson P, et al. Remote sensing supported sea surface pCO2 estimation and variable analysis in the Baltic Sea[J]. Remote Sensing, 2021, 13(2): 259.DOI: 10.3390/rs13020259

Wang Y, Li X, Song J, et al. Carbon sinks and variations of pCO 2 in the Southern Ocean from 1998 to 2018 based on a deep learning approach[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 3495-3503.DOI: 10.1109/JSTARS.2021.3066552

Landschützer P, Gruber N, Bakker D C E, et al. Recent variability of the global ocean carbon sink[J]. Global Biogeochemical Cycles, 2014, 28(9): 927-949.DOI: 10.1002/2014GB004853

Landschützer P, Gruber N, Bakker D C E. Decadal variations and trends of the global ocean carbon sink[J]. Global Biogeochemical Cycles, 2016, 30(10): 1396-1417.DOI: 10.1002/2015GB005359

Gregor L, Lebehot A D, Kok S, et al. A comparative assessment of the uncertainties of global surface ocean CO< sub> 2</sub> estimates using a machine-learning ensemble (CSIR-ML6 version 2019a)–have we hit the wall?[J]. Geoscientific Model Development, 2019, 12(12): 5113-5136.DOI: 10.5194/gmd-12-5113-2019

Guorong Z, Xuegang L, Baoxiao Q, et al. A general regression neural network approach to reconstruct global 1°× 1° resolution sea surface pCO 2[J]. Haiyang Xuebao, 2020, 42(10): 70-79. DOI: 10.3969/j.issn.0253-4193.2020.10.007

Zhong G, Li X, Song J, et al. Reconstruction of global surface ocean pCO 2 using region-specific predictors based on a stepwise FFNN regression algorithm[J]. Biogeosciences, 2022, 19(3): 845-859.DOI: 10.5194/bg-19-845-2022

Wu H, Wang L, Ling X, et al. Spatiotemporal reconstruction of global ocean surface pCO2 based on optimized random forest[J]. Science of The Total Environment, 2024, 912: 169209.DOI: 10.1016/j.scitotenv.2023.169209

Ge W, Patino J, Todisco M, et al. Explaining deep learning models for spoofing and deepfake detection with

SHapley Additive exPlanations[C]//ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2022: 6387-6391.DOI: 10.1109/ICASSP43922.2022.9747476

Chau T T T, Gehlen M, Chevallier F. A seamless ensemble-based reconstruction of surface ocean pCO 2 and air–sea CO 2 fluxes over the global coastal and open oceans[J]. Biogeosciences Discussions, 2021, 2021: 1-30.DOI: 10.5194/bg-19-1087-2022

Chau T T T, Chevallier F, Gehlen M. Global analysis of surface ocean CO2 fugacity and air-sea fluxes with low latency[J]. Geophysical Research Letters, 2024, 51(8): e2023GL106670.DOI: 10.1029/2023GL106670