Spatiotemporal reconstruction of global ocean surface pCO2 based on optimized random forest

Huisheng Wu, Lejie Wang, Xiaochun Ling, Long Cui, Ruixue Sun, Na Jiang

PII: S0048-9697(23)07839-7
DOI: https://doi.org/10.1016/j.scitotenv.2023.169209
Reference: STOTEN 169209

To appear in: Science of the Total Environment

Received date: 12 September 2023
Revised date: 12 November 2023
Accepted date: 6 December 2023

Please cite this article as: H. Wu, L. Wang, X. Ling, et al., Spatiotemporal reconstruction of global ocean surface pCO2 based on optimized random forest, Science of the Total Environment (2023), https://doi.org/10.1016/j.scitotenv.2023.169209

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier B.V.
Spatiotemporal reconstruction of global ocean surface pCO$_2$ based on optimized random forest

Huisheng Wu $^{a,*}$, Lejie Wang $^a$, Xiaochun Ling $^b$, Long Cui $^a$, Ruixue Sun $^a$ and Na Jiang $^b$

$^a$College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, Shandong 266580, China; z22160024@s.upc.edu.cn (L.W.); z22160008@s.upc.edu.cn (L.C.); sunruixue_hy@163.com (R.S.)

$^b$Shandong Provincial Institute of Land Surveying and Mapping, Jinan 250102, China; 29107259@qq.com (X.L.); jiangna123321@163.com (N.J.)

*Corresponding author: wuhuisheng@upc.edu.cn

Abstract: The partial pressure of ocean surface CO$_2$ (pCO$_2$) plays an important role in quantifying the carbon budget and assessing ocean acidification. For such a vast and complex ocean system as the global ocean, most current research practices tend to study the ocean into regions. In order to reveal the overall characteristics of the global ocean and avoid mutual influence between zones, a holistic research method was used to detect the correlation of twelve predictive factors, including chlorophyll concentration (Chlor$_a$), diffuse attenuation coefficient at 490 nm (Kd$_{490}$), density ocean mixed layer thickness (Mlotst), eastward velocity (East), northward velocity (North), salinity (Sal), temperature (Temp), dissolved iron (Fe), dissolved silicate (Si), nitrate (NO$_3$), potential of hydrogen (pH), phosphate (PO$_4$), at the global ocean scale. Based on measured data from the Global Surface pCO$_2$ (LDEO) database, combined with National Aeronautics and Space Administration (NASA) Ocean Color satellite
data and Copernicus Ocean reanalysis data, an improved optimized random forest (ORF) method is proposed for the overall reconstruction of global ocean surface pCO₂, and compared with various machine learning methods. The results indicate that the ORF method is the most accurate in overall modeling at the global ocean scale (mean absolute error of 6.27μatm, root mean square error of 15.34μatm, R² of 0.92). Based on independent observations from the LDEO dataset and time series observation stations, the ORF model was further validated, and the global ocean surface pCO₂ distribution map of 0.25° × 0.25° for 2010 to 2019 was reconstructed, which is of significance for the global ocean carbon cycle and carbon assessment.

**Keywords:** Ocean surface pCO₂, Random Forests, Satellite remote sensing, Machine learning, Geographic detector, Global ocean

1. **Introduction**

   Increases in atmospheric carbon dioxide (CO₂) have led to global climate change. As one of the largest carbon reservoirs on earth, the ocean plays an important role in the response to and regulation of global climate change. Increases in atmospheric CO₂ are partly absorbed by the oceans, leading to changes in ocean acidity and alkalinity (i.e., the acidification effect), which has a significant impact on marine organisms, such as coral reefs, fish, shellfish, and plankton, as well as mollusks (Orr et al., 2005; Doney et al., 2009; Kroeker et al., 2013; Servili et al., 2023; Zhao et al., 2023). The partial pressure of ocean surface CO₂ (pCO₂) is an important indicator of environmental
quality and ocean health. When pCO$_2$ at the sea surface is greater than that in the atmosphere, CO$_2$ is released from the ocean into the air. In contrast, when ocean surface pCO$_2$ on is lower than that in the atmosphere, the ocean absorbs CO$_2$. Monitoring changes in ocean surface pCO$_2$ can help us understand the absorption and release of CO$_2$ in the atmosphere by the ocean, determine the rate and magnitude of the ocean's response to global climate change, and predict and assess the potential impacts of ocean acidification on marine ecosystems. Monitoring long-term changes in ocean surface pCO$_2$ also reflects changes in carbon content in the ocean and the distribution of carbon sinks and carbon sources, which provides an important basis for marine environmental management and policy formulation (Jung et al., 2023).

Oceans are large and complex systems. As actual pCO$_2$ measurement data for the sea surface cannot be fully obtained on a global scale, traditional estimation methods are limited by the amount of CO$_2$ measurement data in the atmosphere (Peylin et al., 2005; Rödenbeck, 2005). The ocean surface pCO$_2$ layer is affected by interactions of thermodynamic effects, biological activities, physical mixing processes, and air–sea exchange effects (Bai et al., 2015; Marrec et al., 2015; Lohrenz et al., 2018; Chen et al., 2019; Fu et al., 2020), which in turn are mainly affected by the pH, temperature, salinity, diffuse reflection attenuation coefficient, mixed layer depth, seawater velocity, and other factors. Advances in remote sensing technology have made it possible to study oceans on a global scale, as these variables can be derived from satellite data and related reanalysis data, providing a powerful tool for large-scale marine science.
In recent years, the application of machine learning in the field of oceanography has developed rapidly, providing effective and accurate solutions for monitoring, simulating, and predicting the ocean surface pCO$_2$. Telszewski et al. (Telszewski et al., 2009) studied the North Atlantic Ocean and trained a self-organizing map (SOM) neural network using chlorophyll a concentration, sea surface temperature, and mixed layer depth; the root mean square error (RMSE) of the fitted data was 11.6 μatm. Landschützer et al. (Landschuetzer et al., 2013) fitted the pCO$_2$ of the Atlantic Ocean using a self-organizing map feedforward network (SOM–FFN) combination method; however, the verification effect of independent observations fluctuated greatly. Nakaoka et al. (Nakaoka et al., 2013) used sea surface temperature, mixed layer depth, chlorophyll a concentration, and sea surface salinity to reconstruct pCO$_2$ in the North Pacific using SOM; the RMSE ranged from 17.6 to 20.2μatm. Chen (Chen et al., 2019) used sea surface temperature, sea surface salinity, chlorophyll a concentration, and the diffuse reflection attenuation coefficient (Kd) combined with various model algorithms to evaluate the Gulf of Mexico, which has a smaller sea area than those used in other studies, and found that random forest (RF) performed best. Fu et al. (Fu et al., 2020) also used these four variables combined with the Cubist machine learning method to improve on the results for the Gulf of Mexico; the overall performance had an RMSE of 8.42μatm and $R^2$ of 0.87. Moreover, the Indian and Southern oceans have been studied (Valsala et al., 2021; Wang et al., 2021). Although the above studies on individual sea areas have good degrees of fitting, the accuracy of model generalization to the global
ocean will be significantly reduced. Landschützer et al. (Landschützer et al., 2016) clustered the global ocean into multiple biogeochemical provinces using an SOM–FFN neural network and reconstructed the nonlinear relationship between driving factors and ocean surface CO₂ observation data for each region. Recently, researchers have combined a stepwise regression algorithm with a feed-forward neural network, defined multiple biogeochemical provinces using the SOM method, and selected different predictors in different regions to model and estimate ocean surface pCO₂. The mean absolute error (MAE) of global estimates is 11.32 μatm and the RMSE is 17.99 μatm (Zhong et al., 2022).

The goal of this study was to develop a machine learning method with general applicability for estimating global ocean surface pCO₂ and filling in some missing data. To this end, we used measured data from the Global Surface pCO₂ (LDEO) Database V2019, National Aeronautics and Space Administration (NASA) Ocean Color satellite data, and Copernicus ocean reanalysis data from 2010 to 2019. After selecting the pH, seawater temperature, seawater salinity, K_d, chlorophyll a concentration, mixed layer depth, seawater velocity, and various salt concentration as prediction factors and analyzing their spatial differences using geographic detectors, an optimized random forest (ORF) was proposed. Comparison with several different machine learning methods, including Multiple Linear Regression (MLR), Convolutional Neural Network (CNN), eXtreme Gradient Boosting (XGBoost), Support Vector Machines (SVM) and Random Forest (RF), our ORF has excellent performance in global ocean surface pCO₂
2. Data and Methods

2.1. Study area

This study was conducted in the global ocean, and different sea areas exhibit different characteristics. The Pacific Ocean has a complex circulation system, including the North Pacific Warm Current, the Equatorial Countercurrent, and the South Pacific Cold Current. Water temperature varies greatly, and there is an obvious temperature gradient from the tropics to the polar regions. The physical and chemical complexity of the Atlantic Ocean is high and includes many ocean current systems and circulations, such as the North Atlantic Current, North–South Equatorial Current, and Brazilian Current. Furthermore, the water depth varies significantly, ranging from shallow to deep troughs, which leads to differences in water temperature, salinity, and biodiversity between different regions. In addition to ocean currents, the Indian Ocean is affected by monsoons, monsoon storms, and tropical cyclones, which are more frequent in the region and affect the physical and biochemical processes of the ocean. The Southern Ocean surrounds Antarctica, and its complexity is mainly reflected in ocean circulation and marine ecosystems. Its cold water sinks from the surface to depth, forming the Southern Ocean circulation system, which plays an important role in global heat distribution and nutrient circulation in the ocean. The Arctic Ocean is located near the Arctic Circle and has a unique geographical location. It is affected by polar drift and ice
sheets, and sea ice coverage varies greatly from season to season. Additionally, the Arctic Ocean is one of the most sensitive areas to global climate change, and changes in temperature and salinity affect the rate of ice melting in the region. Different ocean regions have different characteristics that have different degrees of impact on ocean surface pCO$_2$. The study of the global ocean as a whole can reveal its overall characteristics, which is of great significance for an accurate understanding of ocean carbon sinks.

2.2. Data sources

2.2.1. Measured data

The measured data were obtained from the Ocean Carbon Data System (OCADS) Global Surface pCO$_2$ (LDEO) Database (OCADS - Global Surface pCO$_2$ (LDEO) Database (noaa.gov)). The LDEO dataset, which serves as a source for the SOCAT dataset, provides multiple types of ocean surface pCO$_2$ (Table 1). We selected pCO$_2$ measurements from this dataset for the period 2010 to 2019 with wide global spatial distribution coverage (Fig. 1). The data came from the high seas and coastal waters and were processed into a uniform data file through a series of steps. The collected data only included data measured by the balancer-CO$_2$ analyzer system and underwent strict quality control to ensure the stability of system performance, reliability of CO$_2$ analysis calibration, and internal data consistency, aiming to provide high-quality ocean carbon cycle observation data to promote the study of global climate change and ocean systems.
(Mkitarian, 2020). The equilibrator temperature lags the in situ temperature owing to the transfer time of seawater pumped from the bow to the pCO₂ system and the residence time of seawater in the equilibration tank at the time of the actual measurement. Therefore, we selected pCO₂ TEQ in this dataset as the measured data, which represents the actual measured pCO₂ results.

### 2.2.2. Satellite data and reanalysis data

The NASA Ocean Color project uses satellite sensors to collect global ocean information, provides accurate and reliable data products through advanced remote sensing technology and data processing methods, and is dedicated to the interpretation and analysis of ocean data (NASA Ocean Color). In this study, the 4×4km resolution global Level 3 products of chlorophyll concentration (Chlor_a) and diffuse attenuation coefficient at 490 nm (Kd_490) from the imaging spectroradiometer (MODIS) Aqua sensor were utilized.

Based on data sources from similar studies (Yu et al., 2023), reanalysis data were derived from the Copernicus Ocean Data Product (Copernicus Marine Data Store | Copernicus Marine MyOcean Viewer). It provides rich four-level data products based on the fusion of satellite remote sensing data and observational data. The research uses monthly products with a resolution of 0.083°× 0.083°, including density ocean mixed layer thickness (Mlotst), eastward velocity (East) and northward velocity (North), salinity (Sal), and temperature (Temp). We also use monthly products with a resolution
of 0.25°×0.25°, including dissolved iron (Fe), dissolved silicate (Si) and nitrate (NO₃), potential of hydrogen (pH), and phosphate (PO₄).

In summary, a total of twelve variables were selected for the study on global oceanic carbon dioxide partial pressure: the level-3 products include Chlor_a and Kd_490, while the level-4 products include Mlotst, East, North, Sal, Temp, Fe, Si, NO₃, pH, and PO₄ (Table 2). These data possess good spatiotemporal resolution and strict quality control, meeting our requirements for oceanic information, and providing a reliable database for further marine research.

2.2.3. Data processing

Owing to differences in the spatial resolution of various products, the nearest-neighbor method was used to resample and unify the resolution to 0.25°×0.25°. The measured data were linked to the satellite and reanalysis data of each variable through coordinate association, which was convenient for their analysis within the same framework.

There are differences in the amount of data in space and time among different types of data. In terms of spatial distribution, although the global coverage of the data was relatively wide, the data densities in the different regions varied (Fig. 1). For example, during 2010, the Atlantic Ocean north of the equator contained 130000 measured data, while the Atlantic Ocean south of the equator contained 70000 measured data, resulting in uneven data coverage density. Overall, from 2010 to 2019,
the data density in the North Pacific and North Atlantic is higher than in other oceans, while the data near the Arctic Ocean is the sparest. In terms of time, the amount of data measured in different years was not uniform (Fig. 2). The data volume in 2014 and 2015 exceeded 1 million data points, while the data volume in 2019 only exceeded 200000.

If data are selected and processed only according to the spatial distribution, it will cause an uneven distribution of data in the year and month, which would interfere with the accuracy of ocean surface pCO$_2$ modeling and lead to inaccurate interannual variation analysis. If data are selected only according to the time distribution, it will lead to an uneven spatial distribution of the selected data, resulting in some sea areas not covered or less covered, and other sea areas densely covered, which will also affect the accuracy of the model and make the spatial analysis inaccurate. Therefore, it was necessary to properly screen the data prior to model training. The measured data were collected at different locations and times; that is, time and space were associated, and the density of data in space corresponded to the amount of data in time. Combined with the distribution area and distribution density of these data in the global ocean, an appropriate selection was made such that the selected data could represent the global ocean to a great extent. Create mutually independent training, validation, and testing sets from this data for use in this study.
2.3. Research methods

2.3.1. Geographic detector

Spatially stratified heterogeneity refers to differences or changes between different places or regions within a certain spatial range. At present, although many algorithms such as K-means and SOM, can be used for classification or partitioning, statistical methods for spatial differentiation are still very limited. Geographic detectors provide a statistical method that can break these defects with good results, help detect and analyze spatial variability, and reveal the driving force behind spatial variability (Wang and Xu, 2017).

Differentiation and factor detection were used to explain the degree of spatial differentiation of a variable to pCO$_2$, measured by q value, which is expressed as follows:

\[
q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N \sigma^2} = 1 - \frac{SSW}{SST} \quad 1)
\]

\[
SSW = \sum_{h=1}^{L} N_h \sigma_h^2 \quad 2)
\]

\[
SST = N \sigma^2 \quad 3)
\]

In the formula: $h=1,\ldots, L$ is the number of layers of ocean surface pCO$_2$ or the number of layers of the prediction factor; $N_h$ and $N$ are the number of units in layer $h$ and the whole area, respectively; $\sigma_h^2$ and $\sigma^2$ are the variances of the pCO$_2$ values in
layer $h$ and the whole area, respectively; and $SSW$ and $SST$ are the sum of the intra layer variances (within sum of squares) and the total sum of squares (total sum of squares), respectively. By comparing the $q$ value of each variable, the dominant factor affecting $pCO_2$ was determined.

Interaction detection determined how the effects of different variables on $pCO_2$ interact by analyzing the interactions between variables. The method was to calculate $q(X1)$ and $q(X2)$ of two variables, $X1$ and $X2$, for $pCO_2$, and then to calculate the value of $q(X1 \cap X2)$ when the two variables interact. This interaction detection helped us judge whether the two factors enhanced or weakened the impact on the target, or whether they were independent, in order to understand the relationship between variables more deeply and reveal the causal relationships behind them.

### 2.3.2. Model selection and the principle of ORF

MLR has been used to estimate local ocean surface $pCO_2$ in previous studies (Chierici et al., 2012) and has been used as a reference benchmark to evaluate the performances of models. CNN, XGBoost, and SVM were used for comparison. RF is an ensemble learning method that combines multiple decision trees for training and prediction by randomly selecting features and samples. In an RF, each decision tree is constructed based on a randomly selected subset of features and samples. Specifically, for the training set of each decision tree, the RF draws the same number of samples from the dataset with replacement and randomly selects a part of the features as
candidate features. This random selection ensured that each decision tree is independent and differs only slightly. Each decision tree was trained according to the selected features and samples to establish a prediction model, and the prediction results of each decision tree were averaged to obtain the overall pCO₂ prediction results (Liaw and Wiener, 2001). RF makes full use of the inference ability of multiple decision trees and reduces the risk of overfitting through this integration method; each decision tree is independent; thus, it has good anti-noise ability and stability and has been widely used in previous studies (Chen et al., 2019; Zhang et al., 2021).

As the performance of RF is affected by many hyperparameters, such as the number of trees, maximum depth of each tree, and minimum sample number of each node, to optimize the accuracy of the model, a bayesian optimization algorithm was added to the RF to form the ORF used in this study. Bayesian optimization uses the results to update the posterior probability distribution of the hyperparameters by evaluating the performance metrics of each hyperparameter combination of the RF. The potential optimal hyperparameter configuration is searched step-by-step by continuously adjusting the hyperparameter combination that is most likely to lead to optimal performance in each iteration.

Specifically (Fig. 3), bayesian optimization uses a gaussian process model to estimate the relationship between hyperparameters and performance indicators and establishes a gaussian process model between existing and new candidate points in each iteration to predict performance indicators. This model uses previous observations
to estimate the joint probability distribution between the hyperparameters and performance index. Using this joint probability distribution, the posterior probability distribution for each hyperparameter combination was calculated to quantify its merits of the hyperparameter combination. The sampling function method was used to select the next combination of hyperparameters to be evaluated based on the posterior probability distribution. The sampling function considers the tradeoff between maximizing the predictive performance and maximizing the uncertainty in selecting the hyperparameter combination that is most likely to lead to optimal performance. Through this iterative process, the posterior probability distribution is continuously updated to gradually converge to the optimal hyperparameter configuration (Mockus et al., 1978; Snoek et al., 2012). Therefore, ORF can better deal with high-dimensional, complex, and huge data, improve the performance and generalization ability of the model, and better meet the needs of the complex task of reconstructing global ocean surface pCO$_2$.

2.3.3. Model and accuracy evaluation

During the model development phase, MLR, CNN, XGBoost, SVM, RF, and ORF were used to construct different models. The MAE, RMSE, and coefficient of determination ($R^2$) of the models on the training, validation, and test sets were used to comprehensively evaluate model accuracy. For the developed model, a dataset constructed independently from the model (treated in the same way as in Section 2.2.3)
was input into different models to calculate the output value and compared it with the real value for performance evaluation. Further coastal validation was conducted on the best performing model to test its ability to reconstruct surface pCO₂ in coastal areas. On this basis, it is validated based on independent observations and compared with time series observations for evaluation (Gregor et al., 2019; Zhong et al., 2021).

3. Results

3.1 Spatial differentiation

3.1.1 Differentiation and factor detection

In the study, geographic detectors were used to independently calculate the spatial heterogeneity between all prediction factors and the partial pressure of carbon dioxide on the ocean surface. Due to the large area and scale of the global ocean as the research object, the overall q value is relatively small. Therefore, it is believed that variables with q value > 0.1 have significant spatial differentiation across a wide range of global oceans. From the experimental results (Fig. 4), there was obvious spatial differentiation between both pH and temperature and pCO₂ of the ocean surface, with q value > 0.2. For salinity, the diffuse reflection attenuation coefficient, and phosphate and chlorophyll a concentration, q value were > 0.1. The q value of nitrate, silicate, iron, and the mixed layer thickness were > 0.06, and the q value of the seawater flow rate was > 0.03. In summary, the variables selected in this study had certain spatial differentiation.
3.1.2 Interaction detection

The interaction detection results (Fig. 5) showed that there was a strong interaction between the selected variables. Potential of hydrogen had the strongest interaction effect on the other variables, and most of the interaction values were $> 0.7$. The interaction effect of the density ocean mixed layer thickness on other variables was relatively weak, but it still exceeded 0.4. The interaction between chlorophyll a concentration and diffuse reflection attenuation coefficient is 0.31; however, most of the interactions between the other variables were $> 0.5$. The interaction of these variables can enhance their effect on ocean surface pCO$_2$, meaning that when these variables act together, their effect is more significant than when they act alone.

3.2 Model performance and validation

We model the dataset using different machine learning methods. Owing to the large amount of data, we randomly selected some data from all the fitting results to show the observation performance (Fig. 6). The difference and fluctuation between the fitted values calculated by MLR and CNN and the true values are the greatest, with a fitting effect of 300 to 400μatm being better, while the fitting effect is poor for parts below or above this range; The fitting results calculated by XGBoost and SVM show some improvement in the low and high value parts compared to the previous two methods, but overall, their performance is not excellent. The difference between the
fitted values calculated using RF and ORF and the true values is the smallest and the overall performance is the most stable. We simultaneously calculated the relevant indicators of the different models in the training, validation, and test sets (Table 3 and Fig. 7). MLR has the worst performance among all models (MAE of 33.14μatm, RMSE of 48.95μatm, \( R^2 \) of 0.18), and was used as the baseline for model performance comparison. Comparatively, SVM, XGBoost, and CNN were better than MLR, but not as good as RF and ORF. RF exhibited good performance; however, ORF exhibited the best performance, confirming its suitability for global \( p\text{CO}_2 \) reconstruction under complex conditions (MAE of 6.27μatm, RMSE of 15.34μatm, \( R^2 \) of 0.92).

To further evaluate these models and confirm whether the ORF model was indeed the best, all the models were rigorously validated independently based on a dataset constructed independently of the model. Through a comparison of the results (Fig. 8), only some of the output values of the MLR and CNN models were more accurate than the true values; the rest deviated significantly; the distribution of lateral divergence was on the \( y = x \) line. The deviation between the output values of the XGBoost and SVM models and the true values was gradually reduced, but there were still many instances where the prediction gap was large.

The RF model has made significant improvements compared to the above models, with a large number of output values close to the true value and beginning to converge towards the \( y = x \) line. However, ORF had the best performance, and almost all the output values were fitted with the true values, concentrated on the \( y = x \) line. In
summary, the convergence effect was the best, and the prediction effect was the most accurate and stable. All validation results showed that ORF had the best performance compared with the other models.

3.3 Coastal verification and independent observation verification

Based on the ORF model with the best performance, its accuracy in coastal areas was further validated in the study. A large amount of LDEO measurement data was extracted from six coastal areas (Fig. 9) and compared with the output values of the model for validation. The results indicate that the output values of the ORF model in coastal areas are close to the pCO$_2$ values measured by LDEO (MAE=11.64µatm, RMSE=17.74µatm), and are concentrated on y = x line (Fig. 10).

The output values of the ORF model were compared with the independent observation results of HOT (Dore et al., 2009) and BATS (Bates, 2007). Compared with the independent observation results of HOT station, the interannual changes between the two are similar from 2010 to 2019. The MAE between the ORF output value and the HOT observation value is 8.31µatm, and the RMSE is 12.75µatm (Fig. 11a). Similar to HOT station, the interannual variation trend of BATS station and BATS station is basically consistent, with MAE of 8.17µatm and RMSE of 13.35µatm (Fig. 11b).
4. Discussion

4.1. Analysis of variable action

According to an analysis of the experimental results of the geographic detector, there is a mechanism of mutual influence between the variables, and a complex relationship with the target. CO$_2$ reacts with water (H$_2$O) in seawater to form carbonic acid (H$_2$CO$_3$), which further dissociates into hydrogen (H$^+$) and carbonate (HCO$_3^-$) ions. The change in pH directly affects the equilibrium state of the H$_2$CO$_3$ equilibrium system. When pH decreases, the number of H$^+$ in seawater increases, resulting in the weakening of the buffering effect of HCO$_3^-$ ions on H$^+$, which leads to an increase in pCO$_2$, and vice versa. Various salts in the ocean affect the pH value through interactions, which then affect the pCO$_2$. Chlorophyll, a pigment found in both plants and marine phytoplankton, plays a key role in photosynthesis, converting CO$_2$ into organic matter and releasing oxygen at the same time. Photosynthesis is one of the main CO$_2$ inhibition mechanisms in marine ecosystems. The change in chlorophyll concentration also affects the gas exchange rate between CO$_2$ and the atmosphere, and further affects ocean surface pCO$_2$. Changes in the ocean current velocity can affect ocean surface pCO$_2$ via two aspects: physical and biological processes. In terms of physical processes, an increase in ocean velocity can accelerate the rate of gas diffusion between the ocean and atmosphere. When the ocean flows faster, water in the surface layer will quickly release CO$_2$ into the atmosphere, resulting in a decrease in the pCO$_2$ of the surface layer. In contrast, when ocean flow is slow, the release rate of CO$_2$ from surface water slows,
and the pCO₂ of the surface of the ocean increases. In terms of biological processes, changes in ocean velocity also affect the activities and ecosystems of marine organisms, change the distribution and ecological chain structure of marine organisms, and affect ocean surface pCO₂. The effect is most complex for temperature, which controls the absorption of CO₂ by the ocean by affecting the solubility of the gas in seawater. The absorption and release of CO₂ by marine organisms are also affected by temperature, which affects the pattern of ocean surface currents and indirectly affects pCO₂.

4.2. Global pCO₂ product comparison of ocean surface

In this study, satellite and reanalysis data were used to reconstruct the annual distribution of 0.25° × 0.25° global ocean surface pCO₂ using the ORF model. Moreover, the pCO₂ data provided by the Copernicus Global Ocean Surface Carbon Products were used to generate an annual distribution map, which was compared with the pCO₂ annual product reconstructed in this study, it can be seen that in addition to the improved resolution, the Arctic Ocean region has also undergone reconstruction (Fig. 12). The global distribution pattern based on the ORF reconstruction was consistent with the global distribution pattern in other products, confirming that ORF can accurately reconstruct the spatial distribution of global pCO₂. Furthermore, we selected a large number of points with the same location for the two products in different years and compared the average values of these points to understand the interannual trend of the two products. The change curves of the two products are almost synchronous,
showing the same growth or decline trend over a decade (Fig. 13).

4.3. Spatiotemporal change analysis

In this study, we reconstructed a global ocean surface $pCO_2$ distribution map from 2010 to 2019 (Fig. 14). Additionally, the study has drawn profiles of the Pacific, Atlantic, and Indian Oceans from north to south, and profiles of the Southern Ocean and the Arctic Ocean from west to east, and analyzed the temporal and spatial variations of ocean surface $pCO_2$ in five sea areas (Fig. 15).

Partial pressure of carbon dioxide in the Pacific Ocean is significantly higher than that in other sea areas. This is because more $CO_2$ is released into the atmosphere owing to strong ocean circulation and mixing in the Pacific region, as well as the influence of industrial activities, human emissions, and land-based pollution, $pCO_2$ peaks near the equator and gradually decreases in the form of fluctuations with increasing latitude. Low latitudes areas of the Pacific Ocean exhibited a gradual growth trend from 2010 to 2019, whereas $pCO_2$ at high latitudes fluctuated.

In the Atlantic Ocean, $pCO_2$ is higher at equatorial low latitudes, in the North Atlantic, and in the subtropical South Atlantic. Partial pressure of carbon dioxide values are low in the northern North Atlantic and southern South Atlantic, and in some mid-latitudes; these results are similar to those of previous studies (Landschuetzer et al., 2013). In terms of time variation, the interannual variation of ocean surface $pCO_2$ in the North Atlantic and South Atlantic fluctuates, but overall shows an upward trend year by
Partial pressure of carbon dioxide in the Arabian Sea of the northwest Indian Ocean is significantly higher than that in the Bay of Bengal in the northeast Indian Ocean, and the pCO₂ decreases slightly with decreasing latitude. The change between 2010 and 2019 showed a steady annual upward trend.

Despite belonging to high latitudes, pCO₂ in the Southern and Arctic Oceans are not the lowest. This is because the melting of ice increases the release of CO₂ in seawater, leading to higher pCO₂ near Antarctica and north of the Arctic Circle. Overall, the Southern Ocean showed a steady increase in pCO₂ until 2018, consistent with previous studies (Wang et al., 2021); the Arctic Ocean showed a small increase with fluctuations until 2018. Partial pressure of carbon dioxide in the Southern Ocean and Arctic Ocean began to show a downward trend in 2019.

From a global perspective, ocean surface pCO₂ around continents is higher than that in the high seas. This reflects the dense population and relatively high level of industrialization on the continents. Industrial production, transportation, and energy consumption produce large amounts of CO₂ emissions, which gradually diffuse into the atmosphere and eventually dissolve into the ocean. Simultaneously, rivers transport large volumes of CO₂-rich freshwater into the ocean. Marine ecosystems near continents are relatively prosperous and have more biological activity; for this reason, marine organisms produce more CO₂ through respiration. Therefore, offshore pCO₂ is higher than offshore pCO₂.
5. Conclusions

For large-scale and complex ocean systems like the global ocean, most studies tend to partition the ocean. The aim of this study is to develop a model that can be applied globally and has broad generalization without the need to partition the ocean. We collected predictive variables that affect carbon dioxide partial pressure from previous research and experimental data, and calculated its spatial differentiation at the global ocean scale using geographic detectors. Compared with various machine learning models, the ORF model proposed in this study has superior performance in predicting pCO₂ in the global ocean as a whole. Based on measured data from the LDEO database, combined with NASA Ocean Color Level-3 data and Copernican Ocean Level-4 data, the distribution of global ocean surface pCO₂ from 2010 to 2019 was reconstructed using ORF. Compared with similar products, the spatial distribution pattern and temporal trend of the reconstructed products are consistent, and the research results are reliable.

Through a holistic research approach to the ocean, we aim to reveal the overall characteristics of the global ocean and capture global interactions and dynamic changes, in order to better understand the operational mechanisms of the ocean. Compared to zoning research, the overall research method can provide a comprehensive and comprehensive perspective. Zonal research can easily lead to local understanding, while holistic research can consider the ocean system as a whole, avoiding overlooking
the mutual influence between different regions. This comprehensive perspective helps us gain new insights and provides useful references for subsequent scientific research in related fields, thereby contributing to the protection and management of marine ecosystems.

In future work, we will use the global ocean zoning method from other studies to reconstruct the global ocean surface pCO₂. At the same time, we will compare and analyze the reconstructed results, and explore the advantages and disadvantages of each regional research method and the overall research method.

Acknowledgements

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.

Declaration of competing interests

No potential conflicts of interest were disclosed by all authors.

Author contributions

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

References


https://doi.org/10.1146/annurev.marine.010908.163834


https://doi.org/10.1073/pnas.0906044106


https://doi.org/10.1016/j.scitotenv.2020.140965


https://doi.org/10.5194/gmd-2019-46


https://doi.org/10.1016/j.scitotenv.2023.166270


https://doi.org/10.1111/gcb.12179


https://doi.org/10.5194/bg-10-7793-2013


https://doi.org/10.1016/j.rse.2017.12.039


https://doi.org/10.5194/bg-12-5371-2015


https://doi.org/10.5194/bg-10-6093-2013

Orr, J.C., Fabry, V.J., Aumont, O., Bopp, L., Doney, S.C., Feely, R.A., Gnanadesikan, A., Gruber, N., Ishida, A., Joos, F., Key, R.M., Lindsay, K., Maier-Reimer, E., Matear, R., Monfray, P., Mouchet, A., Najjar, R.G.,
https://doi.org/10.1038/nature04095

https://doi.org/10.1029/2003GB002214


Figures

**Fig. 1.** The global distribution of measured data from 2010 to 2019 (shown in grey).

**Fig. 2.** Histogram of quantity distribution of measured data in each year from 2010 to 2019.
Fig. 3. Optimize random forest process structure diagram (Starting with data set partitioning, random forest hyperparameter adjustment and bayesian optimization were introduced, and parameter indicators of model performance were added for evaluation, and the optimal random forest model was used in the marine pCO₂ reconstruction task of this study).
Fig. 4. Single-factor detection at the global ocean scale, where q value represents the spatial heterogeneity strength between ocean surface pCO$_2$ and various predictive factors.

Fig. 5. Interaction detection at the global ocean scale, this value represents the strength of the influence of the combined interaction between the predictors on the
ocean surface pCO$_2$. 
Fig. 6. The performance of different models on the same dataset is evaluated using independent training, validation, and test sets (a partial representation of the data fitting is randomly shown in the graph).

Fig. 7. Model performance index chart, (a), (c) and (e) are line graphs of mean
absolute error and root-mean-square error of different models on the training set, verification set and test set respectively; (b), (d) and (f) are pie charts of Coefficient of Determination of different models on the training set, verification set and test set respectively.

Fig. 8. The performance between the output values of different models and the
actual values, with colors representing the degree of binary fitting.

Fig. 9. Further validation of the model by selecting measured data from six coastal areas (the latitude and longitude information is shown in the figure).
Fig. 10. Accuracy performance of ORF model in coastal areas.
Fig. 11. Validation based on independent observation from time series stations, (a) the Hawaii Ocean Time-series, (b) the Bermuda Atlantic Time-series.

Fig. 12. Comparing the global ocean surface pCO$_2$ products, the Copernicus global ocean surface pCO$_2$ products are shown on the left, and the optimized random forest reconstruction products are shown on the right.

Fig. 13. Interannual trend comparison between Copernicus global ocean surface pCO$_2$ products and optimized random forest reconfiguration products.
Fig. 14. The partial pressure of carbon dioxide on the global ocean surface during 2010-2019 was obtained in this study.
Fig. 15. Ocean surface pCO$_2$ changes in different sea areas from 2010 to 2019 (the horizontal axis represents the direction of cross-sections).
### Tables

**Table 1.** Types of pCO$_2$ data provided by Global Surface pCO$_2$ (LDEO) Database V2019.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>pCO$_2$ _TEQ</td>
<td>The partial pressure of CO$_2$ (in microatmospheres) in Marine water under a temperature column that measures the partial pressure of CO$_2$ in °C, which is usually the value actually measured</td>
</tr>
<tr>
<td>pCO$_2$ _SST</td>
<td>Partial pressure of CO$_2$ in sea water under sea surface temperature (°C) column (in microatmospheres)</td>
</tr>
<tr>
<td>pCO$_2$ _SSTPA</td>
<td>The partial pressure of carbon dioxide in seawater (in pascals) is the temperature under the sea surface temperature (°C) column</td>
</tr>
</tbody>
</table>

**Table 2.** List of satellite data and reanalysis data.

<table>
<thead>
<tr>
<th>Full name</th>
<th>Abbreviation</th>
<th>Resolution</th>
<th>Producer</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density ocean mixed layer thickness</td>
<td>Mlöst</td>
<td>0.08° × 0.08°, monthly, 2010–2019</td>
<td>CMEMS Global Monitoring and Forecasting Centre</td>
<td><a href="https://doi.org/10.48670/moi-00021">https://doi.org/10.48670/moi-00021</a></td>
</tr>
<tr>
<td>Eastward velocity</td>
<td>East</td>
<td>0.08° × 0.08°, monthly, 2010–2019</td>
<td>CMEMS Global Monitoring and Forecasting Centre</td>
<td><a href="https://doi.org/10.48670/moi-00021">https://doi.org/10.48670/moi-00021</a></td>
</tr>
<tr>
<td>Northward velocity</td>
<td>North</td>
<td>0.08° × 0.08°, monthly, 2010–2019</td>
<td>CMEMS Global Monitoring and Forecasting Centre</td>
<td><a href="https://doi.org/10.48670/moi-00021">https://doi.org/10.48670/moi-00021</a></td>
</tr>
<tr>
<td>Salinity</td>
<td>Sal</td>
<td>0.08° × 0.08°, monthly, 2010–2019</td>
<td>CMEMS Global Monitoring and Forecasting Centre</td>
<td><a href="https://doi.org/10.48670/moi-00021">https://doi.org/10.48670/moi-00021</a></td>
</tr>
<tr>
<td>Temperature</td>
<td>Temp</td>
<td>0.08° × 0.08°, monthly, 2010–2019</td>
<td>CMEMS Global Monitoring and Forecasting Centre</td>
<td><a href="https://doi.org/10.48670/moi-00021">https://doi.org/10.48670/moi-00021</a></td>
</tr>
<tr>
<td>Dissolved Iron</td>
<td>Fe</td>
<td>0.25° × 0.25°, monthly, 2010–2019</td>
<td>CMEMS Global Monitoring and Forecasting Centre</td>
<td><a href="https://doi.org/10.48670/moi-00019">https://doi.org/10.48670/moi-00019</a></td>
</tr>
<tr>
<td>Dissolved Silicate</td>
<td>Si</td>
<td>0.25° × 0.25°, monthly, 2010–2019</td>
<td>CMEMS Global Monitoring and Forecasting Centre</td>
<td><a href="https://doi.org/10.48670/moi-00019">https://doi.org/10.48670/moi-00019</a></td>
</tr>
</tbody>
</table>
Table 3. Specific performance data of different models on training, validation, and testing sets (the precision value refers to the overall reconstruction accuracy within the global ocean range).

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE (μatm)</th>
<th>RMSE (μatm)</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training process</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF</td>
<td>3.23</td>
<td>9.41</td>
<td>0.97</td>
</tr>
<tr>
<td>RF</td>
<td>9.33</td>
<td>17.65</td>
<td>0.90</td>
</tr>
<tr>
<td>SVM</td>
<td>18.45</td>
<td>33.43</td>
<td>0.62</td>
</tr>
<tr>
<td>XGBoost</td>
<td>22.77</td>
<td>37.24</td>
<td>0.53</td>
</tr>
<tr>
<td>CNN</td>
<td>28.08</td>
<td>44.89</td>
<td>0.32</td>
</tr>
<tr>
<td>MLR</td>
<td>33.25</td>
<td>49.37</td>
<td>0.18</td>
</tr>
<tr>
<td><strong>Validation process</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF</td>
<td>6.31</td>
<td>15.29</td>
<td>0.92</td>
</tr>
<tr>
<td>RF</td>
<td>11.37</td>
<td>22.58</td>
<td>0.83</td>
</tr>
<tr>
<td>SVM</td>
<td>18.69</td>
<td>33.86</td>
<td>0.61</td>
</tr>
<tr>
<td>XGBoost</td>
<td>22.70</td>
<td>37.36</td>
<td>0.53</td>
</tr>
<tr>
<td>CNN</td>
<td>27.91</td>
<td>44.57</td>
<td>0.33</td>
</tr>
<tr>
<td>MLR</td>
<td>33.01</td>
<td>48.95</td>
<td>0.18</td>
</tr>
<tr>
<td><strong>Testing process</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF</td>
<td>6.27</td>
<td>15.34</td>
<td>0.92</td>
</tr>
<tr>
<td>RF</td>
<td>11.34</td>
<td>21.63</td>
<td>0.84</td>
</tr>
<tr>
<td>SVM</td>
<td>18.73</td>
<td>33.50</td>
<td>0.62</td>
</tr>
<tr>
<td>XGBoost</td>
<td>22.70</td>
<td>36.94</td>
<td>0.54</td>
</tr>
<tr>
<td>CNN</td>
<td>27.96</td>
<td>44.16</td>
<td>0.34</td>
</tr>
<tr>
<td>MLR</td>
<td>33.34</td>
<td>48.95</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Declaration of 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.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

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.
Graphical abstract
Highlights

• Conduct comprehensive research on the global ocean surface pCO$_2$ from a holistic perspective.
• Analyze ocean surface pCO$_2$ and twelve factors of the ocean using a geographical detector.
• Compare multiple machine learning models of pCO$_2$, including MLR, CNN, XGBoost, SVM, RF, and ORF.
• Reconstruct the resolution of ocean surface pCO$_2$ to $0.25^\circ \times 0.25^\circ$.
• The ORF model improves the reconstruction accuracy of ocean surface pCO$_2$. 