 doi:10.20944/preprints202104.0065.v1

Article

Estimating Sea Surface pCO_2 in the North Atlantic based on CatBoost

Hongwei Sun [1,*], Yihui Chen [1], Lin Li [1] and Boyu Zhao [2]

[1] School of Oceanography, Zhejiang University, 1 Zheda RD, Dinghai District, Zhoushan 316021, China; 21834006@zju.edu.cn
[2] School of Economics, Zhejiang University, 866 Yuhangtang RD, Xihu District, Hangzhou 310058, China; 21901414@zju.edu.cn
* Correspondence: 21834006@zju.edu.cn; Tel: +86-183-2661-1517

Abstract: Sea surface partial pressure of CO_2 (pCO_2) is a critical parameter in the quantification of air-sea CO_2 flux, which plays an important role in calculating the global carbon budget and ocean acidification. In this study, we use chlorophyll-a concentration (Chla), sea surface temperature (SST), absorption due to dissolved and particulate detrital matter (Adg), diffuse attenuation coefficient of downwelling irradiance at 490nm (Kd) and mixed layer depth (MLD) as input data for retrieving the sea surface pCO_2 in the North Atlantic based on a remote sensing empirical approach with the Categorical Boosting (CatBoost) algorithm. The results show that the root mean square error (RMSE) is 8.25μatm, the mean bias error (MAE) is 4.92μatm and the coefficient of determination (R^2) can reach 0.946 in the validation set, which mean that the CatBoost model makes an improvement compared to other models in the published studies. In the further analysis of the spatial and temporal distribution of the sea surface pCO_2 in the North Atlantic, it can be found that the North Atlantic sea surface pCO_2 has a clear trend with latitude variations and have strong seasonal changes. Furthermore, the sea surface pCO_2 in this area is mainly affected by sea temperature and salinity, and influenced by biological activities in some sub-regions.

Keywords: sea surface pCO_2; ocean color remote sensing; CatBoost algorithm; temporal and spatial distribution; influencing factors

1. Introduction

Ocean is one of the destinations for anthropogenic carbon, which takes up about 30% of emissions from pre-industrial times to 1994 [1]. Besides that, CO_2 circulation between air and sea plays an important role in the global carbon budget [2-4]. Atmospheric CO_2 has been increased by 40% because of fossil fuels and the oceanic uptake of CO_2 has been increased by 30% relatively [5-7]. In recent years, regional and global air-sea CO_2 flux has received a lot of attention [8-11], due to more and more measured data about sea surface partial pressure of carbon dioxide (pCO_2).

In recent years, more and more studies about sea surface pCO_2 have achieved some good results with the help of satellite remote sensing on account of its spatial and temporal resolution and coverage. In practice, surface pCO_2 is controlled by four major factors: the thermodynamic process, physical mixing between different water masses, biological activities, and the air-sea gas exchange [12-14]. According to these processes, some satellite-derived parameters can be used in the study for their features are closely related to them. For example, sea surface temperature (SST, ℃) can reflect the thermodynamic process directly. Besides that, some inherent optical properties (IOPs) and apparent optical properties (AOPs), such as absorption due to dissolved and particulate detrital matter (Adg, m^{-1}), particulate backscattering (bbp, m^{-1}), absorption by phytoplanktonic (Aph, m^{-1}) and Diffuse Attenuation Coefficient of downwelling irradiance (Kd, m^{-1}), can measure the mass of carbon which influences the sea surface pCO_2 through the process of water mixing and biological activities. In addition to these factors, some elements such as sea surface salinity (SSS, dimensionless) and surface chlorophyll-a concentration (Chla,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

$mg \cdot m^{-3}$) can be deduced by the biological activities and other variables like wind speed ($m \cdot s^{-1}$) and mixed layer depth (MLD, m) based on the process of the air-sea gas exchange are used in the study of pCO_2 as well [8, 10, 15-20]. Specifically, first of all, the process of the horizon and vertical mixing among the water masses, which have different characteristics such as total alkalinity (TA, $\mu mol \cdot kg^{-1}$) and dissolved inorganic carbon (DIC, $\mu mol \cdot kg^{-1}$) [21, 22], can influence the distribution of sea surface pCO_2, and SSS and SST can be calculated in a carbonate system [23]. Secondly, respiration, photosynthesis and calcification can affect surface pCO_2 directly because they can deplete both TA and DIC in a 2 to 1 ratio [10, 24]. Furthermore, extreme events like hurricane and storms also affect surface pCO_2 through the process of air-sea CO_2 exchange [25-27].

In limited cases, only one or two processes control the changes of surface pCO_2 [8]. There are many efforts have been made to estimate sea surface pCO_2 based on the regression analyses, including linear regression, multiple regression and nonlinear regression, to find a correlation between pCO_2 and less than three factors [28-35]. However, in a complex ocean environment, it is difficult to find a relationship based on these regression methods because many processes are interacted [36], so a simple and direct relationship between multiple factors is not existed. In recent years, some non-traditional empirical methods mainly about machine learning methods are used in the study such as multilayer perceptron neural network, self-organizing mapping, supporting vector machines, principle component regression and random forests [9, 18, 37-40]. In these studies, the model produced to predicting pCO_2 can have relative precise results in specific regions.

Although these methods had made a progress in the study of estimating surface pCO_2, the study is still a hard task because of the complexity and dynamics in the physical process in the open ocean water. Some problems need to be solved in the region which dominated by multiple processes. Specifically, Hales developed a semi-analytical method for the US West Coast but the model of the specific parameters may have poor applicability for other regions [41]. Bai proposed a mechanistic semi-analytic algorithm makes the model more applied in other regions and improve the accuracy across short timescales and small spatial scales [8], but the ocean process is difficult to quantify in practice by the algorithm. According to Chen, a method based on regression tree and ensemble learning[9], which is called RFRE, had great potential to be a robust approach for regional pCO_2 modeling in the Gulf of Mexico across many different water types, yet it may be poor in decadal long-term scale or in the region which has non-optimal satellite observing conditions.

In our study, the main purpose, is to overcome the problems in published studies and using a new empirical method to develop a model with a more precious result and general applicability. We select a new machine-learning method called Categorical Boosting (CatBoost) which often improves model performance in the regression learning in other subjects [42-45]. The aim of this method is to get a good performance in the North Atlantic and can be generalized in other regions all over the world. Apart from that, we will analyze the distribution and variation of the surface pCO_2 in the North Atlantic based on the results of our model.

2. Materials and Methods

2.1. Study region

The region of the North Atlantic, bounded by 15°N–55°N and 80°W–0°W, was selected in this study. In recent years, the North Atlantic region has become a research hotspot, because of its complex climate pattern. In different regions. The North Atlantic have different climate models. In the tropics, low-frequency climate variability is closed to Atlantic sea surface temperature (SST) fluctuations; In mid-latitudes, the North Atlantic Oscillation (NAO) is the leading mode of variability, and its effect is far-reaching and significant [46].

Furthermore，climate change will feed back into biogeochemical processes by affecting chemical and biological processes on the surface of the North Atlantic that are critical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

to the absorption of carbon from the ocean. The North Atlantic is considered to be the most important sink of carbon dioxide in the world's oceans, storing 23 percent of the world's anthropogenic carbon, even though it covers only 15 percent of the world's oceans [47]. For the carbon change in the North Atlantic, there are some causes possibly include temperature rise [48], lower water ventilation rate [46], changes in biological activity [49], and anthropogenic carbon dioxide uptake (emissions from fossil fuels) [50]. The uptake of carbon dioxide in the North Atlantic are different in space and time, but few observations cover large areas [51] and long periods of time [52]. Therefore, the study of sea surface pCO_2 in the North Atlantic is of great significance to the study of carbon sink and carbon source in the North Atlantic.

2.2. Data sources

The cruise data used in the study come from the Atlantic oceanographic & meteorological laboratory of the national oceanic and atmospheric administration (NOAA/AOML), which belongs to the NOAA Ocean Acidification Project, has 20 cruise data in the North Atlantic from March 2010 to September 2013. The space range covers most regions, which are in the range of 15°N to 55°N and 80°W to 0°W. Figure 1 shows the cruise data of sea surface pCO_2.

The satellite data used in the study includes Chla, Kd, SST, Adg and MLD. The Chla, Kd, SST and Adg data derived from National Aeronautics and Space Administration (NASA) standard 8-days Modis/Aqua Level-3 data products and they are calculated from spectral remote sensing reflectance (Rrs, sr^{-1}) in 7 bands between 412 and 878 nm. These data come from the NASA Ocean Color Processing Center (OCDPS, http://ocean-color.gsfc.nasa.gov/). The MLD products used in this study come from the 8-day average product data of HYCOM model (January 2003-December 2020). The data was derived from the depth of sea-water mixed layer in the mixed coordinate sea-ocean model established in the United States global ocean data homogenization (www.science.oregonstate.edu/ocean/productivity/). The remote sensing data used in this study are equidistant cylindrical projection raster data with spatial resolution of 4km × 4km. In particular, SST was used to capture the thermodynamic effects, MLD was used to monitor the freshwater characteristics of multiple river inputs and thermodynamic effects, and in addition, Chla, Adg and Kd were reflected the effects of biological activities and physical mixing of water masses.

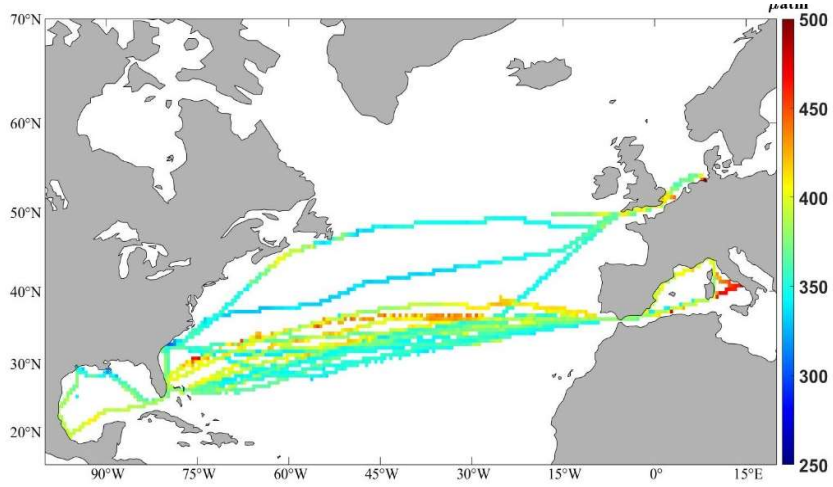

Figure 1. Sea surface pCO_2 from 20 cruises in the North Atlantic from 2010 to 2013.

2.3. Methods

In many published studies, many approaches were used to build the sea surface pCO_2 model based on satellite data. These studies mainly include empirical and semi-analytical approaches. For example, Chen showed the advantage of the semi-analytical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

approaches is that can explain the results and mechanism clearly [10], nevertheless, it is the disadvantage that can't achieve the accuracy of the model that use the empirical method. In recent years, there are lots of studies which used machine learning based on empirical approach, such as Support Vertical Machine (SVM) [9], neural network [18, 37, 53], regression tree [16] and Random Forest (RF) [9]. Besides that, some traditional empirical methods were also used, i.e., multi-linear regression (MLR), multi-nonlinear regression (MNR), principal component regression (PCR) [10, 54].

Among these approaches, we found RF had a highest precision in the most of the time [9] and it is an empirical method which is based on regression tree and an ensemble learning named bagging. In other words, RF takes the advantage of each regression tree via bagging to improve model generalization [55, 56]. Besides that, there are more studies use an empirical method called Gradient Boost Decision Tree (GBDT) based on regression tree but combined another ensemble learning named boosting [57, 58]. The distinction between RF and GBDT, or, the difference between bagging and boosting is the way of sampling. When we want to generate a model which uses ensemble learning, we need sample from the trained data. Bagging is the way that uses uniform sampling but boosting prefer to sample according to higher error rate based on last training (negative down sampling). Beyond that, bagging treats every dataset equally but boosting is based on the weight of different basic classifier. According these factors, RF is easy to meet the problems about over-fitting, when the dataset has a large deviation, which has a good result in training data except for testing data. Furthermore, RF has randomness to an extent, that makes an obscure explanation for the model in the study. In that case, we choose an algorithm named CatBoost, which is an optimization for GBDT method [42], and it has many advantages for evaluating the value of pCO_2.

For the CatBoost, there are two improvements compared to GDBT [42], that is, feature combination and ordered boosting. Firstly, CatBoost model will do some related work on categorical feature, on account of their different cardinality, including calculating the frequency of a category, considering using different combinations of categorical features to build regression trees. Secondly, in order to solve the problem of prediction shift caused by gradient deviation, CatBoost model replaces the gradient estimation method in some traditional algorithms by using ordered boosting. Through the above improvements, CatBoost can achieve progress in regression learning.

Furthermore, we choose three commonly used statistical indicators as the standard and measure to evaluate and compare the models' accuracy and performance, including coefficient of determination (R^2), root mean square error (RMSE), mean bias error (MAE), the statistical indicators are described below [59]:

$$\mathrm{R}^2 = \frac{\sum_{n=1}^{i}(y_{i,m} - y_{i,e})^2}{\sum_{n=1}^{i}(y_{i,m} - \bar{y}_{i,m})^2} \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{n=1}^{i}(y_{i,m} - y_{i,e})^2} \quad (2)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_{i,m} - y_{i,e}| \quad (3)$$

where $y_{i,m}$, $y_{i,e}$ and $\bar{y}_{i,m}$ are the measured data, estimated data, and mean of measured data, respectively and n is the number of data set. We prefer to get a higher R^2. For the others, if the indicators are lower, the model performs better.

In order to decrease the risk of over-fitting, 10-fold cross-validation was chosen in the process of the training. It was this way which means dividing the dataset randomly into 10 equal-size subsamples and putting different 9 subsamples for training and 1 subsample for testing each time that can ensure every data was chosen to train and the risk of over-fitting would be significantly reduced.

3. Results

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

3.1. CatBoost Model performance

This study applies the CatBoost algorithm to all measured data. First, each cruise is matched with the remote sensing product data including SST, Chla, Adg, Kd and MLD at the same time and location as the cruise change. After deleting some extreme data, a total of 51,270 pieces of data are obtained and divided into training sets and test sets. The inputs of the model are SST, Chla, Adg, Kd and MLD and the output is sea surface pCO_2. In order to make the model results achieve the best results, the maximum depth of the model is set to 1-15 layers, and the number of iterations are set to 100, 200, 500, 800, 1000 and 1500 respectively. The test results of the model data validation set are shown in the Figure 2. As the maximum depth increases and the number of iterations increases, the RMSE of the validation set are decreasing and the model effect is better. When the maximum depth is greater than 10 and the number of iterations is greater than 500, the RMSE of the verification set tends to decrease slowly and stabilizes below 10μatm.

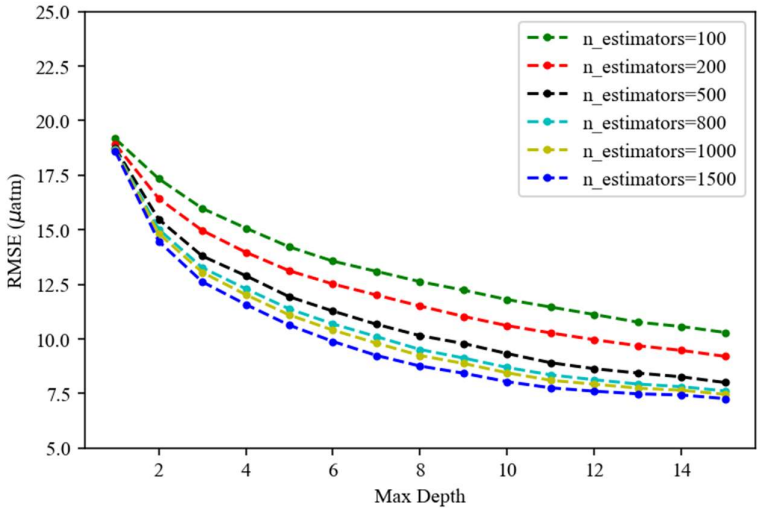

Figure 1. The performance of the CatBoost model in model validation in different max-depths and different number of estimators.

Figure 3 and Figure 4 show the performance of the CatBoost model in training and validation datasets respectively and color coded by data density, when the number of iterators is 1500 and maximum depth of the model is set to 15. Clearly, most data are closely along the 1:1 line, that means we can get a better result with the CatBoost model that makes predicted data are close to the field data. Statistically, the performance of the CatBoost model with the training dataset shows the result that R^2=0.991, RMSE=3.28μatm and MAE=2.36μatm. Similarly, the validation dataset has reached a good result that R^2=0.943, RMSE=8.25μatm and MAE=4.92μatm, which show a good performance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

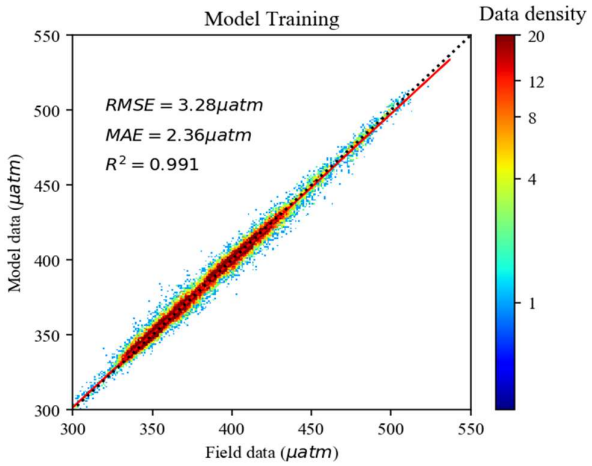

Figure 3. CatBoost model performance in estimating surface pCO_2 in the North Atlantic with training set.

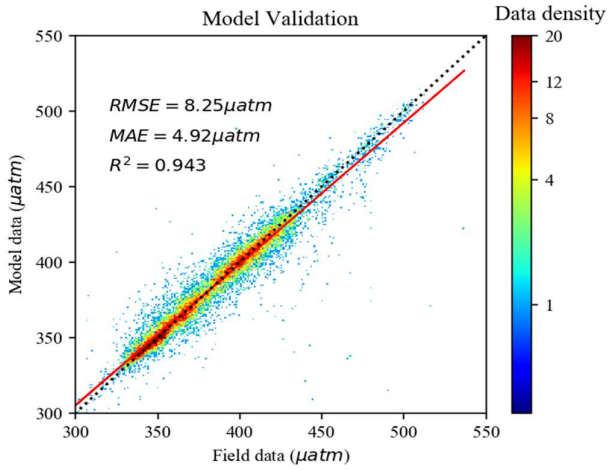

Figure 4. CatBoost model performance in estimating surface pCO_2 in the North Atlantic with validation set.

Besides that, in order to compare the differences and advantages between CatBoost and other model, this study will use some regression algorithms in machine learning, including linear regression, support vector machine, neural network, k-nearest neighbor and some ensemble learning like Random Forest, Gradient Boosting Regression Tree, Adaboost and XGBoost, to build the corresponding sea surface pCO_2 inversion models with same batch of cruise data. The performance of the inversion model with validation dataset based on different algorithms are shown in Table 1, and the results including RMSE, R^2 and MAE.

In Table 1, it can be found that traditional algorithms such as linear regression is poor in the study of the sea surface pCO_2 in the North Atlantic, where R^2 is only 0.31, MAE is 21.95μatm and RMSE is 28.35μatm. Besides that, the performance of using a single weak learner is much lower than that of the ensemble learning method combining multiple learners. Among them, support vector machine, neural network and k-nearest neighbor in regression learning is slightly lower than the method of tree learner, on account of their problems of overfitting. In contrast, the R^2 of regression tree model can reach 0.86, and RMSE less than 14μatm, which has relatively good inversion results compared to other methods. In addition, the ensemble learning method can significantly improve the accuracy of the model. For example, in the model of Random Forest, Bagging Regression and XGBoost, R^2 can reach 0.86 and RMSE less than 10μatm, the performance is slightly lower

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

than CatBoost. Among these algorithms, the performance of the CatBoost model is superior to other algorithms (RMSE=8.25μatm, R^2=0.94, MAE=4.92μatm), so use the CatBoost model can reflect sea surface pCO_2 in the North Atlantic better.

Table 1. Comparison of RMSE, R^2 and MAE between predicted pCO_2 and measured pCO_2 by using different machine learning methods in the verification set.

Algorithm	RMSE (μatm)	R^2	MAE (μatm)
Linear Regression	28.35	0.31	21.95
k-Nearest Neighbor	15.46	0.80	10.07
Neural Network	19.28	0.68	13.73
Regression Tree	13.07	0.86	6.03
Support Vector Machine (Gaussian kernel function)	18.35	0.71	12.28
Support Vector Machine (Linear kernel function)	29.10	0.31	21.45
Random Forest	9.75	0.92	5.57
Bagging Regression	9.69	0.92	5.59
Adaboost	19.44	0.68	14.91
Gradient Boosting Regression Tree	16.87	0.76	12.22
XGBoost	9.75	0.92	6.16
Catboost	8.25	0.94	4.92

3.2. Independent validation

Besides that, in order to compare the differences and advantages between CatBoost and other model, this study will use some regression algorithms in machine learning, including linear regression, support vector machine, neural network, k-nearest neighbor and some ensemble learning like Random Forest, Gradient Boosting Regression Tree, Adaboost and XGBoost, to build the corresponding sea surface pCO_2 inversion models with same batch of cruise data. The performance of the inversion model with validation dataset based on different algorithms are shown in Table 1, and the results including RMSE, R^2 and MAE.

In order to prevent the problem of overfitting, in addition to the cross-validation in the model development, the measured data of each cruise will be individually verified. The timeseries between measured value and predicted value of part of cruises are shown in Figure 5 and Table 2 shows the magnitude of R^2, RMSE and MAE between the pCO_2 measured data and pCO_2 predicted data for each cruise.

Figure 5 shows that the almost agreement with the field pCO_2 and modeled pCO_2 during the cruise except extreme abnormal data. It can be seen from Table 2 that, except for the R^2 between inversion pCO_2 and the measured pCO_2 in cruise BMBE20120418 is 0.45, the other cruises are all above 0.7, and RMSE is less than 10μatm and MAE is less than 5μatm. Besides that, R^2 of 16 cruises higher than 0.85 among all 20 cruises and MAE of all cruises are lower than 10μatm. It suggests that the CatBoost model has good results for different seasons and areas.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

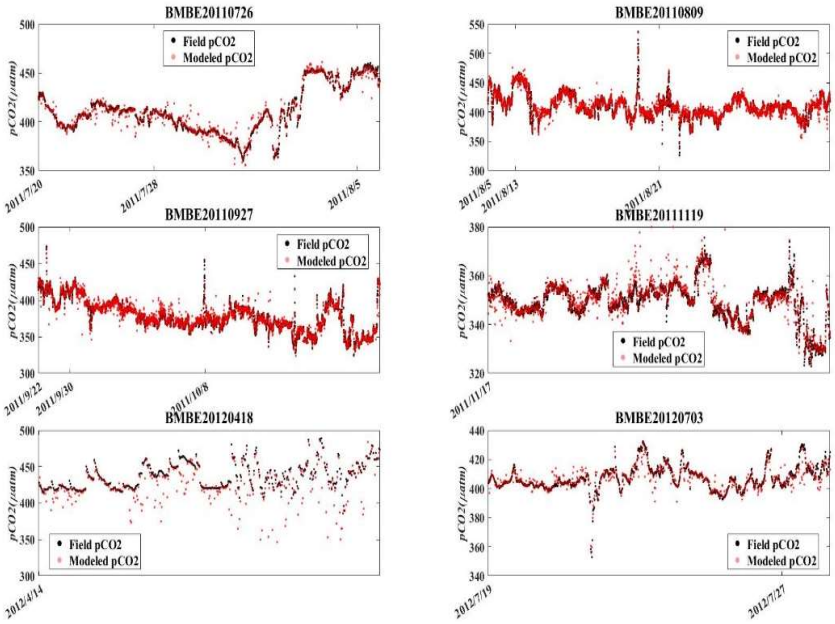

Figure 5. Comparison between field-measured surface pCO_2 and derived pCO_2 in part of cruise.

Table 2. The performance of the CatBoost model between predicted pCO_2 and measured pCO_2 for each cruise.

CRUISE ID	R^2	RMSE (μatm)	MAE (μatm)
BMBE20100302	0.91	5.01	3.11
BMBE20100326	0.89	7.03	3.25
BMBE20101014	0.84	4.96	2.93
BMBE20101202	0.96	4.42	2.88
BMBE20110726	0.94	5.63	3.60
BMBE20110809	0.91	5.72	3.27
BMBE20110927	0.94	5.25	3.47
BMBE20111119	0.80	3.51	2.39
BMBE20120418	0.54	16.46	8.83
BMBE20120703	0.78	5.00	3.22
BMBE20120913	0.88	4.35	2.96
BMBE20121220	0.91	3.69	2.32
BMBE20130207	0.91	5.57	3.55
BMBE20130220	0.90	4.18	2.70
BMBE20130329	0.83	5.78	3.84
BMBE20130515	0.74	4.46	3.01
BMBE20130528	0.85	5.33	3.27
BMBE20130702	0.89	10.81	4.02
BMBE20130716	0.97	5.50	3.57
BMBE20130903	0.87	4.62	3.14

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

Through the independent validation, the result suggest that the CatBoost algorithm has a good generalization ability surface pCO_2 inversion model in these cruises, and even for the extensive data coverage (both spatially and temporally) in the North Atlantic.

3.3. Model sensitivity

Figure 6 shows the sensitivity of the CatBoost model to uncertainties of each input variable (Kd, Chla, Adg, SST and MLD). Similarly, Table 3 shows the differences between original pCO_2 data and new pCO_2 data which some uncertainties noise added. It can be found that the model is more sensitive to uncertainties is SST and MLD than in Chla, Kd and Adg.

Statistically, with -20% uncertainties added in Kd, the model shows a slight change (RMSE=5.62μatm, R^2=0.97), and it has a similar result with +20% uncertainties added in Kd (RMSE=5.32μatm, R^2=0.974). When the uncertainties added in Adg and Chla, the result indicates that the model also has little sensitivity, which R^2 of the cases above 0.97 and their RMSE less than 6μatm. It seems to be acceptable, because the uncertainties in the MODIS satellite-derived products are generally within ±20% in the North Atlantic.

Compared to above variables, SST and MLD has a larger sensitivity in the CatBoost model. When -20% uncertainties added in SST, it shows the result that RMSE is 7.88μatm and R^2 is 0.94. Similarly, it has a significant difference between original data and new data (RMSE=7.33μatm, R^2=0.95) when +20% uncertainties added in SST. When it comes to MLD, the R^2 are both 0.93 and RMSE are less than 9μatm when +20% or -20% uncertainties added. Overall, although MLD and SST are more sensitive to the model compared to others, the sensitivity is acceptable and tolerate to the study in the North Atlantic. Although uncertainties were implicitly included in the developed model when satellite data of each variable were used directly in the model development, and these uncertainties would be cancelled to a large extent when applying the CatBoost model to the same satellite data products.

Table 3. Statistics of the sensitivity of the RFRE pCO2 model to uncertainties in each of the satellite-derived environment variables (Kd, Chla, Adg, SST and MLD) based on the validation dataset.

Cases	RMSE (μatm)	R^2
+20% in Kd	5.32	0.97
-20% in Kd	5.62	0.97
+20% in Chla	5.38	0.97
-20% in Chla	4.45	0.98
+20% in Adg	4.86	0.98
-20% in Adg	4.74	0.98
+20% in SST	7.88	0.94
-20% in SST	7.33	0.95
+20% in MLD	8.66	0.93
-20% in MLD	8.67	0.93

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

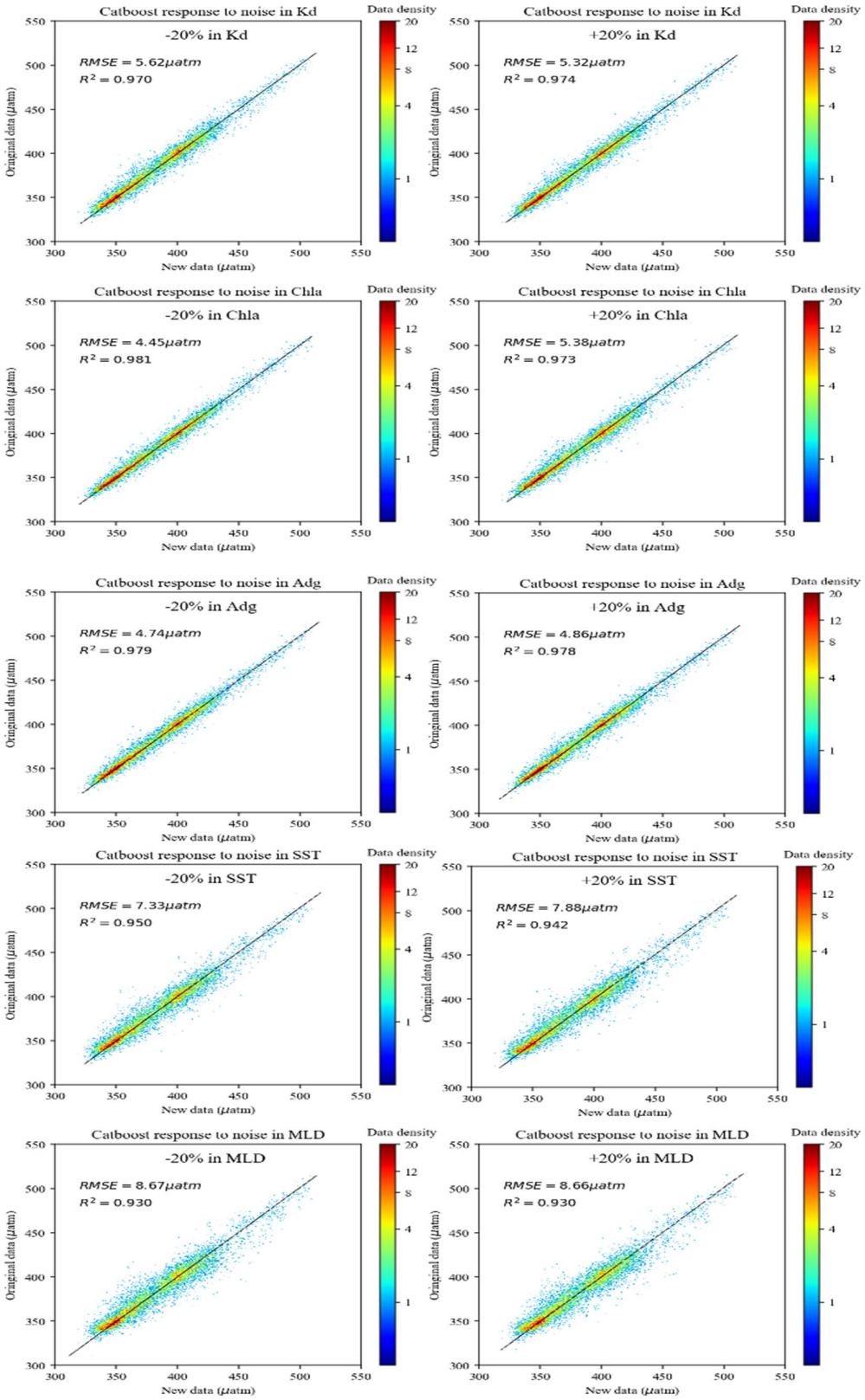

Figure 6. CatBoost pCO_2 model sensitivity to changes in the input SST, Adg, Chla, Kd and MLD, based on the dataset used to develop the pCO_2 model.

4. Discussion

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

4.1. Seasonal and interannual variations of surface pCO_2

In this section, we analyze the spatial and temporal variation patterns of sea surface pCO_2 in the North Atlantic. In addition, we analyze the seasonal and interannual variabilities based on monthly mean surface pCO_2 from January 2003 to December 2020.

Figure 7 shows the annual climatological maps of surface pCO_2 in the North Atlantic based on CatBoost model with MODIS data between January 2003 and December 2020. As shown in the Figure 7, the average sea surface pCO_2 in the low latitude (south of 30°N) is the highest, while the average sea surface pCO_2 in high latitudes (north of 45°N) is slightly higher than in mid-latitudes (between 30°N and 45°N) and pCO_2 in coastal area is higher than that far from the continental shelf. During the period of 2003-2020, the annual average sea surface pCO_2 in North Atlantic did not show a significant trend. In different sub-regions, there are distinct trends and distributions. For example, it has great change for different years in the Canary Sea (15°W-35°W, 20°N-30°N). There was a relatively low situation in the year of 2009, 2014 and 2015. Besides that, in the low latitude area, the annual average sea surface pCO_2 of the US east coast is about 30-50μatm higher than the Canary Sea at the same latitude. In the region of higher latitudes, the sea surface pCO_2 is about 20μatm higher than in the mid-latitudes. Since the annual average distribution of pCO_2 can only roughly judge the inter-annual variation, it cannot reflect the specific seasonal variation trend and the relevant impact of the climate characteristics. The monthly average distribution and variation of sea surface pCO_2 in the North Atlantic is shown in Figure 8.

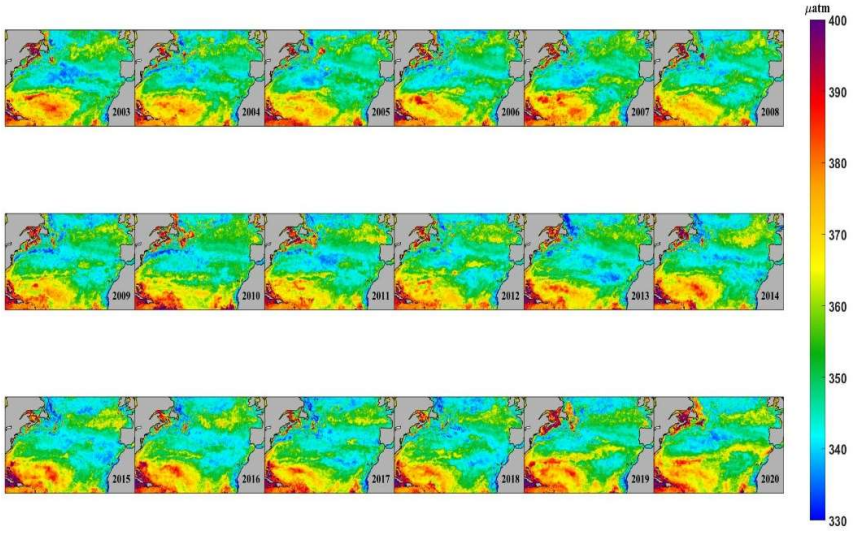

Figure 7. Annual mean distribution of sea surface pCO_2 from 2003 to 2020.

Figure 8 shows the monthly mean distribution of sea surface pCO_2 from 2003 to 2020. It can be seen that the sea surface pCO_2 varies significantly with the seasons in the mid-latitude area (between 30°N and 45°N). The average pCO_2 is lower than 300μatm in winter (from December to February), while the data is higher than 400μatm in summer (from June to August). Secondly, from May to October, the differences of sea surface pCO_2 in the south of 60°N are more obvious, and temperature may play a major role in the process of changing. In addition, the significant changing areas are mainly concentrated in the Gulf stream and the North Atlantic warm current, while the summer pCO_2 in the area with Canary cold current is significantly lower than the same latitude area in the southwestern North Atlantic. Through the influence of cold and warm currents on seawater temperature, the sea surface pCO_2 in the North Atlantic has seasonal changes significantly.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

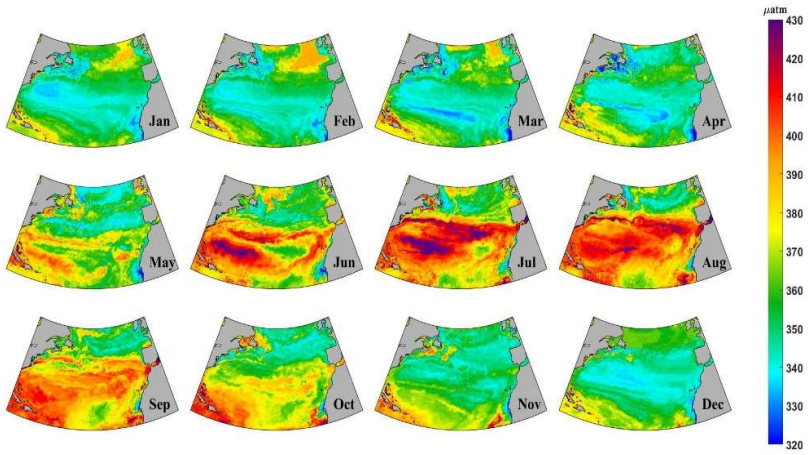

Figure 8. Monthly mean distribution of sea surface pCO_2 from 2003 to 2020.

Figure 9 (a) shows three sub-regions that will be studied separately. They are around the different continental shelves and are affected and controlled by different ocean currents, so they are representative in studying. Sub-region1 (10°W-35°W, 15°N-30°N) is near the African continental shelf and affected by the Canary current. Sub-region2 (60°W-70°W, 20°N-40°N) is near the East American continental shelf and controlled by the Gulf Stream. Sub-region3 (10°W-35°W, 45°N-55°N) is near the European continental shelf and it is the convergence location of the Atlantic current and the Arctic current. Figure 9 (b) - (d) show the time series in the whole North Atlantic and three sub- regions. It is shown that surface pCO_2 in sub-region1 and sub-region2 have same trend with the whole region except for sub-region3. From Figure 9 (b), it can be found surface pCO_2 in sub-region1 is 15-20μatm lower than the whole region in winter and similar in summer. On the contrary, Figure 9 (c) shows surface pCO_2 in sub-region2 is 20μatm higher than the whole region or more in summer but similar in winter. Besides that, Figure 9 (d) shows the trend that higher in winter and lower in summer in the sub-region3 and it suggests that surface pCO_2 in this area may be affected by other factors and processes. Since the North Atlantic has a large range area and the climate patterns are complex, the variations of surface pCO_2 in each subregion is also different.

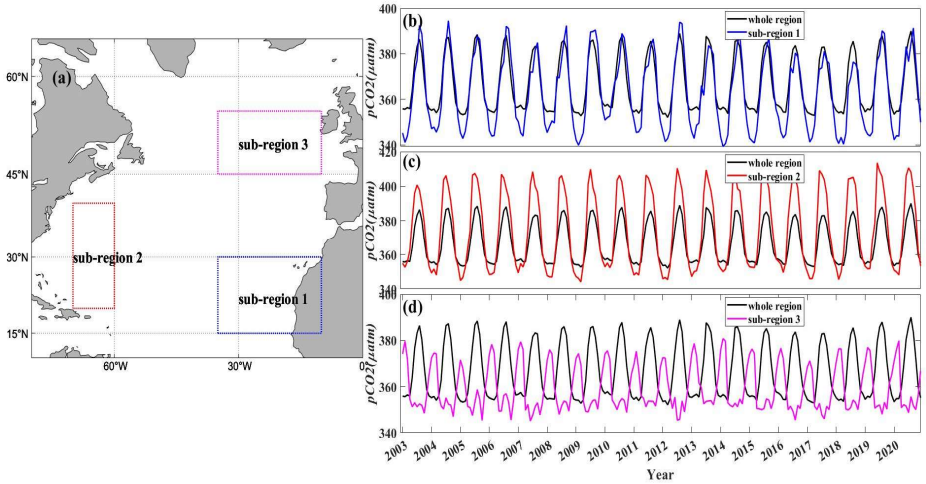

Figure 9. (a) Location of each sub-region. (b) - (d) Monthly surface pCO_2 time series in the whole North Atlantic and in the three sub-regions from Jan 2003 to Dec 2020.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

It can be found that the results of the Empirical Orthogonal Function (EOF) analysis about annual mean surface pCO_2 in the North Atlantic from 2003 to 2010 in Figure 10 and they can show the spatial and temporal variations of each sub-region. First of all, Figure 10 (a) and Figure 10 (b) show the spatial and temporal distribution of the first principal component (EOF1) respectively, and variance contribution rate of EOF1 reaches 0.86. The figures show that there is an opposite trend between southwestern of the North Atlantic and other regions and that is the main spatial distribution in the whole region. In terms of timescale, 2013-2018 has a different trend with other years. Besides that, from Figure 10 (c) and Figure 10 (d), they show that the spatial and temporal distribution of EOF2 and variance contribution rate has 0.10 in this component. Figure 10 (d) suggests that 2019 and 2020 have strong converse trend compared with other years. In addition, different sub-regions have distinct tendencies which sub-region1 and sub-region3 may have diverse variations between 2003 and 2020. Through the EOF analysis, the distribution and variation of each sub-region can be observed separately from the perspective of time and space.

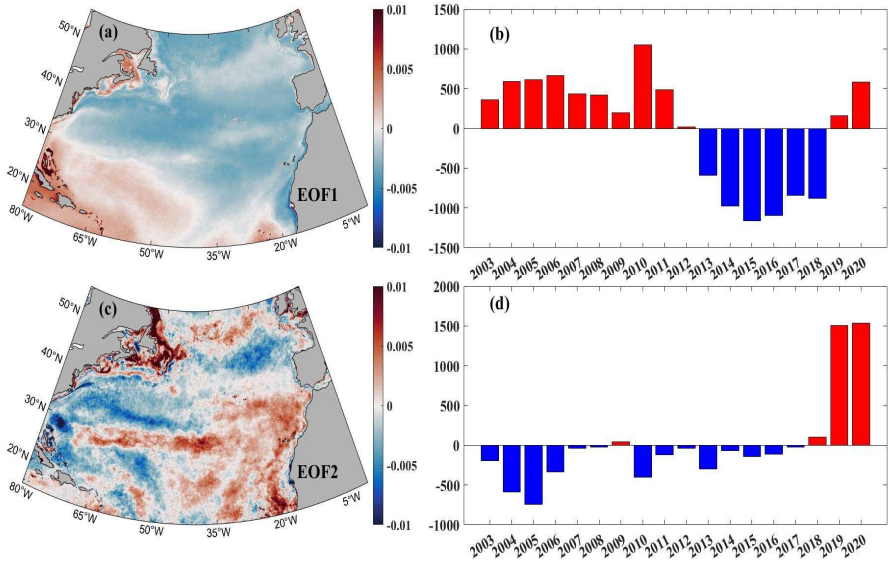

Figure 10. The results of EOF analysis of surface pCO_2 in the North Atlantic from 2003 to 2020. (a) The spatial distribution of the first principal component (b) The temporal distribution of the first principal component (c) The spatial distribution of the second principal component (d) The temporal distribution of the second principal component.

4.2. Comparison between surface pCO_2 and different environmental variables

In this study, we use four environmental variables, including Adg, Chla, Kd, SST and MLD, to develop the CatBoost model in order to invert surface pCO_2 in the North Atlantic. In the published studies, Moussa et al. believed that the influencing factors including SST, SSS and MLD (Moussa, Benallal et al. 2016). Chen et al. found that Kd at 490nm has an effect on surface pCO_2 (Chen, Hu et al. 2016). Figure 11 shows the proportion of each remote sensing variable in the CatBoost model to invert sea surface pCO_2. It can be found the proportion of MLD and SST are above 30%. On the contrary, Adg, Chla and Kd account for only about 10%. Because MLD is mainly determined by the salinity and temperature of seawater, and Chla, Adg and Kd reflect the effects of biological activities, it suggests that salinity and temperature are the most important factor to influence pCO_2, and biology maybe one of the factors affecting sea surface pCO_2.

Figure 12 shows the correlation between annual mean SST, Chla, Kd, MLD and surface pCO_2 from 2003 to 2020 respectively. It can be found that SST and surface pCO_2 have strong positive correlation in most regions. On the contrary, Chla, Kd and MLD have strong negative correlations with surface pCO_2 except high-latitude region. The most obvious area is the sub-region3 in Figure 9 (a) and it may be more affected by biological activities or other factors, so it can explain the opposite trend about sub- region3 and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2021 doi:10.20944/preprints202104.0065.v1

whole region in Figure 9 (d). Except for that, it is shown that SST is still the most important factor affecting surface pCO_2 and biological activities are related to surface pCO_2 variations more or less in some regions.

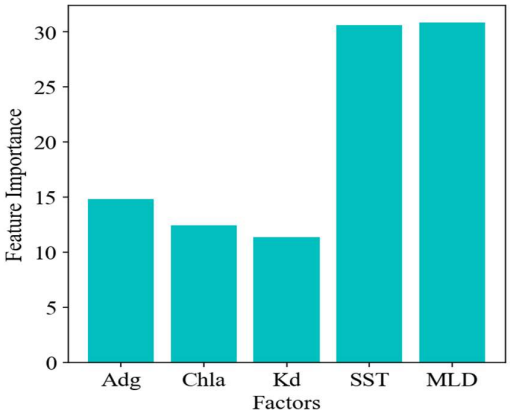

Figure 11. The proportion of each remote sensing variable in the results of the inversion with CatBoost model.

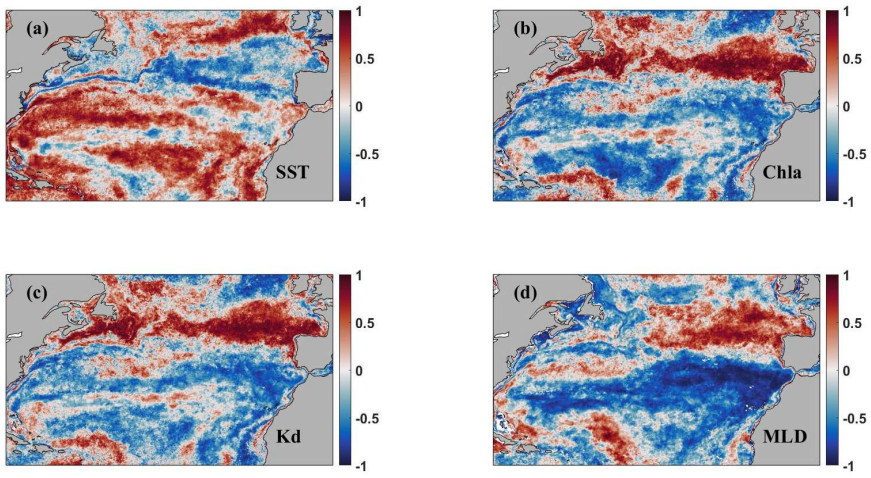

Figure 12. Maps of correlation coefficients between annual mean SST (a), Chla (b), Kd (c), MLD (d), and surface pCO_2 respectively.

4.3. Advantages and limitations of the Catboost

Through the evaluation results, it can be found that empirical CatBoost algorithm can estimate surface pCO_2 in the North Atlantic with the uncertainty of less than 5μatm. Comparing to the published studies, CatBoost model shows great advantages in different environments of North Atlantic. Although the CatBoost model works like a "black box", the mechanisms of estimating surface pCO_2 by each of the input variables are clear. Besides that, the model shows the tolerance to uncertainties in the input satellite variables in different-process-dominated regions of the North Atlantic. Overall, the CatBoost approach shows great advantages over other empirical approaches in satellite mapping of surface pCO_2.

Although the CatBoost model has shown to be applicable in the North Atlantic, but it is unknown whether it works for the water with surface pCO_2 outside the range due to its empirical nature. In our study, the cruise data and satellite data are limited. Furthermore, the model has a satisfactory does not indicate the model is applicable in all types of waters driven by different processes. In addition to the model applicability range, another problem is the CatBoost model works like a "black box", so it cannot understand of the

 doi:10.20944/preprints202104.0065.v1

driving mechanisms between input and output variables explicitly like the semi-analytical approaches and it is difficult to explain clearly about each process influence the surface pCO_2 variations. Besides that, different oceanic processes may not be independent from each other and surface pCO_2 maybe driven by these processes collectively, so empirical approach cannot explain the interaction of different processes but can achieve a better model accuracy.

5. Conclusions

In this study, based on the CatBoost algorithm, an inversion model is established for the sea surface pCO_2 in the North Atlantic. The remote sensing product data such as SST, MLD, Adg and Kd are the input data, and the output is the measured data of the sea surface pCO_2 from 20 cruises. The model shows good performance in both the training set and the validation set, even better in comparisons with other empirical approaches based on different algorithms. Through the model, the monthly and annual average spatial distribution of sea surface pCO_2 in the North Atlantic Ocean from 2003 to 2020 are reversed, and the main influencing factors of surface pCO_2 are analyzed. The following conclusions are drawn:

1. The interannual variation of sea surface pCO_2 in the North Atlantic is relatively stable, and the quarterly variation is more pronounced especially in mid-latitudes. Since various parts of the North Atlantic are affected by different ocean currents and dominated by complex climate patterns, different regions show different trends. In general, the average sea surface $pCOl_2$ in low latitude regions is the highest, while the average sea surface pCO_2 in high latitude regions is slightly higher than that in mid-latitude regions; while at the same latitude, the sea surface pCO_2 in mid-high latitude areas is roughly similarly. But in low latitudes, the pCO_2 in the eastern Atlantic Ocean is obviously lower than that in the western.
2. The main influencing factors of surface pCO_2 in the North Atlantic are SST and SSS. In addition, biological activities also play a role in affecting pCO_2 variations in some regions. The influencing factors are different in each sub-region, on account of complex climate patterns.
3. CatBoost model can invert surface pCO_2 with a wide range of applications and high accuracy results, and future research needs to be focused on improving the capability of the CatBoost pCO_2 model in tracing long-term scale variations and explain the interaction mechanism of each process.

Author Contributions: Conceptualization, S.H. Methodology, S.H., Z.B.; Analysis, S.H., C.Y. and Z.B.; Resources, L.L; Writing and Editing, S.H., L.L.; Supervision, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China 2016YFC1400905.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: Not applicable.

Acknowledgments: The authors are grateful to the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this manuscript and the National Key Research and Development Program of China 2016YFC1400905.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Menon, S., et al., *Climate Change 2007: The Physical Science Basis, Couplings Between Changes in the Climate System and Biogeochemistry*. 2007.
2. Cai, W.-J., *Estuarine and Coastal Ocean Carbon Paradox: CO2 Sinks or Sites of Terrestrial Carbon Incineration?*, in *Annual Review of Marine Science, Vol 3*, C.A. Carlson and S.J. Giovannoni, Editors. 2011. p. 123-+.
3. Wei-Jun and C.J.G.R. Letters, *Air-sea exchange of carbon dioxide in ocean margins: A province-based synthesis*. 2006.
4. Chen, F., et al., *Sea surface pCO(2)-SST relationships across a cold-core cyclonic eddy: Implications for understanding regional variability and air-sea gas exchange.* Geophysical Research Letters, 2007. **34**(10).
5. Sun, Q., D. Tang, and S. Wang, *Remote-sensing observations relevant to ocean acidification.* International Journal of Remote Sensing, 2012. **33**(23): p. 7542-7558.
6. Doney, S.C., et al., *OCEAN ACIDIFICATION: A CRITICAL EMERGING PROBLEM FOR THE OCEAN SCIENCES.* Oceanography, 2009. **22**(4): p. 16-+.
7. Orr, J.C., et al., *Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms.* Nature, 2005. **437**(7059): p. 681-686.
8. Bai, Y., et al., *A mechanistic semi-analytical method for remotely sensing sea surface pCO2 in river-dominated coastal oceans: A case study from the East China Sea.* 2015. **120**(3): p. 2331-2349.
9. Chen, S., et al., *A machine learning approach to estimate surface ocean pCO(2) from satellite measurements.* Remote Sensing of Environment, 2019. **228**: p. 203-226.
10. Chen, S., et al., *Estimating surface pCO(2) in the northern Gulf of Mexico: Which remote sensing model to use?* Continental Shelf Research, 2017. **151**: p. 94-110.
11. Le, C., et al., *Estimating summer sea surface pCO(2) on a river-dominated continental shelf using a satellite-based semi-mechanistic model.* Remote Sensing of Environment, 2019. **225**: p. 115-126.
12. Fennel, K., et al., *Denitrification effects on air-sea CO2 flux in the coastal ocean: Simulations for the northwest North Atlantic.* Geophysical Research Letters, 2008. **35**(24).
13. Ikawa, H., et al., *Air-sea exchange of CO2 at a Northern California coastal site along the California Current upwelling system.* Biogeosciences, 2013. **10**(7): p. 4419-4432.
14. Xue, L., et al., *Sea surface carbon dioxide at the Georgia time series site (2006-2007): Air-sea flux and controlling processes.* Progress in Oceanography, 2016. **140**: p. 14-26.
15. Chen, S., et al., *Remote estimation of surface pCO(2) on the West Florida Shelf.* Continental Shelf Research, 2016. **128**: p. 10-25.
16. Lohrenz, S.E., et al., *Satellite estimation of coastal pCO(2) and air-sea flux of carbon dioxide in the northern Gulf of Mexico.* Remote Sensing of Environment, 2018. **207**: p. 71-83.
17. Marrec, P., et al., *Dynamics of air-sea CO2 fluxes in the northwestern European shelf based on voluntary observing ship and satellite observations.* Biogeosciences, 2015. **12**(18): p. 5371-5391.
18. Moussa, H., et al., *Satellite-derived CO2 fugacity in surface seawater of the tropical Atlantic Ocean using a feedforward neural network.* International Journal of Remote Sensing, 2016. **37**(3): p. 580-598.
19. Fay, A.R. and G.A. McKinley, *Correlations of surface ocean pCO(2) to satellite chlorophyll on monthly to interannual timescales.* Global Biogeochemical Cycles, 2017. **31**(3): p. 436-455.
20. Zhu, Y., et al., *Satellite-derived surface water pCO(2) and air-sea CO2 fluxes in the northern South China Sea in summer.* Progress in Natural Science-Materials International, 2009. **19**(6): p. 775-779.
21. Lee, K., et al., *Global relationships of total alkalinity with salinity and temperature in surface waters of the world's oceans.* Geophysical Research Letters, 2006. **33**(19).
22. Yang, B., R.H. Byrne, and R. Wanninkhof, *Subannual variability of total alkalinity distributions in the northeastern Gulf of Mexico.* Journal of Geophysical Research-Oceans, 2015. **120**(5): p. 3805-3816.
23. Pierrot, D., et al., *MS Excel Program Developed For CO2 System Calculations.* 2006.

24. Reynaud, S., et al., *Interacting effects of CO2 partial pressure and temperature on photosynthesis and calcification in a scleractinian coral.* Global Change Biology, 2003. **9**(11): p. 1660-1668.
25. Bates, N.R., et al., *Variability of pCO(2) on diel to seasonal timescales in the Sargasso Sea near Bermuda.* Journal of Geophysical Research-Oceans, 1998. **103**(C8): p. 15567-15585.
26. Turk, D., J.W. Book, and W.R.J.J.o.M.S. Mcgillis, *PCO2 and CO2 exchange during high bora winds in the Northern Adriatic.* 2013. **s 117–118**: p. 65–71.
27. Bates, N.R. and L. Merlivat, *The influence of short-term wind variability on air-sea CO2 exchange.* Geophysical Research Letters, 2001. **28**(17): p. 3281-3284.
28. Sarma, V.V.S.S., et al., *Basin-scale pCO(2) distribution using satellite sea surface temperature, Chla, and climatological salinity in the North Pacific in spring and summer.* Global Biogeochemical Cycles, 2006. **20**(3).
29. Stephens, M.P., et al., *SEA-AIR FLUX OF CO2 IN THE NORTH PACIFIC USING SHIPBOARD AND SATELLITE DATA.* Journal of Geophysical Research-Oceans, 1995. **100**(C7): p. 13571-13583.
30. Jamet, C., C. Moulin, and N. Lefevre, *Estimation of the oceanic pCO(2) in the North Atlantic from VOS lines in-situ measurements: parameters needed to generate seasonally mean maps.* Annales Geophysicae, 2007. **25**(11): p. 2247-2257.
31. Olsen, A., J.A. Trinanes, and R. Wanninkhof, *Sea-air flux of CO2 in the Caribbean Sea estimated using in situ and remote sensing data.* Remote Sensing of Environment, 2004. **89**(3): p. 309-325.
32. Ono, T., et al., *Basin-scale extrapolation of shipboard pCO(2) data by using satellite SST and Ch1a.* International Journal of Remote Sensing, 2004. **25**(19): p. 3803-3815.
33. Rangama, Y., et al., *Variability of the net air-sea CO2 flux inferred from shipboard and satellite measurements in the Southern Ocean south of Tasmania and New Zealand.* Journal of Geophysical Research-Oceans, 2005. **110**(C9).
34. Chen, L., et al., *Estimation of monthly air-sea CO2 flux in the southern Atlantic and Indian Ocean using in-situ and remotely sensed data.* Remote Sensing of Environment, 2011. **115**(8): p. 1935-1941.
35. Sarma, V., *Monthly variability in surface pCO(2) and net air-sea CO2 flux in the Arabian Sea.* Journal of Geophysical Research-Oceans, 2003. **108**(C8).
36. Memery, L., et al., *The relevant time scales in estimating the air-sea CO2 exchange in a mid-latitude region.* Deep-Sea Research Part Ii-Topical Studies in Oceanography, 2002. **49**(11): p. 2067-2092.
37. Friedrich, T. and A. Oschlies, *Neural network-based estimates of North Atlantic surface pCO(2) from satellite data: A methodological study*. Journal of Geophysical Research-Oceans, 2009. **114**.
38. Landschuetzer, P., et al., *A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink.* Biogeosciences, 2013. **10**(11): p. 7793-7815.
39. Nakaoka, S., et al., *Estimating temporal and spatial variation of ocean surface pCO(2) in the North Pacific using a self-organizing map neural network technique.* Biogeosciences, 2013. **10**(9): p. 6093-6106.
40. Telszewski, M., et al., *Estimating the monthly pCO(2) distribution in the North Atlantic using a self-organizing neural network.* Biogeosciences, 2009. **6**(8): p. 1405-1421.
41. Hales, B., et al., *Satellite-based prediction of pCO(2) in coastal waters of the eastern North Pacific.* Progress in Oceanography, 2012. **103**: p. 1-15.
42. Prokhorenkova, L., et al., *CatBoost: unbiased boosting with categorical features*, in *Advances in Neural Information Processing Systems 31*, S. Bengio, et al., Editors. 2018.
43. Huang, G., et al., *Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions.* Journal of Hydrology, 2019. **574**: p. 1029-1041.
44. Zhang, Y., Z. Zhao, and J. Zheng, *CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China.* Journal of Hydrology, 2020. **588**.

45. Li, S., et al., *Quantification of chlorophyll-a in typical lakes across China using Sentinel-2 MSI imagery with machine learning algorithm.* The Science of the total environment, 2021. **778**: p. 146271-146271.
46. Marshall, J., et al., *North Atlantic climate variability: Phenomena, impacts and mechanisms.* International Journal of Climatology, 2001. **21**(15): p. 1863-1898.
47. Petit, J.R., et al., *Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica.* Nature, 1999. **399**(6735): p. 429-436.
48. Bates, N.R., *Interannual variability of the oceanic CO2 sink in the subtropical gyre of the North Atlantic Ocean over the last 2 decades.* Journal of Geophysical Research-Oceans, 2007. **112**(C9).
49. Olsen, A., et al., *Interannual variability in the wintertime air-sea flux of carbon dioxide in the northern North Atlantic, 1981-2001.* Deep-Sea Research Part I-Oceanographic Research Papers, 2003. **50**(10-11): p. 1323-1338.
50. Luger, H., et al., *CO2 fluxes in the subtropical and subarctic North Atlantic based on measurements from a volunteer observing ship.* Journal of Geophysical Research-Oceans, 2006. **111**(C6).
51. Corbiere, A., et al., *Interannual and decadal variability of the oceanic carbon sink in the North Atlantic subpolar gyre.* Tellus Series B-Chemical and Physical Meteorology, 2007. **59**(2): p. 168-178.
52. Canadell, J.G., et al., *Contributions to accelerating atmospheric CO2 growth from economic activity, carbon intensity, and efficiency of natural sinks.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(47): p. 18866-18870.
53. Jo, Y.-H., et al., *On the variations of sea surface pCO(2) in the northern South China Sea: A remote sensing based neural network approach.* Journal of Geophysical Research-Oceans, 2012. **117**.
54. Lohrenz, S.E. and W.J. Cai, *Satellite ocean color assessment of air-sea fluxes of CO2 in a river-dominated coastal margin.* Geophysical Research Letters, 2006. **33**(1).
55. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
56. James, G., et al., *Tree-Based Methods*. 2013: An Introduction to Statistical Learning.
57. Friedman, J.H., *Greedy function approximation: A gradient boosting machine.* Annals of Statistics, 2001. **29**(5): p. 1189-1232.
58. Chen, T., H. Tong, and M. Benesty, *xgboost: Extreme Gradient Boosting*. 2016.
59. Barnes, B.B. and C. Hu, *Cross-Sensor Continuity of Satellite-Derived Water Clarity in the Gulf of Mexico: Insights Into Temporal Aliasing and Implications for Long-Term Water Clarity Assessment.* Ieee Transactions on Geoscience and Remote Sensing, 2015. **53**(4): p. 1761-1772.