Article

Space-Time Sea Surface pCO$_2$ Estimation in the North Atlantic Based on CatBoost

Hongwei Sun $^{1,*}$, Junyu He $^{1,2,3}$, Yihui Chen $^1$ and Boyu Zhao $^3$

$^1$ Ocean College, Zhejiang University, Zhoushan 316021, China; jyhe@zju.edu.cn (J.H.); chenyihui@zju.edu.cn (Y.C.)
$^2$ Ocean Academy, Zhejiang University, Zhoushan 316021, China
$^3$ Economics College, Zhejiang University, 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 used chlorophyll-a concentration (Chla), sea surface temperature (SST), dissolved and particulate detrital matter absorption coefficient (Adg), the diffuse attenuation coefficient of downwelling irradiance at 490 nm (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 showed that the root mean square error (RMSE) is 8.25 µatm, the mean bias error (MAE) is 4.92 µatm and the coefficient of determination ($R^2$) can reach 0.946 in the validation set. Subsequently, the proposed algorithm was applied to the sea surface pCO$_2$ in the North Atlantic Ocean during 2003–2020. 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, through variance analysis and EOF (empirical orthogonal function) analysis, the sea surface pCO$_2$ in this area is mainly affected by sea temperature and salinity, while it can also be influenced by biological activities in some sub-regions.

Keywords: sea surface pCO$_2$; ocean color remote sensing; CatBoost algorithm; temporal and spatial distribution

1. Introduction

The ocean is one of the destinations for anthropogenic carbon, which takes up about 30% of emissions from pre-industrial times to 1994 [1]. The CO$_2$ cycle 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 since the industrial revolution [5–7]. In recent years, regional and global air–sea CO$_2$ flux has received a lot of attention [8–11], therefore more and more measured data about sea surface partial pressure of carbon dioxide (pCO$_2$).

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, as follows: (I) sea surface temperature (SST, °C) can reflect the thermodynamic process directly. (II) Some inherent optical properties (IOPs) and apparent optical properties (AOPs), such as dissolved and particulate detrital matter absorption coefficient (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. (III) Some elements such as sea surface salinity (SSS, dimensionless) and surface chlorophyll-a concentration (Chla, mg·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]. (IV) The process of the horizon and vertical mixing among the water masses, which have different characteristics such as total alkalinity (TA, $\mu$mol·kg$^{-1}$) and dissolved inorganic carbon (DIC, $\mu$mol·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]. (V) Extreme events like hurricanes and storms also affect surface $pCO_2$ through the process of air–sea CO$_2$ exchange [24–26].

Given the above considerations, many regression models (including linear regression, multiple regression and nonlinear regression) have been made to estimate sea surface $pCO_2$ by using environmental factors, such as SST, Chla and MLD [27–34], e.g., T. Ono used second-order multiple regression equations of SST and Chla with a regression error of $\pm 14 \mu$atm and $\pm 17 \mu$atm in the subtropical and subarctic domain, respectively [35]. However, these regression models are not sufficient to model a complex ocean environment, because many natural processes interact and a direct relationship between multiple factors does not exist [35]. In recent years, machine learning techniques (such as multi-layer perceptron neural network, self-organizing mapping, supporting vector machines, principal component regression and random forests) have been widely used to study the complex environmental systems due to their characteristics about self-learning and self-adapting [9,18,36–40]. In these studies, the model produced to predicting $pCO_2$ can have relative precise results in specific regions, e.g., Lohrenz showed the result of $R^2$ reaches 0.743 when predicting $pCO_2$ in the northern Gulf of Mexico when using principal component analysis and multiple regression [40]; Moussa used a feedback neural network to predict $pCO_2$ in the tropical Atlantic Ocean with a good result (RMSE = 8.7 $\mu$atm) [18].

Although these methods had made progress, it is still a challenge in estimating surface $pCO_2$ because of the complexity and dynamics in the physical process in open ocean water. Some problems need to be solved in the region which is 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 et al. [8] proposed a mechanistic semi-analytic algorithm that makes the model more applied in other regions and improves the accuracy across short timescales and small spatial scales, 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.

To overcome the issues mentioned above, this study employed a new empirical method to develop a machine learning model for obtaining high precise remote $pCO_2$ product in North Atlantic during 2003–2020 by using Chla, Kd, SST, Adg and MLD, and we select the CatBoost method 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 to other regions all over the world. Further, we will analyze the distribution and variation of the surface $pCO_2$ in the North Atlantic.

### 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 has 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 feedback into biogeochemical processes by affecting chemical and biological processes on the surface of the North Atlantic that are critical 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 is 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 and meteorological laboratory of the national oceanic and atmospheric administration (NOAA/AOML), which belongs to the NOAA Ocean Acidification Project. It included 20 cruise data in the North Atlantic from March 2010 to September 2013. This cruise data covers an area with the longitude ranging from $15^\circ$N to $55^\circ$N and latitude ranging from $80^\circ$W to $0^\circ$W, shown in Figure 1.

![Figure 1. Sea surface pCO$_2$ from 20 cruises in the North Atlantic from 2010 to 2013.](image)

In particular, Chla, Adg and Kd can reflect the effects of biological activities and physical mixing of water masses; while SST and MLD are able to capture the thermodynamic effects and monitor the freshwater characteristics of multiple river inputs and thermodynamic effects, respectively. Therefore, they are selected for further model building purposes in this study. The sources of these data were listed in Table 1.

Table 1. Data source about remote sensing product in the model.

<table>
<thead>
<tr>
<th>Products</th>
<th>Time Scale</th>
<th>Resources</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chla</td>
<td>8 days</td>
<td>NASA Modis/Aqua Level-3 data</td>
<td>January 2003–December 2020</td>
</tr>
<tr>
<td>SST</td>
<td>8 days</td>
<td>NASA Modis/Aqua Level-3 data</td>
<td>January 2003–December 2020</td>
</tr>
<tr>
<td>Adg</td>
<td>8 days</td>
<td>NASA Modis/Aqua Level-3 data</td>
<td>January 2003–December 2020</td>
</tr>
<tr>
<td>Kd</td>
<td>8 days</td>
<td>NASA Modis/Aqua Level-3 data</td>
<td>January 2003–December 2020</td>
</tr>
<tr>
<td>MLD</td>
<td>8 days</td>
<td>HYCOM model</td>
<td>January 2003–December 2020</td>
</tr>
</tbody>
</table>
2.3. Methods

In recent years, there are lots of studies that used machine learning based on an empirical approach, such as support vertical machine (SVM) [9], neural network [18,36,53], 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) [10,40]. Among these approaches, we found RF had the highest precision most of the time [9] and it is an empirical method that is based on a 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 [54,55]. Besides that, there are more studies that use an empirical method called gradient boost decision tree (GBDT) based on regression tree but combined another ensemble learning named boosting [56,57]. 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 a sample from the trained data. Bagging is the way that uses uniform sampling but boosting prefers to sample according to a higher error rate based on last training (negative downsampling). Beyond that, bagging treats every dataset equally but boosting is based on the weight of different basic classifiers. According to 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 Categorical Boosting (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 GBDT [42], that is, feature combination and ordered boosting. Firstly, the CatBoost model will do some related work on categorical features, 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, the 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 [58]:

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

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{i,m} - y_{i,e})^2} \]  

\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_{i,m} - y_{i,e}| \]  

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. 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 pCO2. 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.

Since the satellite input variables have inherent uncertainties and deviations, the study fed the uncertainties of each MODIS-derived variable into the CatBoost model, in order to determine the model’s sensitivity through the comparison between sea surface pCO2 using original inputs and inputs with extra uncertainties added, separately. Specifically, MODIS-derived SST has an uncertainty of ≤1 °C [59], and Chla has an uncertainty of 5–35% [60–62]. Kd has an uncertainty of 13% [63]. Therefore, in our study, ±20% uncertainties were added to the input variables to explore the impact of uncertainties on our retrieval.

Further, the proposed algorithm was applied to retrieve the sea surface pCO2 in the North Atlantic during 2003–2020 and the space-time variations and patterns of sea surface pCO2 were investigated.

3. Results
3.1. CatBoost Model Performance

The test results of the model data validation set are shown in Figure 2. As the maximum depth increases and the number of iterations increases, the RMSE of the validation set is 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 10 µatm.

Figure 2. The performance of the CatBoost model in model validation in different max-depths and a different number of estimators.

Figure 3 shows 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 the maximum depth of the model is set to 15. In the two figures, we can find that the CatBoost model-estimated pCO2 is very close to the in-situ observed pCO2 within both training and validating data set, since the scatter points are distributed along the 1:1 line. Specifically, the accuracy indicators of the training and validating data set are 0.991 vs. 0.943, 3.28 vs. 4.92 µatm and 2.36 vs. 4.92 µatm in terms of R2, RMSE and MAE, respectively.
Besides that, in order to compare the differences and advantages between CatBoost and other models, 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 pCO2 inversion models with the same batch of cruise data. The performance of the inversion model with validation dataset based on different algorithms are shown in Table 2, and the results including RMSE, $R^2$ and MAE.

Table 2. The performance of different machine learning methods on pCO2 estimation using the validation set.

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

In Table 2, it can be found that traditional algorithms such as linear regression are poor in the study of the sea surface pCO2 in the North Atlantic, where $R^2$, MAE and RMSE are 0.31, 21.95 µatm and 28.35 µatm, respectively. 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, because of the overfitting issue. In contrast, the $R^2$ of the regression tree model can reach 0.86, and RMSE less than 14 µatm, which has relatively good inversion performance 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 than CatBoost. Among these algorithms, the performance of the CatBoost model is superior to other algorithms (RMSE = 8.25 µatm, $R^2 = 0.94$, MAE = 4.92 µatm).
3.2. Independent Validation

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 independent validation results are shown in Figure 4 and Table 3.

![Figure 4. Comparison between field-measured surface pCO$_2$ and derived pCO$_2$ in part of cruise.](image)

Table 3. The performance of the CatBoost model between predicted pCO$_2$ and measured pCO$_2$ for each cruise.

<table>
<thead>
<tr>
<th>CRUISE ID</th>
<th>$R^2$</th>
<th>RMSE (µatm)</th>
<th>MAE (µatm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMBE20100302</td>
<td>0.91</td>
<td>5.01</td>
<td>3.11</td>
</tr>
<tr>
<td>BMBE20100326</td>
<td>0.89</td>
<td>7.03</td>
<td>3.25</td>
</tr>
<tr>
<td>BMBE20101014</td>
<td>0.84</td>
<td>4.96</td>
<td>2.93</td>
</tr>
<tr>
<td>BMBE20101202</td>
<td>0.96</td>
<td>4.42</td>
<td>2.88</td>
</tr>
<tr>
<td>BMBE20110726</td>
<td>0.94</td>
<td>5.63</td>
<td>3.60</td>
</tr>
<tr>
<td>BMBE20110809</td>
<td>0.91</td>
<td>5.72</td>
<td>3.27</td>
</tr>
<tr>
<td>BMBE20110927</td>
<td>0.94</td>
<td>5.25</td>
<td>3.47</td>
</tr>
<tr>
<td>BMBE20111119</td>
<td>0.80</td>
<td>3.51</td>
<td>2.39</td>
</tr>
<tr>
<td>BMBE20120418</td>
<td>0.54</td>
<td>16.46</td>
<td>8.83</td>
</tr>
<tr>
<td>BMBE20120703</td>
<td>0.78</td>
<td>5.00</td>
<td>3.22</td>
</tr>
<tr>
<td>BMBE20120913</td>
<td>0.88</td>
<td>4.35</td>
<td>2.96</td>
</tr>
<tr>
<td>BMBE20121220</td>
<td>0.91</td>
<td>3.69</td>
<td>2.32</td>
</tr>
<tr>
<td>BMBE20130207</td>
<td>0.91</td>
<td>5.57</td>
<td>3.55</td>
</tr>
<tr>
<td>BMBE20130220</td>
<td>0.90</td>
<td>4.18</td>
<td>2.70</td>
</tr>
<tr>
<td>BMBE20130329</td>
<td>0.83</td>
<td>5.78</td>
<td>3.84</td>
</tr>
<tr>
<td>BMBE20130515</td>
<td>0.74</td>
<td>4.46</td>
<td>3.01</td>
</tr>
<tr>
<td>BMBE20130528</td>
<td>0.83</td>
<td>5.33</td>
<td>3.27</td>
</tr>
<tr>
<td>BMBE20130702</td>
<td>0.89</td>
<td>10.81</td>
<td>4.02</td>
</tr>
<tr>
<td>BMBE20130716</td>
<td>0.97</td>
<td>5.50</td>
<td>3.57</td>
</tr>
<tr>
<td>BMBE20130903</td>
<td>0.87</td>
<td>4.62</td>
<td>3.14</td>
</tr>
</tbody>
</table>
Figure 4 shows that most estimated pCO$_2$ values are in line with the in-situ field pCO$_2$ sampling data except for some extreme abnormal data. Specifically, the $R^2$ ranges from 0.74 to 0.97, except the cruise BMBE20120418 with $R^2$ 0.54; and the RMSE and MAE ranges from 3.51 to 10.81 µatm, and from 2.32 to 4.01 µatm, respectively, except BMBE20120418 (due to its extreme abnormal from its cruise route).

Through the independent validation, the results 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 5 shows the sensitivity of the CatBoost model to uncertainties of each input variable (Kd, Chla, Adg, SST and MLD). Similarly, Table 4 shows the performance of the CatBoost model by adding the noise of each input variable. It can be found that the model is more sensitive to uncertainties in SST and MLD than in Chla, Kd and Adg. Statistically, with −20% uncertainties added in Kd, the model shows a slight change (RMSE = 5.62 µatm, $R^2 = 0.970$), and it has a similar result with +20% uncertainties added in Kd (RMSE = 5.32 µatm, $R^2 = 0.974$). When the uncertainties are 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 6 µatm. It seems to be acceptable because the uncertainties in the MODIS satellite-derived products are generally within ±20% in the North Atlantic.

<table>
<thead>
<tr>
<th>Cases</th>
<th>RMSE (µatm)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>+20% in Kd</td>
<td>5.32</td>
<td>0.97</td>
</tr>
<tr>
<td>−20% in Kd</td>
<td>5.62</td>
<td>0.97</td>
</tr>
<tr>
<td>+20% in Chla</td>
<td>5.38</td>
<td>0.97</td>
</tr>
<tr>
<td>−20% in Chla</td>
<td>4.45</td>
<td>0.98</td>
</tr>
<tr>
<td>+20% in Adg</td>
<td>4.86</td>
<td>0.98</td>
</tr>
<tr>
<td>−20% in Adg</td>
<td>4.74</td>
<td>0.98</td>
</tr>
<tr>
<td>+20% in SST</td>
<td>7.88</td>
<td>0.94</td>
</tr>
<tr>
<td>−20% in SST</td>
<td>7.33</td>
<td>0.95</td>
</tr>
<tr>
<td>+20% in MLD</td>
<td>8.66</td>
<td>0.93</td>
</tr>
<tr>
<td>−20% in MLD</td>
<td>8.67</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Compared to the above variables, SST and MLD have a larger sensitivity in the CatBoost model. When −20% uncertainties are added in SST, it shows that RMSE is 7.88 µatm and $R^2$ is 0.94. Similarly, it has a significant difference between original data and new data (RMSE = 7.33 µatm, $R^2 = 0.95$) when +20% uncertainties are added in SST. When it comes to MLD, the $R^2$ values are both 0.93 and RMSE are less than 9 µatm 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 tolerable 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 canceled to a large extent when applying the CatBoost model to the same satellite data products.
Figure 5. 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.
3.4. 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 6 shows the annual climatological maps of surface pCO$_2$ in the North Atlantic based on the CatBoost model with MODIS data between January 2003 and December 2020. Although no significant temporal trend was found with the annual mean values of sea surface pCO$_2$ in the North Atlantic ocean, some spatial patterns can be concluded as follows: (a) highest annual mean values of sea surface pCO$_2$ was detected in the low latitude areas (south of 30°N); (b) the annual mean values of sea surface pCO$_2$ in high latitude areas (north of 45°N) are slightly higher than the values in the mid-latitude areas (between 30°N and 45°N); (c) the annual mean values of sea surface pCO$_2$. 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 level in the years 2009, 2014 and 2015. Besides that, in the low latitude area, the annual mean sea surface pCO$_2$ value on the east coast of the US is about 30–50 µatm higher than the Canary Sea at the same latitude. In the region of higher latitudes (45°N–60°N), the sea surface pCO$_2$ is about 20 µatm higher than in the mid-latitudes (30°N–45°N).

![Figure 6. Annual mean distribution of sea surface pCO$_2$ from 2003 to 2020.](image)

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 7. 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 300 µatm in winter (from December to February), while the data is higher than 400 µatm 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.
Further, three sub-regions from various continental shelves in the North Atlantic Ocean were chosen for comparison purposes, shown in Figure 8a, as follows: 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 8b–d show the time series of sea surface pCO2 in the whole North Atlantic and three sub-regions. Several conclusions can be summarized: (i) Surface pCO2 in sub-region1 and sub-region2 have the same trend compared with the whole region except for sub-region3. (ii) Surface pCO2 in sub-region1 is 15–20 µatm lower than the whole region in winter and similar in summer (Figure 8b). (iii) Surface pCO2 in sub-region2 is 20 µatm higher than the whole region or more in summer but similar in winter (Figure 8c). (iv) Regarding sub-region3, high sea surface pCO2 values were found in the winter season while low values were found in summer seasons, suggesting that surface pCO2 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 pCO2 in each sub-region are also different.
Empirical orthogonal function (EOF) analysis was employed to detect the spatial and temporal variation of the annual mean sea surface pCO$_2$ in the North Atlantic Ocean during the period 2003–2020. First of all, Figure 9a,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 the opposite trend with other years. Besides that, Figure 9c,d show that the spatial and temporal distribution of EOF2, respectively, and variance contribution rate has 0.10 in this component. Figure 9d suggests that 2019 and 2020 have strong converse trends compared with other years. In addition, different sub-regions have opposite tendencies. For example, 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 9. 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. Discussion
4.1. 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. [18] believed that the impact factors including SST, SSS and MLD. Chen et al. [15] found that Kd at 490 nm has an effect on surface pCO$_2$. Figure 10 shows the proportion of each remote sensing variable in the CatBoost model to invert sea surface pCO$_2$. It can be found that the proportion of MLD and SST are above 30%. On the contrary, Adg, Chla and Kd account for only about 10%. As 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 factors to influence pCO$_2$, and biology may be one of the factors affecting sea surface pCO$_2$. From the conclusion of Luger [64], temperature and “biology” are major forcings, except for Chla, which confirmed our results.
Figure 10. The proportion of each remote sensing variable in the results of the inversion with the CatBoost model.

Figure 11 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 correlations in most regions. On the contrary, Chla, Kd and MLD have strong negative correlations with surface pCO$_2$ except for high-latitude regions. The most obvious area is the sub-region3 in Figure 8a and it may be more affected by biological activities or other factors, so it can explain the opposite trend about sub-region3 and the whole region in Figure 8d. 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. Maps of correlation coefficients between annual mean SST (a), Chla (b), Kd (c), MLD (d), and surface pCO$_2$ respectively.

Additionally, since SST is a major contributor to pCO$_2$, the association between SST and North Atlantic Oscillation (NAO) pattern are needed to be considered. Generally, the SST is associated with NAO, and it can be found that associations are more obvious in the part of the North Atlantic in some seasons, e.g., in the winter, pCO$_2$ in the Canary Sea has a relatively complicated trend. It is relatively high in 2009–2011 and 2018–2019 when the NAO has a higher trend to other periods. However, for the whole area, the relationship between the SST pattern and NAO is not obvious. More detailed analysis is beyond the scope of our study and it will be discussed in further research.
4.2. Advantages and Limitations of the CatBoost

Through the evaluation results, it can be found that the empirical CatBoost algorithm can estimate surface pCO$_2$ in the North Atlantic with the uncertainty of less than 5 µatm. Comparing to the published studies, the CatBoost model shows great advantages in different environments of North Atlantic, e.g., Friedrich showed the result (RMSE = 19.0 µatm) with neural network in the North Atlantic [36], and in the same area, Telszewski showed RMSE = 11.6 µatm with a self-organizing neural network [39]. Since the CatBoost algorithm solves the problem of gradient bias and prediction shift, 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, a problem is the CatBoost model works like a “black box”, so it cannot understand the 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$_2$ may be driven by these processes collectively, so an empirical approach cannot explain the interaction of different processes but can achieve a better model accuracy.

Additionally, the CatBoost model can be applied globally if two questions have been considered. For one thing, sufficient data measured by using the same method are needed. For another, because of the mediocre result for some extreme data, the model needs more parameters and adjustments if it is to be applied globally. These approaches need to be considered in further research.

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 comparison 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 impact 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 pCO$_2$ 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 similar. However, in low latitudes, the pCO$_2$ in the eastern Atlantic Ocean is obviously lower than that in the western.

2. The main impact 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 impact factors are different in each sub-region, on account of complex climate patterns.

The 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, H.S.; Methodology, H.S., B.Z.; Analysis, H.S., Y.C. and B.Z.; Writing and editing, H.S. and J.H.; Supervision, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the China Postdoctoral Science Foundation (2020M681825).

Data Availability Statement: Not applicable.

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

References


5. Fennel, K.; Wilkin, J.; Previdi, M.; Najjar, R. Denitrification effects on air-sea CO₂ fluxes in the northwestern European shelf based on voluntary observing ship and satellite observations. Prog. Oceanogr. 2017, 157, 94–110. [CrossRef]


