Estimating spatial and temporal variation in ocean surface pCO2 in the Gulf of Mexico using remote sensing and machine learning techniques

Zhiyi Fu, Linshu Hu, Zhende Chen, Feng Zhang, Zhou Shi, Bifeng Hu, Zhenhong Du, Renyi Liu

Science of the Total Environment 745 (2020) 140965

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Estimating spatial and temporal variation in ocean surface pCO_2 in the Gulf of Mexico using remote sensing and machine learning techniques

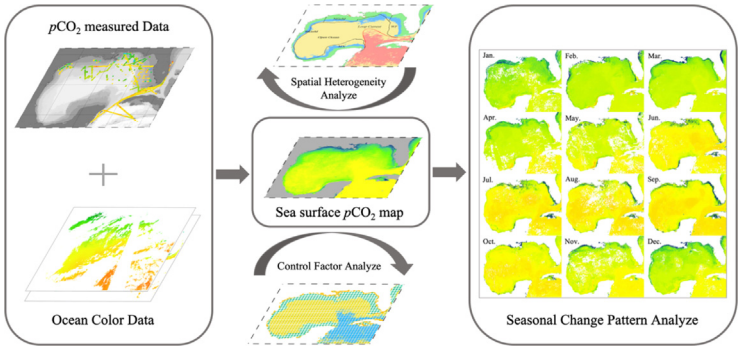
Zhiyi Fu [a], Linshu Hu [a], Zhende Chen [a], Feng Zhang [a,b,*], Zhou Shi [c], Bifeng Hu [d,e], Zhenhong Du [a,b], Renyi Liu [a,b,f]

[a] School of Earth Sciences, Zhejiang University, Hangzhou 310027, China
[b] Zhejiang Provincial Key Laboratory of Geographic Information Science, Hangzhou 310028, China
[c] Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
[d] Unité de Recherche en Science du Sol, INRAE, Orléans 45075, France
[e] Sciences de la Terre et de l'Univers, Orléans University, Orléans 45067, France
[f] Ocean Academy, Zhejiang University, Zhoushan 316021, China

HIGHLIGHTS

- A satellite-based surface pCO_2 model with good performance is proposed.
- Six sub-regions in study area were divided according to pCO_2 spatial heterogeneity.
- The specific control variables for different regional pCO_2 were identified.
- Variable changes and correlations were used to explain pCO_2 seasonal variation.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 24 March 2020
Received in revised form 9 July 2020
Accepted 12 July 2020
Available online 19 July 2020

Editor: Christian Herrera

Keywords:
Surface pCO_2
Remote sensing
Gulf of Mexico
Data mining

ABSTRACT

Research on the carbon cycle of coastal marine systems has been of wide concern recently. Accurate knowledge of the temporal and spatial distributions of sea-surface partial pressure (pCO_2) can reflect the seasonal and spatial heterogeneity of CO_2 flux and is, therefore, essential for quantifying the ocean's role in carbon cycling. However, it is difficult to use one model to estimate pCO_2 and determine its controlling variables for an entire region due to the prominent spatiotemporal heterogeneity of pCO_2 in coastal areas. Cubist is a commonly-used model for zoning; thus, it can be applied to the estimation and regional analysis of pCO_2 in the Gulf of Mexico (GOM). A cubist model integrated with satellite images was used here to estimate pCO_2 in the GOM, a river-dominated coastal area, using satellite products, including chlorophyll-a concentration (Chl-a), sea-surface temperature (SST) and salinity (SSS), and the diffuse attenuation coefficient at 490 nm (Kd-490). The model was based on a semi-mechanistic model and integrated the high-accuracy advantages of machine learning methods. The overall performance showed a root mean square error (RMSE) of 8.42 μatm with a coefficient of determination (R^2) of 0.87. Based on the heterogeneity of environmental factors, the GOM area was divided into 6 sub-regions, consisting estuaries, near-shores, and open seas, reflecting a gradient distribution of pCO_2. Factor importance and correlation analyses showed that salinity, chlorophyll-a, and temperature are the main controlling environmental variables of pCO_2, corresponding to both biological and physical effects. Seasonal changes in the GOM region were also analyzed and explained by changes in the environmental variables.

* Corresponding author at: School of Earth Sciences, Zhejiang University, Hangzhou 310027, China.
E-mail address: zfcarnation@zju.edu.cn (F. Zhang).

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

Therefore, considering both high accuracy and interpretability, the cubist-based model was an ideal method for pCO_2 estimation and spatiotemporal heterogeneity analysis.

1. Introduction

Coastal carbon fluxes account for 25–50% of the ocean's absorption of anthropogenic carbon dioxide (Sarmiento, 1993), playing a significant role in the global carbon cycle (Thomas et al., 2004; Laruelle et al., 2010; Lee et al., 2011). However, due to the complexity of its ecosystem, the current estimation of air-sea CO_2 flux in coastal areas is biased and inaccurate (Shim et al., 2007). Unlike the atmospheric pCO_2, which is fairly uniform, oceanic pCO_2 varies both spatially and temporally (Mahadevan et al., 2004). Physical and biogeochemical factors, including surface temperature-driven solubility, biological processes, fall-to-winter vertical mixing, ocean circulation, river runoff, and shelf-ocean exchange will influence the temporal and spatial variability of the surface-ocean pCO_2 on coastal areas (Sergio R. Signorini, 2013).

Environmental variables related to four pCO_2 controlling processes (e.g., biological effects, thermodynamic effects, mixing effects and air-sea exchange effects), such as sea-surface temperature (SST), sea-surface salinity (SSS), concentration (Chl-a) and diffuse attenuation coefficient (Kd), can be good indicators of sea surface pCO_2 due to the capture of environmental changes. Ocean thermodynamic effect is dependent on SST, an exponential function ($\boldsymbol{pCO_{2@T2} = pCO_{2@T1} \times e^{0.0423\times(T_2-T_1)}}$) were estimated to describe the relationship between surface pCO_2 and SST (Takahashi et al., 2002; Takahashi et al., 2009). The solubility of CO_2 and the dissociation constants of the carbonate system can be influenced by SST and SSS (Weiss, 1974; Lee et al., 1998; Millero et al., 2006). Physical processes (horizontal and vertical mixing) of different water masses that have distinct carbonate characteristics such as total alkalinity (TA) and dissolved inorganic carbon (DIC) can also be well tracked by SST and SSS (Lee et al., 2006; Yang and Byrne, 2015). Biological activities such as photosynthesis, respiration and calcification, can be implicitly expressed by optical parameters of Chl and Kd (Friedrich and Oschlies, 2009; Jamet et al., 2012; Lei et al., 2020). Most studies that estimate pCO_2, whether based on remote sensing images or observation data sets, employed statistical approaches such as multiple linear regression (MLR) (Lefèvre, 2002; Olsen et al., 2004; Jamet et al., 2007; Marrec et al., 2015), multiple polynomial regression (MPR) (Stephens et al., 1995; Ono et al., 2004; Zhu et al., 2009; Qin et al., 2014) and principle component regression (PCR) (Lohrenz and Cai, 2006; Lohrenz et al., 2010). Machine learning methods, such as self-organizing maps (SOMs) (Telszewski et al., 2009; Nakaoka et al., 2013), regression trees (Lohrenz et al., 2018), and feedforward neural networks (Jo et al., 2012; Moussa et al., 2016), have also performed well in estimating pCO_2. However, statistical-based research still lacks information related to the mechanistic processes that explain why coastal areas play a role as sinks or sources for carbon fluxes (Dai et al., 2013). At the same time, the predicted results cannot be explained by physical, chemical, and biological methods either (Chen et al., 2017).

The geographical conditions of the Gulf of Mexico (GOM) are complex. The north GOM is dominated by the Mississippi and Atchafalaya River system (MARS), which has been ranked as the seventh largest freshwater discharge system in the world (Milliman and Meade, 1983). "The dead zone" is adjacent to the outflows of the Mississippi and Atchafalaya Rivers (Rabalais et al., 2002), and GOM coastal wetlands provide many ecosystem services as well (Engle, 2011). The plume of MARS extends westward (Huang et al., 2013), while the Loop Current flows through the Florida Strait, into the Gulf Stream, and heads north up the eastern coast of the U.S. The diversity and heterogeneity of coastal ecosystems, which has been emphasized in earlier work (Borges et al., 2005), is particularly prominent in the Gulf of Mexico. Therefore, it is necessary to explain the physical and biogeochemical processes that control surface pCO_2 in the GOM area.

In order to overcome the inherent disadvantages of empirical methods, a nonlinear semi-mechanistic model was developed by Hales et al. (2012) to study the upwelling-dominated U.S. western margins. This model reproduces changes in DIC and TA caused by mixing processes and thermal forcing, and then uses CO_2 System Program (CO2SYS) (Pierrot and Wallace, 2006) to calculate surface pCO_2 from DIC and TA (Hales et al., 2012). The mechanistic semi-analytic algorithm (MeSAA) was developed to parameterize and quantify the contribution of major controlling factors in a mechanistic manner, like the inherent nonlinearities of the carbonate system, and was used to model summertime surface pCO_2 in the East China Sea (ECS), a river-dominated coastal ocean (Bai et al., 2015). Similarly, a satellite-based semi-mechanistic model was developed to consider surface pCO_2 as the sum of the pCO_2 caused by biological and mixing effects, and was applied to the north GOM area to estimate summer sea-surface pCO_2 (Le et al., 2019). Since mechanism-based methods can be effective and are more meaningful than simple empirical regression methods, semi-mechanistic methods have been increasingly applied to nearshore areas in recent years (Song et al., 2016). Although the semi-mechanistic model has an explanatory advantage, it has greater uncertainty due to uncertainties caused by river end-member properties and satellite-derived variables (Le et al., 2019); and because the strong surface delamination in the summer minimizes the vertical mixing effect, the model cannot be applied well in areas or seasons with strong upwelling (Chen et al., 2017).

A random-forest-based regression ensemble (RFRE) model, proposed by Chen et al., (2019), combined machine learning and semi-mechanistic methods to overcome the problems of high uncertainty and insufficient explanatory power. The random forest method, while dealing with complex problems, also provides the role of distinguishing factor differences. Therefore, a model using SST, SSS, Chl and diffuse attenuation coefficient at 488 nm (Kd, m^{-1}) using the semi-analytical algorithm developed by Lee (2005) (Kd_Lee) to implicitly interpret controlling processes can apply to different GOM partitions and seasons with satisfactory performance (Fennel et al., 2008; Ikawa et al., 2013). Developing a model that combines semi-mechanistic and machine learning methods is an effective way to estimate sea-surface pCO_2 in complex coastal areas. In practice, however, considerations of partitioning and spatial heterogeneity remain inadequate in most coastal ocean studies (Lohrenz et al., 2018). There is controversy over whether an estuary is a weak source or sink (Ternon et al., 2000; Borges and Abril, 2012), if the GOM's inner shelf is a moderate seasonal sink (Cai, 2003; Lohrenz et al., 2010; Huang et al., 2015), and if mid-latitude open oceans are usually a net sink for atmospheric CO_2 (Takahashi et al., 2009; Landschützer et al., 2014; Takahashi et al., 2014). Hence, more semi-mechanistic methods should be introduced to adequately characterize seasonal changes and underlying spatial patterns in the GOM, as previously emphasized (Robbins et al., 2014), and to explain the heterogeneity of factors and the natural laws of partitioning as well.

The cubist method is widely used for terrestrial carbon digital mapping and is a more popular machine learning method because of its high explanatory power (Adhikari et al., 2014; Gray et al., 2016; Rudiyanto et al., 2018). Cubist uses related environmental variables to build an estimating model; it divides subsets according to their geographic similarity (Ma et al., 2017a) and provides results on the relative importance of each variable in the model (Pouladi et al., 2019; Yan et al., 2020). Thus, the application of the Cubist model can reasonably reflect the spatial

heterogeneity of the GOM and explain the importance of interpretation factors by region.

In this study, we will combine the semi-mechanistic and cubist methods to predict ocean surface pCO_2 of the GOM with higher accuracy, using the four environmental variables: SST, SSS, Chl and Kd-490 (for brevity it is simply called Kd in this study). The model divide GOM area based on environmental variable differences rather than geographic locations for the first time. Spatiotemporal heterogeneity analyze of pCO_2 in the coastal area and main controlling environmental variables analyze are on the basis of model zoning. The main objectives of this study are to (1) develop a pCO_2-estimating model for a coastal area by using environmental variables; (2) divide the GOM into subregions according to model rules, analyze the spatial heterogeneity of pCO_2, and discuss the main factors used to indicate pCO_2-controlling processes based on the rules' linear equation; and (3) analyze seasonal changes in pCO_2 for the GOM and discuss its causes. This study can contribute to propose a model with general applicability to estimate surface pCO_2 from satellites for coastal areas and facilitate variables and forms selection of pCO_2 model construction in different regions of the GOM. Relating seasonal changes of environmental variable and pCO_2 will provide an idea for explaining reasons for the coastal areas act as a carbon source or carbon sink changing by seasons as well. Furtherly, will contribute to estimate the near-shore carbon flux, quantify the role of coastal area in the carbon cycle, and provide effective information for understanding the mechanism of ocean acidification.

2. Dataset and method

2.1. Study area

The Gulf of Mexico (GOM) has an area of 1.6 million km^2, including the West Florida shelf (WFS), the Louisiana shelf, the Texas shelf, the Mexico shelf, and the open bay. The Louisiana Continental Shelf (LCS) in the north is a typical river-dominated continental shelf, dominated by the Mississippi and the Atchafalaya River System (MARS). The GOM shoreline includes a variety of coastal habitats, costal strand beaches, adjacent marshes, and subaqueous habitats, and extends approximately 18–30.5°N and 80.6–98°W (Mendelssohn et al., 2017). “The dead zone” is the world's second largest dead oxygen zone, caused by nutritional fluxes from the Mississippi-Atchafalaya Basin coupled with temperature and density-induced stratification, and it has expanded in recent years (Larsen, 2004). The Loop Current flows into the bay through the Yucatan Channel and flows eastward from the Florida Straits. The general coastal climate is subtropical with warm to hot summers and cool winters (Ellis and Dean, 2012). Sea surface temperatures are lowest in February and highest in August. Surface winds are directed south-southwest in the summer, while mostly from the east in non-summer seasons. Typhoons also bring phytoplankton blooms into the northern GOM area (Shi and Wang, 2007).

2.2. Field data

The in-situ measured pCO_2 used in this study was downloaded from the Ocean Carbon Data Systems (OCADS, https://www.nodc.noaa.gov/ocads/). Twelve cruises collected data throughout the year of 2018 and are described in Table 1. These are data mainly distributed in the northern GOM area and partly distributed in the open sea (Fig. 2). The in-situ sea-surface properties include pCO_2, sea-surface salinity (SSS), and sea-surface temperature (SST). Sea-surface pCO_2 data were measured by using non-dispersive infrared analyzer Li-COR (model 7000) with a measurement frequency of 2 min and an accuracy of 2 uatm. The SSS and SST data were obtained ~3 m below the sea surface, using a CTD (SBE-38 or SBE-45, Seabird Inc.,) integrated in the underway pCO_2 system. All cruise data underwent quality control, with quality control flags for fCO_2 values (2 = good, 3 = questionable). The details for data sampling, processing, and quality control can be found in Pierrot et al., 2009.

Table 1
Underway pCO_2 measurements used for the development and validation of the pCO_2 model. These surface pCO_2 measurements were collected from different cruises, covering the entire year of 2018.

Cruise ID	Ship name	Date range	# of observations
EQ17	M/V Celebrity Equinox	1/1/2018–1/6/2018	2179
AS17	M/V Allure of the Seas	1/4/2018–1/7/2018	1198
GU1801_Leg1	R/V Gordon Gunter	1/14/2018–1/22/2018	4178
GU1801_Leg2	R/V Gordon Gunter	1/26/2018–2/9/2018	7421
GU1801_Leg3	R/V Gordon Gunter	2/12/2018–2/27/2018	5428
GU1801_Leg4	R/V Gordon Gunter	3/1/2018–3/16/2018	7941
GU1802	R/V Gordon Gunter	6/24/2018–7/9/2018	7609
GU1803-transit	R/V Gordon Gunter	7/11/2018–7/14/2018	1340
GU1803-Leg1	R/V Gordon Gunter	7/20/2018–8/3/2018	7196
GU1803-Leg2	R/V Gordon Gunter	8/6/2018–8/19/2018	4727
GU1804	R/V Gordon Gunter	8/23/2018–8/31/2018	4445
GU1805-Leg1	R/V Gordon Gunter	9/2/2018–9/9/2018	3563
GU1805-Leg2	R/V Gordon Gunter	9/11/2018–9/30/2018	9659
EQ18	M/V Celebrity Equinox	1/6/2018–12/22/2018	872
GU1806	R/V Gordon Gunter	11/10/2018–12/4/2018	10,127
Total from all cruises			77,883
Total used in model development and validation			7963

2.3. Satellite data

Daily standard NASA level-2 data products, obtained by the Moderate Resolution Imaging Spectroradiometer (MODIS/Aqua), were downloaded from NASA's ocean color website (http://oceancolor.gsfc.nasa.gov/). These level-2 ocean color data included properties such as chlorophyll-a and spectral remote sensing reflectance (Rrs) between 412 and 678 nm. We directly obtained chlorophyll-a and the diffuse attenuation coefficient at 490 nm (Kd-490) for the level-2 products calculated from spectral Rrs data. For details on algorithms, refer to the official NASA ocean color document (https://oceancolor.gsfc.nasa.gov/atbd/). SSS was calculated by using an empirical algorithm developed by Chen and Hu (2017), this multilayer perceptron neural network-based (MPNN) SSS model use Rrs at 412, 443, 488, 555, 667 nm and SST as input and has been proven to perform satisfactorily in both coastal and plume areas. SeaWiFS Data Analysis System software (Version 7.5) with up-to-date calibration coefficients were used to reprocess the images. According to NASA's recommendation, images with quality level > 1 were discarded from the daily level-2 SST data. The spatial resolution of the field data was averaged at 1 km in order to match the satellite data's resolution of approximately 1 km. Conjugate matching was possessed between daily images and in situ underway measurements. During the matching process, a time window of ± 6 h between in situ and MODIS measurements was used, and a median value from a 3×3 pixel box centered at each sampling site was used to filter sensor and algorithm noise.

2.4. Cubist model

Cubist is a spatial data mining algorithm based on the M5 algorithm and is essentially similar to regression trees. It recursively partitions the predictor variables in a divide-and-conquer way, discovering the unknown relationships between predictor and predicted variables, and then generates a rule-based prediction model (Quinlan, 1992; Quinlan, 1993; Holmes et al., 1999). The predicted variable will be divided into subsets that are more internally homogenous with respect to the target variable and covariates than the dataset as a whole (Ma et al., 2017b; Liang et al., 2019; Peng et al., 2019). Unlike the regression trees in CART (Classification and Regression Trees) (Breiman et al., 1984a, 1984b), Cubist's terminal nodes are multiple linear regression

equations rather than predictions, thus constructing regression models as piecewise linear models. Cubist models generally provide better results than those produced by simple techniques, such as multivariate linear regression, and are easier to understand than neural networks.

Cubist improves the accuracy of rule-based models by combining a composite model with an instance-based or nearest-neighbor algorithm. On this basis, a committee model constituted by multiple rules can be generated for each rule of the composite model; the subsequent rule in the committee model will be used to correct the predicted value of the previous rule (Walton, 2008; Pouladi et al., 2019). Cubist uses heuristic algorithms to automatically integrate both rule-based and composite models, yielding the smallest average absolute error value for modeling prediction results (Henderson et al., 2005; Miller et al., 2015).

Absolute|error|:

$$\frac{|T_1-P_1|+.....|T_n-P_n|}{n}$$

A series of rules used to define partitions are sorted in descending order of importance by cubist with the form: if {conditions} then linear model. This means that the first rule has the greatest contribution to accuracy when modeling the training data, while the last rule has the least. Thus, the distribution of the dataset on the variables are indicated by the rules. The variables used in the model are built on this subset and can be used to interpret the main variables that affect the subset in the meantime (Lacoste et al., 2014).

Therefore, we use cubist to (1) estimate the sea-surface pCO_2 of the GOM region according to the input environment variables; (2) analyze the distribution heterogeneity of the entire GOM region's pCO_2 on a spatial scale according to model rules; (3) analyze the main physical and biological processes affecting pCo_2 and their related environmental variables in different regions according to the most significant variable of each rule equation; (4) analyze patterns of seasonal change in pCO_2 across the GOM region based on pCO_2 seasonal maps predicted by the model. In this study, the cubist algorithms were employed in the R studio software package "*caret*".

3. Result and discussions

3.1. Model performance

The number of rule sets is a critical parameter for cubist; therefore, firstly, we compared multiple metrics to select the optimal number of rule sets. Fig. 3 shows that the root mean square error (RMSE) and coefficient of determination (R^2) was lowest and highest, respectively, with six rule sets; RMSE and R^2 increases and decreases, respectively, with more than six rule sets. This indicates that there is model over-fitting when exceeding six rules. Thus, six rules were selected when building the cubist prediction model to balance both model accuracy and simplicity.

The dataset was sorted and grouped, each group data was randomly divided by 8:2, after compositing group data, the dataset was divided in two parts. The 80% of dataset was used to build the prediction model and the rest was used to validate the model. The RMSE and R^2 were calculated to evaluate model performance. The performance of the cubist model in both the training and validation datasets is shown in Fig. 4, colored by data density. The R^2 is 0.93 and 0.87 for model development and validation, respectively, with an RMSE of 5.52 and 8.42 μatm. A histogram of the residuals (difference between measured and predicted pCO_2) for the combined datasets (both model training and validation data) is shown in Fig. 4c. The histogram shows that 98.4% of the residuals were smaller than 20.5 μatm pCO_2 standard deviation.

The other standard statistical measures, mean bias (MB) and mean absolute error (MAE), were also used to compare the cubist model with other methods. As presented in Table 2, the performances of machine learning methods were better, especially cubist, each index is significantly better than the other methods. Since MLR cannot simulate nonlinear characteristics well, it yielded greater error with an RMSE of 16.94 μatm and R^2 of 0.47 for validation. Although MNR method constructed model in a non-linear way, it cannot fully consider the inherent differences in the dataset, resulting in RMSE of 15.44 μatm and R^2 of 0.56. Worse performance shown in semi-mechanistic method with an RMSE of 37.79 μatm and R^2 of 0.27 was due to the great impact brought by the uncertainty of remote sensing products and river-end members on the model accuracy, and the fact that semi-mechanistic method was not applicable for other seasons except summer (a season when seawater stratification is evident). The R^2 and RMSE of the cubist model are equivalent to that of Chen's latest research (Chen et al., 2019) which showed an showed an overall performance of a root mean square difference (RMSD) of 9.1 μatm, with a R^2 of 0.95; the accuracy was lower but his model's correctness was higher. Considering its sufficient ability to explain factor heterogeneity, compared with other machine learning methods, cubist model can be viewed as a favorable method for estimating pCO_2 in the GOM.

Table 2
Comparison table of pCO_2 estimation approaches in the GOM, all methods used the same training dataset and validation dataset.

Approach		RMSE (μatm)	R^2	MB (μatm)	MAE (μatm)
MLR	Training	15.54	0.52	0.00	8.75
	Validation	16.94	0.47	−0.23	9.09
MNR	Training	14.65	0.57	0.00	8.81
	Validation	15.44	0.56	−0.24	8.89
Semi-mechanistic (Chen et al., 2017)	Training	37.91	0.28	0.00	27.15
	Validation	37.79	0.27	−0.20	26.81
Machine Learning (Cubist in this study)	Training	5.52	0.93	−0.07	3.18
	Validation	8.42	0.87	−0.41	4.55
Machine Learning (SVM)	Training	14.64	0.59	1.55	7.63
	Validation	15.94	0.55	1.34	7.93
Machine Learning (Neural Network)	Training	12.02	0.71	0.03	7.09
	Validation	13.11	0.69	−0.13	7.17

The year pCO_2 map estimated by the cubist model is shown in Fig. 5. It can be seen that the overall trend of pCO_2 showed a ring-shaped inward value increase. In the northern region, due to the physical and biological effects brought by river water mixing, low salinity and strong photosynthesis make pCO_2 was lower throughout the whole year (Lohrenz and Cai, 2006; Lohrenz et al., 2018; Le et al., 2019). In the open sea area, the pCO_2 was usually higher because of the minimal river influence. Meanwhile, as a result of the inflow of high salinity warm current water, high pCO_2 values extended eastward into the open sea area as well.

In summary, the cubist model estimated pCO_2 by fully considering the regional heterogeneity and effectively predicted pCO_2 within the range of 200–550 μatm. Despite the biological and physical effects on pCO_2 that were considered at the same time, the consequence of these were not considered separately, thus avoiding the propagation error generated when quantifying the two effects. Cubist model shows advantages over other methods in accuracy, and the correlation and contribution of the environmental variables can be more intuitively understood due to its tree structure that provides a linear model for each node.

3.2. Rule-based GOM partition

The cubist model divided the dataset by six rules and developed linear models for each subset (Table 3), each model has independently verified before analyzing pCO_2 differences in specific regions (Table 4). These subsets correspond to the rule-based sub-regions, which indicate the distribution of different spatially-heterogeneous regions in the GOM. Annual average images of four variables (SST, SSS, Kd and Chl)

Table 3
Cubist-generated linear models for each region, explaining the relationship between environmental variables and surface pCO_2.

Region	Linear models
1	2.2144 + 2.6 SSS − 218.8 Kd + 239.6 Chl
2	−467.385 + 228.7 Chl - 300.9 Kd + 8.58 SST + 9.6 SSS
3	−44.7571 + 92.1 Chl −132.9 Kd + 7.95 SST + 2.5 SSS
4	−216.4958 + 1.1 SSS − 547.8 Kd + 279.9 Chl + 0.74 SST
5	295.8847 + 2.84 SST + 0.3 SSS
6	91.5145 + 3.08 SST + 5.8 SSS - 20.5 Chl

were composited to correspond to the conditions of each region rule and apply in rule-based region division. Combining Figs. 1 and 6, it can be seen that the characteristics of the region are closely related to the distribution of the continental shelf and water bodies (wetlands and river), and the depth of the ocean as well.

On the whole, the dividing area of the cubist model is similar to the GOM continental shelf partition used by (Xue et al., 2016). Rules 2, 3, and 5 divided the GOM region mainly into three sub-regions, the near-shore, offshore, and open sea areas. The near-shore area surrounds the outer circle of the GOM, which is mostly concentrated in the north—that is, the river-dominated NGOM (northern GOM) area, whose depth is less than 90 m. The other area of region 2 closely corresponded to the distribution of wetlands (for example, the southern and eastern coastal wetland areas).

Regions 1 and 4 were distributed within the vicinity of surrounding water bodies and are areas that need great attention. Geographically, region 1 mainly distributed in estuaries, while region 4 has very little distribution and was basically in the surrounding of wetlands. And, from the perspective of rule division, the difference lies in the value of the environmental variable (Chl) used to indicate biological effects. The offshore area with a depth of less than 200 m, compared to region 2, was less affected by rivers or wetlands. However, region 3 still distributed within an area that is rich in phytoplankton, and its edges are roughly consistent with the water depth boundary, reaching the boundary between the near-shore and open seas. In addition, the different water mixing caused by different water flow directions (Fig. 1b) may be one of the reasons for the difference. The different water properties of mixed water in region 3 and region 2 may be reflected in SST and SSS. Since the former was the mixing with the Loop Current's high-temperature water caused by the single transportation from the north to the southeast, while the latter were mixed with low-salinity wetland waters on the edge of Florida. Near-shore regions 1, 2, 3 and the offshore region 4 nearly occupied the GOM's entire continental shelf. Region 5 occupied the largest area and covered the open sea, which was consistent with the conceptual open sea area. It separated from the coast, and the depth was at least greater than 200 m. The region that required particular attention was the scattered area clustered near the Yucatan. Evidently, the distribution of region 6 captured the trajectory of the Loop Current, and that along the tracks of the Florida Current, extending to the east as well.

3.3. Regional dominant factor

Only one or two processes may dominantly control the changes of surface pCO_2 in specific region (Bai et al., 2015), thus analyzing which corresponding environmental variables were the main dominant factors can well indicate the controlling processes in sub-regions. From the valid variables selected to build the linear equation of each subset, we can analyze the importance of each environmental variable. From the linear equations of each subset (Table 3), it can be seen that the SST variable participated in the construction of each rule equation and was the most significant environmental variable in the model. After normalization, the most important parameters of each equation can be de-

Table 4
Independent accuracy verification of each region model for both train and validation.

Region	Train					Validation				
	Count	RMSE (μatm)	R^2	MAE (μatm)	MB (μatm)	Count	RMSE (μatm)	R^2	MAE (μatm)	MB (μatm)
1	90	19.35	0.94	13.58	1.84	26	32.55	0.82	23.60	−7.44
2	636	7.97	0.95	5.02	−0.38	155	16.07	0.79	10.32	−1.35
3	624	5.13	0.96	2.98	−0.27	156	6.59	0.93	4.18	−0.84
4	176	13.03	0.80	7.98	−0.36	45	16.81	0.75	13.11	0.72
5	3802	3.66	0.67	2.48	−0.01	960	3.99	0.65	2.68	0.02
6	1015	4.66	0.97	3.01	−0.07	276	6.79	0.94	4.81	−0.66

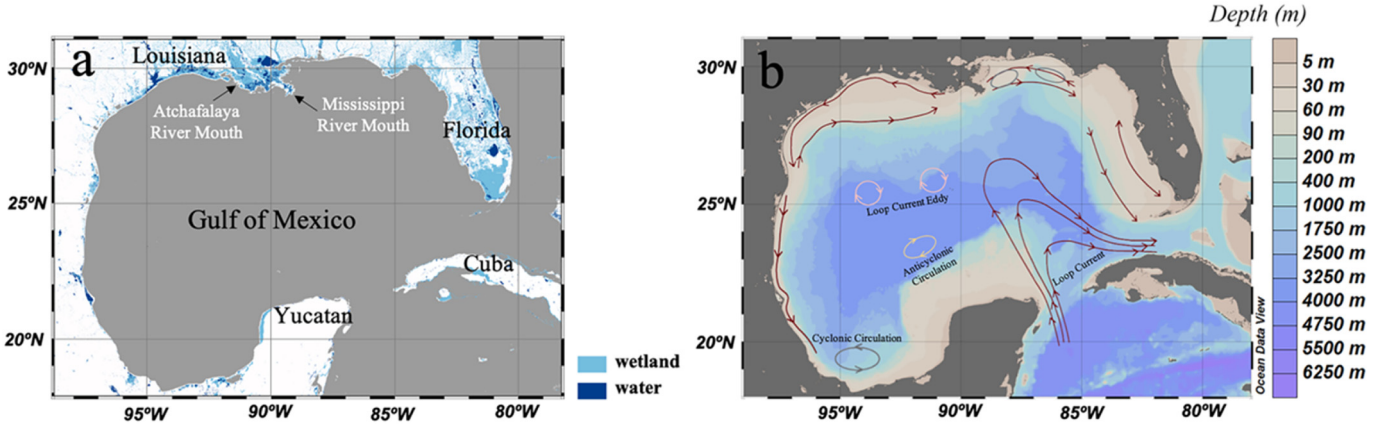


Fig. 1. (a) Geographical location of the GOM and distribution of its river and wetlands; (b) Water depth and typical Loop Current path in the GOM.

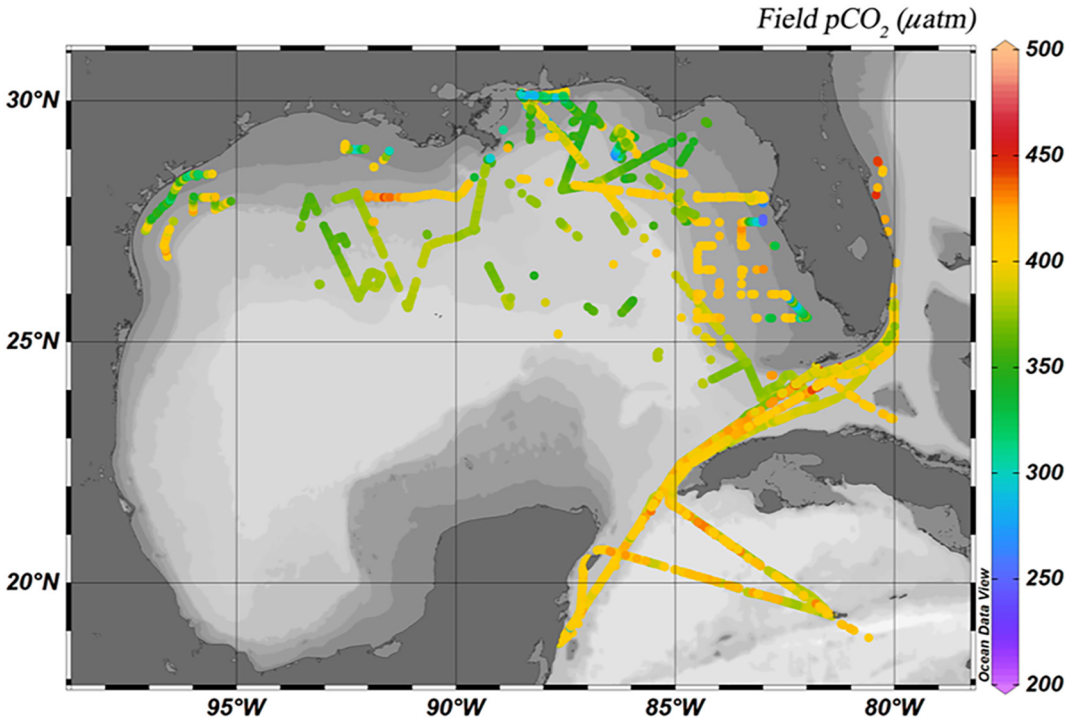


Fig. 2. Spatial distribution of conjugate samples of remote sensing and in-situ data in the GOM.

termined by the coefficients of the parameters. The parameters Kd and Chl always appeared simultaneously, and their coefficients were much larger than those of the remaining parameters. The region of rule 5 was determined to have a strong correlation with SST, since SST is the variable with largest weight among the two variables in the rule 5 equation. The subset of rule 1 and 3 has less samples, thus resulting in greater

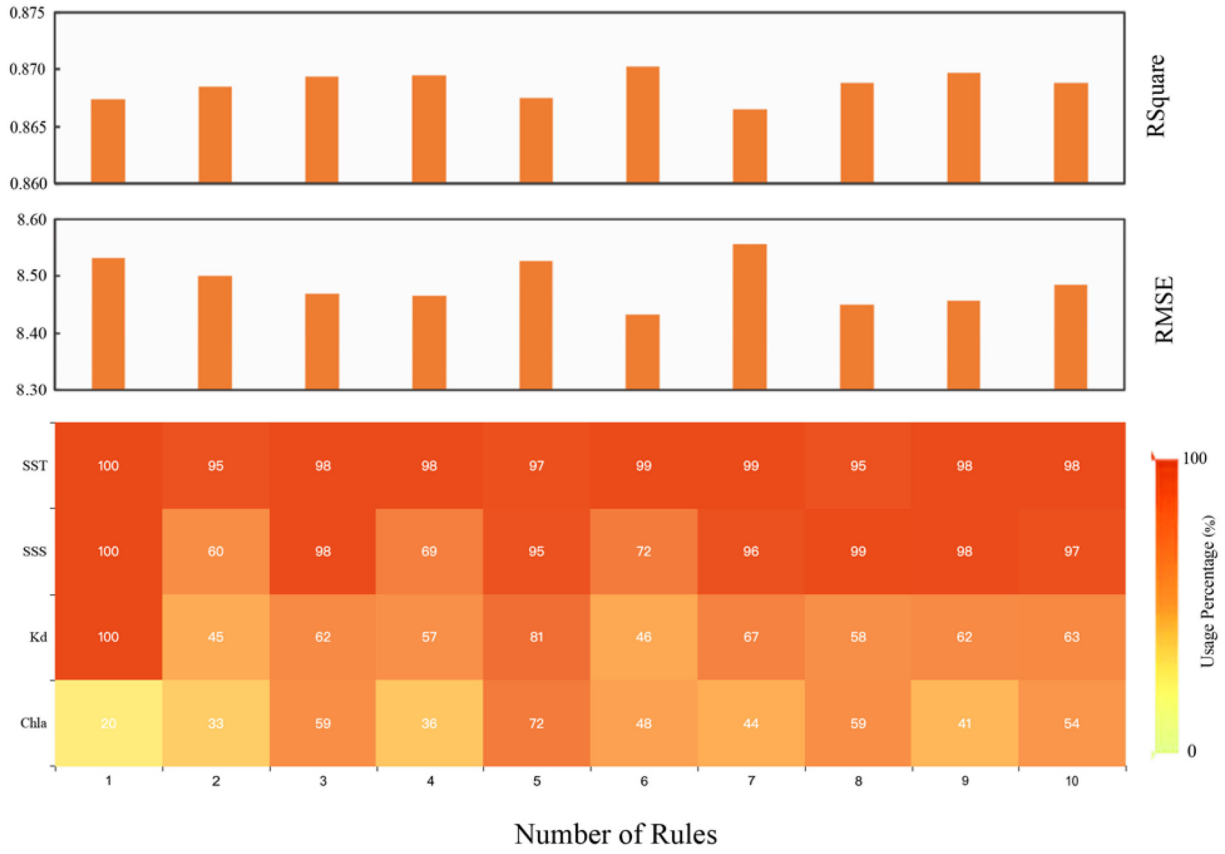


Fig. 3. Trend in the RMSE and R^2 when evaluating surface pCO_2 accuracy and interpretability as rule number increases, while listing the usage percentage of each environmental variable in model development.

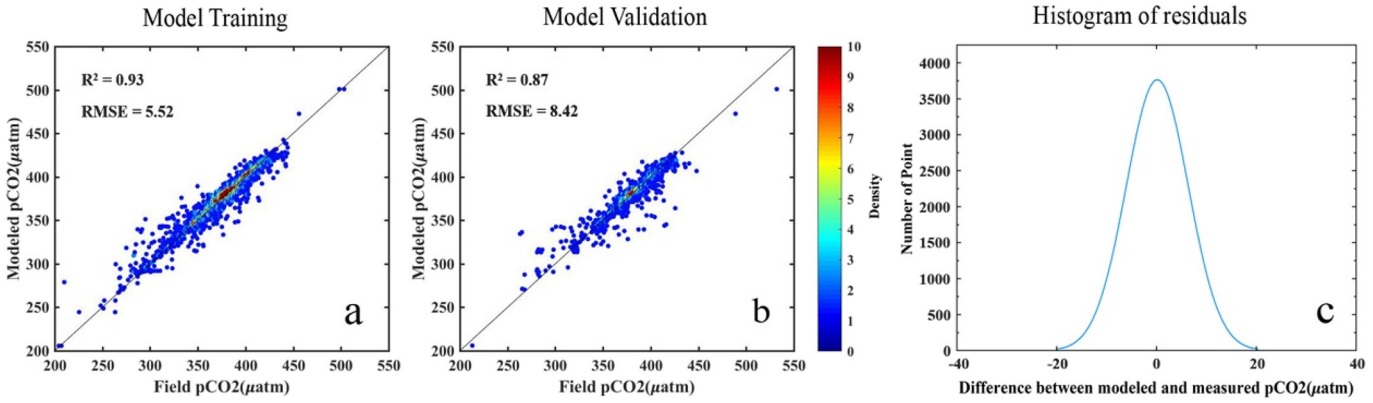


Fig. 4. Cubist model performance in estimating pCO_2 in both training (a) and validation (b); data pairs are color coded by data density. (c) Histogram of error distributions for both model development and validation.

uncertainties reflecting in higher RMSE of rule model verification. The most influential environmental variables in each rule regions are shown in Fig. 7. Based on the map, the entire Gulf area of the GOM was strongly affected by environmental variable Chl. The dominate factor of the open sea area was SST, and the area reflecting the Loop Current was affected by both SSS and SST. Combining Spatial correlation analyze can verify the rationality of the zoning from another perspective and contribute to analysis the main control factors.

In fact, the correlation between the four variables and pCO_2 also has the same distribution trend as the partitions (Fig. 8). The distribution of the correlation between Chl and Kd is basically the same, strong negative correlations appear in most areas. But the negative correlations were weak in the dead oxygen area and the circulation area, unlike the other regions. Compared to Chl and Kd, the entire GOM has a strong positive correlation with SST, with the exception of the estuary region and the region less than 30 m being weakly positively correlated. A strong correlation with salinity occurs mainly at the edge of the GOM, which corresponds to the distribution of wetlands and water bodies. On the contrary, a negative correlation appears in the Loop Current area, which is different from the open sea.

As for the NGOM shelf, the Mississippi-Atchafalaya River and associated plume play the most significant role in determining the distribution of pCO_2 (Chen et al., 2019). The significant influence of the river on the biogeochemistry of the estuary region is that rivers bring a large amount of inorganic and organic carbon (Cai, 2003; Bianchi et al., 2013) and an inflow of low-salinity river water rich in nutrients at the same time (Guo et al., 2012). The former increases the pCO_2 value, while the latter increases phytoplankton production. In addition, some chemical characteristics of river water (such as DIN loading) have been shown to positively correlate with pCO_2 (Lohrenz et al., 2008). Therefore, reasonable consideration should be given to the effects of the strong biological uptake caused by the input of river-borne nutrients as well as the rapid change in salinity caused by the mixing of river and seawater on pCO_2 (Lohrenz et al., 2018). The relationship between pCO_2 and salinity was a downward shape with higher values at low salinities, corresponding to the high fluvial inputs of DIC and AT, and higher values

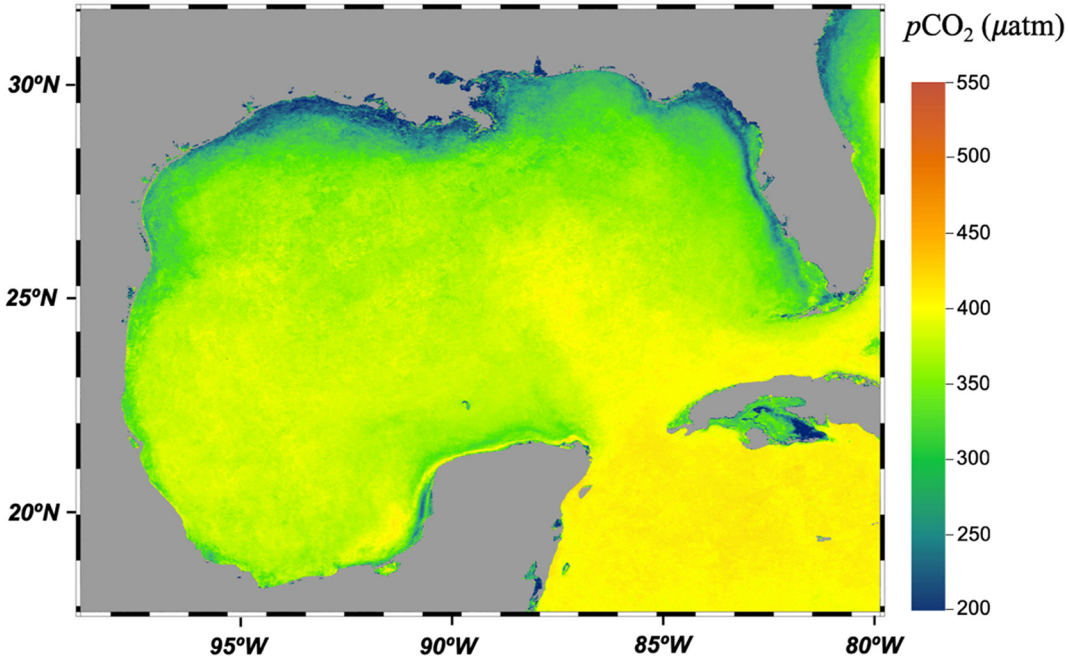


Fig. 5. Annual surface pCO_2 map generated by the cubist model, averaged from cruises.

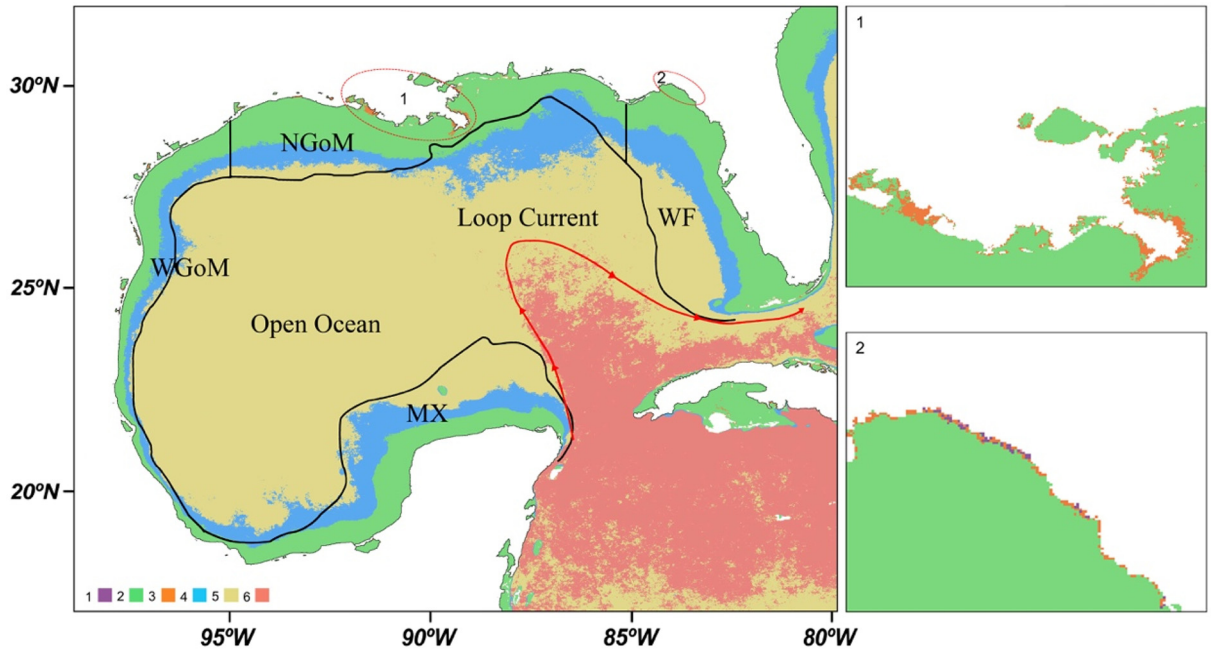


Fig. 6. Partitioning according to the model rules; the Loop Current and GOM geographic partitioning are marked on the figure.

corresponding to the high temperature and low production in the open sea (Guo et al., 2012; Huang et al., 2015). The strong drawdown of CO_2 caused by high productivity can be used to explain the lower values at intermediate salinities (Lohrenz et al., 1999; Lohrenz et al., 2008; Guo et al., 2012). However, this cannot be used to determine whether mixing or biological effects were the main processes for controlling pCO_2, clues can be provided by spatial correlation analyses. The estuary area has a strong correlation with salinity, but the variables Chl and Kd do not show a high correlation (the correlation coefficient is close to 0). Combined with the results of semi- mechanistic model research (Bai et al., 2015; Chen et al., 2017; Le et al., 2019), which quantified the two effects respectively, we can hypothesize that the variability of pCO_2 was controlled more strongly by mixing than biological effects in the near-shore plume waters. Further speculation can be drawn from salinity being the main environmental factor affecting estuary pCO_2. Future investigations can collect more remote sensing data and measure more conjugate samples to determine the dominant factors of pCO_2.

The inner shelf area (nearshore and offshore areas) is affected by both harmful and harmless algae blooming (Walsh et al., 2006). The strong biological effect of chlorophyll, due to its negative correlation with pCO_2 (Sarma et al., 2006), can be used to track when biological processes become dominant in the area. At the same time, due to the

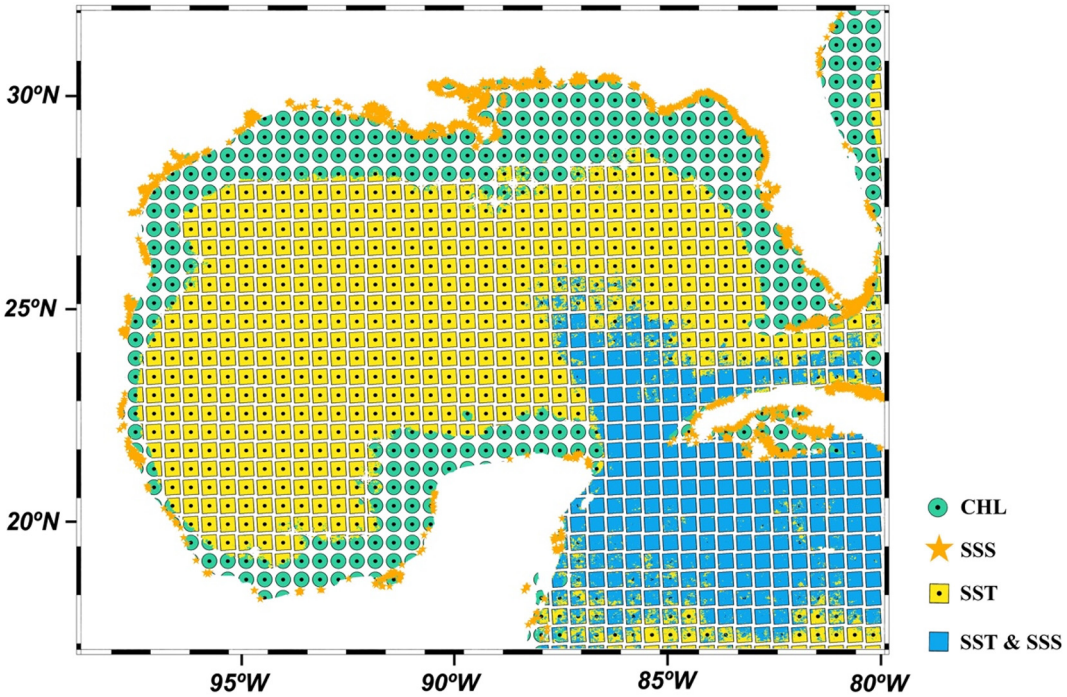


Fig. 7. Distribution of main environmental factors in the different regions, derived from the main variables of each linear model after normalization.

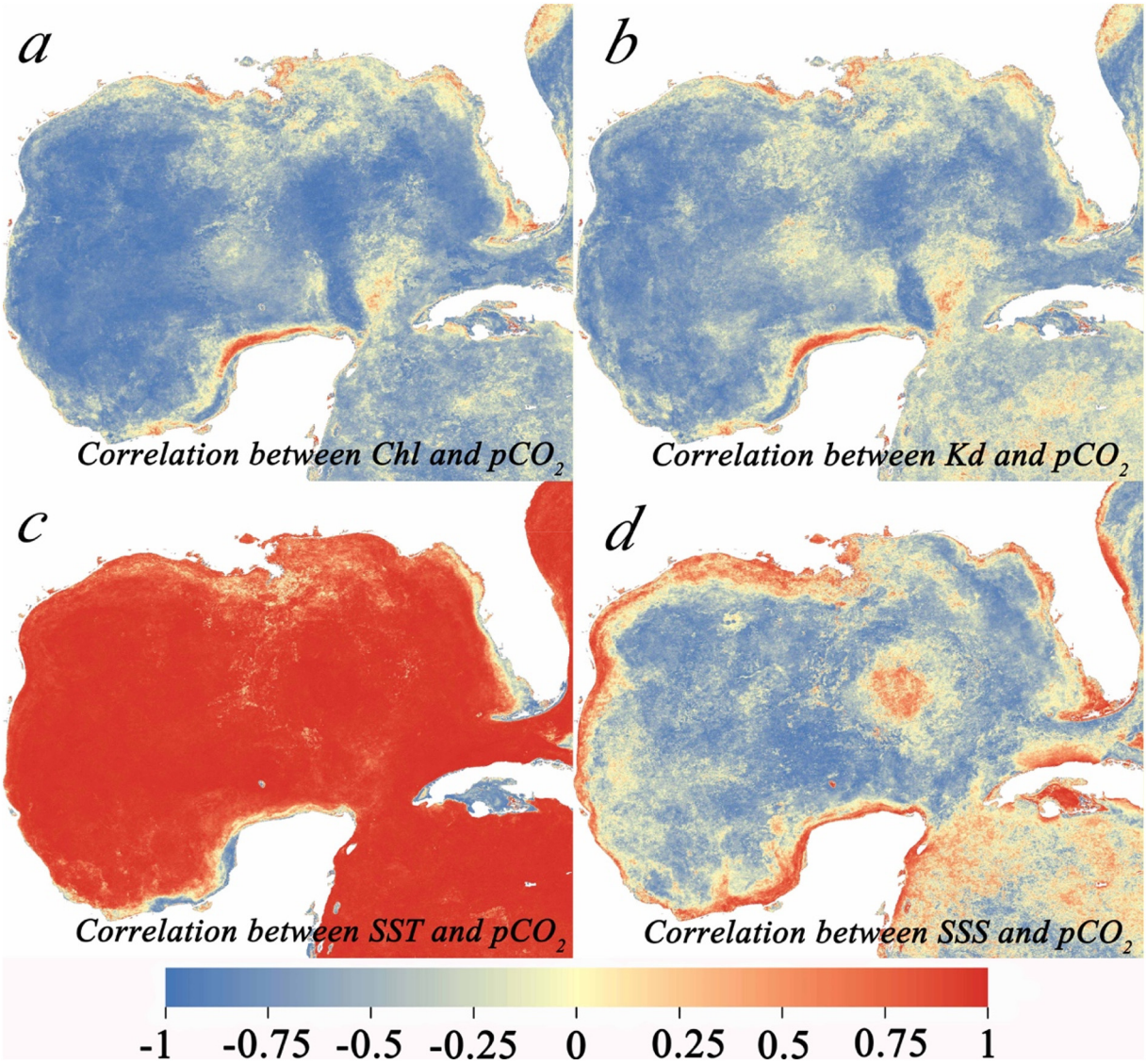


Fig. 8. Maps of correlation coefficients at 1 km resolution between Chl (a), Kd (b), SST (c), SSS (d), and surface pCO_2, respectively.

impact of abnormal weather and ocean currents in the summer, the dead oxygen area has expanded further. Abnormal algae breeding brought serious blooms and red tide problems, making the biological effects on pCO_2 more significant in the inner shelf area. Besides, some different features in the offshore area may indicate that circulation phenomena (Oey et al., 2013), such as eddy currents, will affect the value of pCO_2 in the offshore area. Therefore, areas with a depth of less than 90 m can be seen as a moderate seasonal CO_2 sink (Cai, 2003; Lohrenz et al., 2010; Huang et al., 2015).

Open seas are not affected by rivers; their seasonal changes are mainly controlled by seasonal temperature changes. It should be noted that the coefficient of temperature and salinity in the circulation zone are almost the same. The influence of high temperature on solubility (Chen et al., 2016) and the increase in seawater evaporation caused by high temperature (Takahashi et al., 2014) make the salinity in the circulation zone higher than in other open sea areas. Therefore, the circulation zone has high amounts of pCO_2 all year round under the combined effect of temperature and salinity.

3.4. Seasonal variations of surface pCO_2

The seasonal variation in the GOM is characterized by higher pCO_2 in summer, lower values in spring and winter, and lower or median values in the fall. The high values in the Loop Current are reflected throughout the year. This is consistent with most studies (Signorini et al., 2013; Chen et al., 2016; Lohrenz et al., 2018; Chen et al., 2019) (Fig. 9).

During non-summer seasons, the outflow from MARS is usually distributed westward along the Louisiana shelf, due to the force of wind directed downcoast from east to west (Feng et al., 2012). However, during the summer, strong winds forcing out of the south and west drive river plumes to distribute eastward and reverse the shelf cycle (Wiseman et al., 1997). The enhanced chlorophyll signal is mainly transmitted westward in the spring and early summer (Walker et al., 2005). Surface water dissolved inorganic carbon (DIC) also showed a westward increase in coastal distribution (Guo et al., 2012). This may explain why areas with low pCO_2 expand westward in non-summer months, opposite to that of summer.

In winter, the thermocline breakdown, MLD increase, and the decrease of SST occur at the same time (Liu and Weisberg, 2007). On one hand, as nutrients are brought to the surface by upwelling, phytoplankton blooms can occur and their negative correlation with chlorophyll can cause low pCO_2 values (Fig. 8a). The low pCO_2 value brought by lower temperatures is supported by the strong positive correlation between temperature and pCO_2 throughout the entire region (Fig. 8c). Thus, surface pCO_2 would significantly decrease by combining both biological and temperature effects. With the transition to spring, the

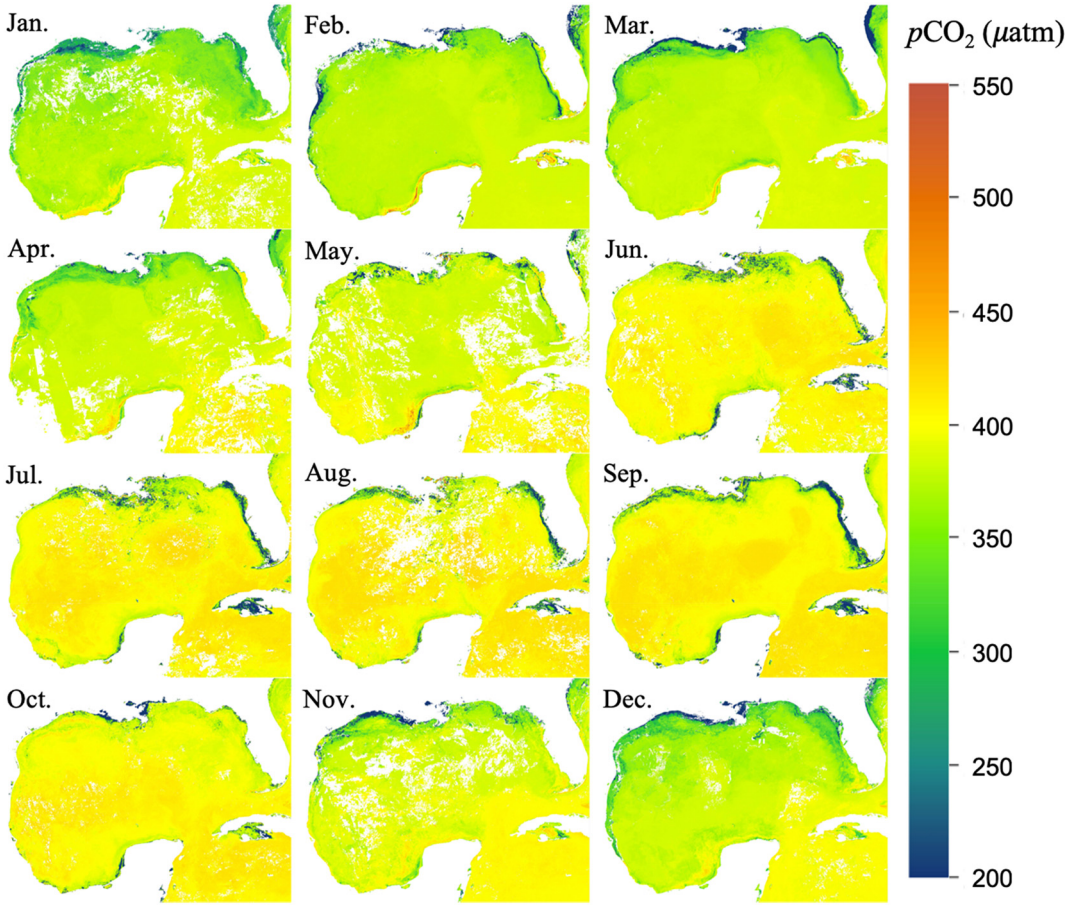


Fig. 9. Monthly map of surface pCO_2 in the GOM, derived from MODIS using cubist model for 2018.

estuary ushers in an increasing river flow, and the wind-induced plume was gradually directs to the east (Salisbury et al., 2004; Teague et al., 2006). The input of river water, which is rich in dissolved organic matter and nutrients, will increase phytoplankton productivity (Wawrik et al., 2003; Yuan et al., 2004), from which a bloom of phytoplankton usually occurs in March. As the temperature rises, pCO_2 also begins to rise, except in the estuary area, which is mostly controlled by biological effects.

In summer, the CO_2 from photosynthesis began to decrease as the main production was inhibited by nutritional depletion from seawater stratification. The increase in sea temperature and the greater respiration, compared with photosynthesis, increased pCO_2 (Chen et al., 2016). However, greater river discharge in late spring and summer caused strong mixing in the estuary while biological effects remained active at the same time. Thus, the estuary during summer still showed low pCO_2 (300–350 μatm), like in other seasons, due to a combination of biological and mixed effects. The pCO_2 maintained a high value in early autumn and then decreased as the water temperature became colder. The biological absorption of surface water CO_2 is suppressed with the consumption of nutrients, keeping the pCO_2 from decreasing further (Huang et al., 2012; Guo et al., 2012). Oxygen and organic matter promote the growth of bacteria, and bacteria break down organic matter in the water column to release carbon dioxide back into the seawater (Cai et al., 2011). Absorption decreased and release increased, resulting in the continued pCO_2 trend of late summer in early autumn.

4. Conclusion

In this study, we applied a cubist model to estimate the sea-surface pCO_2 from MODIS images for the GOM area and obtained a satisfied performance with an RMSE of 8.42 μatm and R^2 of 0.87. Model rules divided the GOM into six sub-regions. The circulation, river influence, and the distribution of wetlands could be used to explain the rationality of the zoning. The linear equations established for each region, with different dominating processes (e.g., physical and biological processes, etc.), made the prediction of pCO_2 in each zone more accurate. In addition, it provided clues for analyzing the dominant environmental factors of each district, helping explain the temporal and spatial variability of pCO_2 from the characteristics and temporal changes of environmental factors, such as the biological and physical factors of chlorophyll and temperature. Nutrient flux and salinity mix caused by river inflow resulted in strong physical and biological effects, making pCO_2 low all year round in the estuary area. In contrast, the flow of water with higher temperatures brought by the Loop Current distinguished the pCO_2 within the Loop Current from that of the open sea, which was not only controlled by temperature, resulting in a higher pCO_2 throughout the whole year. The difference between two regions in the inner shelf is the circulation trend and the influence of the wetland on the coastline, reflected in the gradient distribution of pCO_2. Phytoplankton blooms, red tides, and eddies result in chlorophyll concentration changes, making the internal shelf a seasonal sink. Our model performed well in estimating pCO_2, which provides a solid foundation for extending its application to others area with similar environmental and geographical conditions. The model also provides effective information for explaining the spatial heterogeneity of pCO_2 and exploring seasonal changes of pCO_2 in coastal areas.

CRediT authorship contribution statement

Zhiyi Fu: Investigation, Methodology, Writing - original draft. **Linshu Hu:** Methodology, Software, Validation. **Zhende Chen:** Data

curation, Visualization. **Feng Zhang:** Conceptualization, Formal analysis, Writing - review & editing. **Zhou Shi:** Conceptualization. **Bifeng Hu:** Writing - review & editing. **Zhenhong Du:** Funding acquisition. **Renyi Liu:** Supervision.

Declaration of competing interest

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.

Acknowledgments

This work was supported by the National Key R&D Program of China (2018YFB0505000, 2017YFB0503604), National Natural Science Foundation of China (41671391, 41922043, 41871287). We also thank NOAA NCEI and LDEO for providing all the available cruise data and thank NASA for providing MODIS satellite data and processing software.

References

Adhikari, K., Hartemink, A.E., Minasny, B., Bou Kheir, R., Greve, M.B., Greve, M.H., 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. PLoS One https://doi.org/10.1371/journal.pone.0105519.

Bai, Y., Cai, W.J., He, X., Zhai, W., Pan, D., Dai, M., Yu, P., 2015. A mechanistic semi-analytical method for remotely sensing sea surface pCO2 in river-dominated coastal oceans: a case study from the East China Sea. J. Geophys. Res. Ocean. 120, 2331–2349. https://doi.org/10.1002/2014JC010632.

Bianchi, T.S., Garcia-Tigreros, F., Yvon-Lewis, S.A., Shields, M., Mills, H.J., Butman, D., Osburn, C., Raymond, P., Shank, G.C., DiMarco, S.F., Walker, N., Reese, B.K., Mullins-Perry, R., Quigg, A., Aiken, G.R., Grossman, E.L., 2013. Enhanced transfer of terrestrially derived carbon to the atmosphere in a flooding event. Geophys. Res. Lett. https://doi.org/10.1029/2012GL054145.

Borges, A.V., Abril, G., 2012. Carbon dioxide and methane dynamics in estuaries. Treatise on Estuarine and Coastal Science https://doi.org/10.1016/B978-0-12-374711-2.00504-0.

Borges, A.V., Delille, B., Frankignoulle, M., 2005. Budgeting sinks and sources of CO2 in the coastal ocean: diversity of ecosystem counts. Geophys. Res. Lett. https://doi.org/10.1029/2005GL023053.

Breiman, L., Friedman, J.H., Olhsen, R.A., Stone, C.J., 1984a. Classification and Regression Trees. Wadsworth, Monterrey, CA.

L. Breiman, J.H. Friedman, R.A. Olhsen, C.J.S., 1984b. Classification and Regression Trees.

Cai, W.J., 2003. Riverine inorganic carbon flux and rate of biological uptake in the Mississippi River plume. Geophys. Res. Lett. https://doi.org/10.1029/2002GL016312.

Cai, W.J., Hu, X., Huang, W.J., Murrell, M.C., Lehrter, J.C., Lohrenz, S.E., Chou, W.C., Zhai, W., Hollibaugh, J.T., Wang, Y., Zhao, P., Guo, X., Gundersen, K., Dai, M., Gong, G.C., 2011. Acidification of subsurface coastal waters enhanced by eutrophication. Nat. Geosci. https://doi.org/10.1038/ngeo1297.

Chen, S., Hu, C., 2017. Remote sensing of environment estimating sea surface salinity in the northern Gulf of Mexico from satellite ocean color measurements. Remote Sens. Environ. 201, 115–132. https://doi.org/10.1016/j.rse.2017.09.004.

Chen, S., Hu, C., Byrne, R.H., Robbins, L.L., Yang, B., 2016. Remote estimation of surface pCO2 on the West Florida Shelf. Cont. Shelf Res. 128, 10–25. https://doi.org/10.1016/j.csr.2016.09.004.

Chen, S., Hu, C., Cai, W., Yang, B., 2017. Estimating surface pCO2 in the northern Gulf of Mexico: which remote sensing model to use? Cont. Shelf Res. J. 151, 94–110. https://doi.org/10.1016/j.csr.2017.10.013.

Chen, S., Hu, C., Barnes, B.B., Wanninkhof, R., Cai, W., Barbero, L., Pierrot, D., 2019. A machine learning approach to estimate surface ocean pCO2 from satellite measurements. Remote Sens. Environ. 228, 203–226. https://doi.org/10.1016/j.rse.2019.04.019.

Dai, M., Cao, Z., Guo, X., Zhai, W., Liu, Z., Yin, Z., Xu, Y., Gan, J., Hu, J., Du, C., 2013. Why are some marginal seas sources of atmospheric CO2? Geophys. Res. Lett. https://doi.org/10.1002/grl.50390.

Ellis, J.T., Dean, B.J., 2012. Gulf of Mexico processes. J. Coast. Res. https://doi.org/10.2112/si_60_2.

Engle, V.D., 2011. Estimating the provision of ecosystem services by Gulf of Mexico coastal wetlands. Wetlands https://doi.org/10.1007/s13157-010-0132-9.

Feng, Y., Dimarco, S.F., Jackson, G.A., 2012. Relative role of wind forcing and riverine nutrient input on the extent of hypoxia in the northern Gulf of Mexico. Geophys. Res. Lett. https://doi.org/10.1029/2012GL051192.

Fennel, K., Wilkin, J., Previdi, M., Najjar, R., 2008. Denitrification effects on air-sea CO2 flux in the coastal ocean: simulations for the northwest North Atlantic. Geophys. Res. Lett. https://doi.org/10.1029/2008GL036147.

Friedrich, T., Oschlies, A., 2009. Neural network-based estimates of North Atlantic surface pCO2 from satellite data: a methodological study. J. Geophys. Res. Ocean. https://doi.org/10.1029/2007JC004646.

Gray, J.M., Bishop, T.F.A., Smith, P.L., 2016. Digital mapping of pre-European soil carbon stocks and decline since clearing over New South Wales, Australia. Soil Res 54, 49–63. https://doi.org/10.1071/SR14307.

Guo, X., Cai, W.J., Huang, W.J., Wang, Y., Chen, F., Murrell, M.C., Lohrenz, S.E., Jiang, L.Q., Dai, M., Hartmann, J., Lin, Q., Culp, R., 2012. Carbon dynamics and community production in the Mississippi river plume. Limnol. Oceanogr. https://doi.org/10.4319/lo.2012.57.1.0001.

Hales, B., Strutton, P.G., Saraceno, M., Letelier, R., Takahashi, T., Feely, R., Sabine, C., Chavez, F., 2012. Satellite-based prediction of pCO2 in coastal waters of the eastern North Pacific. Prog. Oceanogr. 103, 1–15. https://doi.org/10.1016/j.pocean.2012.03.001.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma https://doi.org/10.1016/j.geoderma.2004.06.007.

Holmes, G., Hall, M., Prank, E., 1999. Generating Rule Sets from Model Trees, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/3-540-46695-9_1.

Huang, W.J., Cai, W.J., Powell, R.T., Lohrenz, S.E., Wang, Y., Jiang, L.Q., Hopkinson, C.S., 2012. The stoichiometry of inorganic carbon and nutrient removal in the Mississippi River plume and adjacent continental shelf. Biogeosciences https://doi.org/10.5194/bg-9-2781-2012.

Huang, W.J., Cai, W.J., Castelao, R.M., Wang, Y., Lohrenz, S.E., 2013. Effects of a wind-driven cross-shelf large river plume on biological production and CO2 uptake on the Gulf of Mexico during spring. Limnol. Oceanogr. https://doi.org/10.4319/lo.2013.58.5.1727.

Huang, W., Cai, W., Lohrenz, S.E., 2015. The carbon dioxide (CO2) system on the Mississippi River-dominated continental shelf in the northern Gulf of Mexico: 1. Distribution and air-sea CO2 flux. J. Geophys. Res. Ocean. https://doi.org/10.1002/2014JC010498.

Ikawa, H., Faloona, I., Kochendorfer, J., Paw .K.T., U, Oechel, W.C., 2013. Air-sea exchange of CO2 at a Northern California coastal site along the California Current upwelling system. Biogeosciences https://doi.org/10.5194/bg-10-4419-2013.

Jamet, C., Moulin, C., Lefévre, N., 2007. Estimation of the oceanic pCO2 in the North Atlantic from VOS lines in-situ measurements: parameters needed to generate seasonally mean maps. Ann. Geophys. https://doi.org/10.5194/angeo-25-2247-2007.

Jamet, C., Loisel, H., Dessailly, D., 2012. Retrieval of the spectral diffuse attenuation coefficient K d (λ) in open and coastal ocean waters using a neural network inversion. J. Geophys. Res. Ocean 117. https://doi.org/10.1029/2012jc008076 n/a-n/a.

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma https://doi.org/10.1016/j.geoderma.2013.07.002.

Landschützer, P., Gruber, N., Bakker, D.C.E., Schuster, U., 2014. Recent variability of the global ocean carbon sink. Glob. Biogeochem. Cycles https://doi.org/10.1002/2014GB004853.

Larsen, J., 2004. Dead zones increasing in world's coastal waters. USA TODAY, NEW YORK, pp. 29–54.

Laruelle, G.G., Dürr, H.H., Slomp, C.P., Borges, A.V., 2010. Evaluation of sinks and sources of CO2 in the global coastal ocean using a spatially-explicit typology of estuaries and continental shelves. Geophys. Res. Lett. https://doi.org/10.1029/2010GL043691.

Le, C., Gao, Y., Cai, W., Lehrter, J.C., Bai, Y., Jiang, Z., 2019. Estimating summer sea surface pCO2 on a river-dominated continental shelf using a satellite-based semi-mechanistic model. Remote Sens. Environ. 225, 115–126. https://doi.org/10.1016/j.rse.2019.02.023.

Lee, Z.-P., 2005. Diffuse attenuation coefficient of downwelling irradiance: an evaluation of remote sensing methods. J. Geophys. Res. 110, C02017. https://doi.org/10.1029/2004JC002573.

Lee, K., Wanninkhof, R., Takahashi, T., Doney, S.C., Feely, R.A., 1998. Low interannual variability in recent oceanic uptake of atmospheric carbon dioxide. Nature https://doi.org/10.1038/24139.

Lee, K., Tong, L.T., Millero, F.J., Sabine, C.L., Dickson, A.G., Goyet, C., Park, G.H., Wanninkhof, R., Feely, R.A., Key, R.M., 2006. Global relationships of total alkalinity with salinity and temperature in surface waters of the world's oceans. Geophys. Res. Lett. 33, 1–5. https://doi.org/10.1029/2006GL027207.

Lee, K., Sabine, C.L., Tanhua, T., Kim, T.W., Feely, R.A., Kim, H.C., 2011. Roles of marginal seas in absorbing and storing fossil fuel CO2. Energy Environ. Sci. https://doi.org/10.1039/c0ee00663g.

Lefèvre, N., 2002. Observations of pCO 2 in the coastal upwelling off Chile: spatial and temporal extrapolation using satellite data. J. Geophys. Res. https://doi.org/10.1029/2000jc000395.

Lei, S., Xu, J., Li, Y., Lyu, H., Liu, G., Zheng, Z., Xu, Y., Du, C., Zeng, S., Wang, H., Dong, X., Cai, X., Li, J., 2020. Temporal and spatial distribution of Kd(490) and its response to precipitation and wind in lake Hongze based on MODIS data. Ecol. Indic. 108, 105684. https://doi.org/10.1016/j.ecolind.2019.105684.

Liang, Z., Chen, S., Yang, Y., Zhou, Y., Shi, Z., 2019. High-resolution three-dimensional mapping of soil organic carbon in China: effects of SoilGrids products on national modeling. Sci. Total Environ. 685, 480–489. https://doi.org/10.1016/j.scitotenv.2019.05.332.

Liu, Y., Weisberg, R.H., 2007. Ocean currents and sea surface heights estimated across the west Florida shelf. J. Phys. Oceanogr. https://doi.org/10.1175/JPO3083.1.

Lohrenz, S.E., Cai, W.J., 2006. Satellite ocean color assessment of air-sea fluxes of CO2 in a river-dominated coastal margin. Geophys. Res. Lett. https://doi.org/10.1029/2005GL023942.

Lohrenz, S.E., Fahnenstiel, G.L., Redalje, D.G., Lang, G.A., Dagg, M.J., Whitledge, T.E., Dortch, Q., 1999. Nutrients, irradiance, and mixing as factors regulating primary production in coastal waters impacted by the Mississippi River plume. Cont. Shelf Res. https://doi.org/10.1016/S0278-4343(99)00012-6.

Lohrenz, S.E., Redalje, D.G., Cai, W.J., Acker, J., Dagg, M., 2008. A retrospective analysis of nutrients and phytoplankton productivity in the Mississippi River plume. Cont. Shelf Res. https://doi.org/10.1016/j.csr.2007.06.019.

Lohrenz, S.E., Cai, W.J., Chen, F., Chen, X., Tuel, M., 2010. Seasonal variability in air-sea fluxes of CO2 in a river-influenced coastal margin. J. Geophys. Res. Ocean. https://doi.org/10.1029/2009JC005608.

Lohrenz, S.E., Cai, W., Chakraborty, S., Huang, W., Guo, X., He, R., Xue, Z., Fennel, K., Howden, S., Tian, H., 2018. Satellite estimation of coastal pCO2 and air-sea flux of carbon dioxide in the northern Gulf of Mexico. Remote Sens. Environ. 207, 71–83. https://doi.org/10.1016/j.rse.2017.12.039.

Ma, Z., Shi, Z., Zhou, Y., Xu, J., Yu, W., Yang, Y., 2017a. A spatial data mining algorithm for downscaling TMPA 3B43 V7 data over the Qinghai–Tibet Plateau with the effects of systematic anomalies removed. Remote Sens. Environ. 200, 378–395. https://doi.org/10.1016/j.rse.2017.08.023.

Ma, Z., Zhou, Y., Hu, B., Liang, Z., Shi, Z., 2017b. Downscaling annual precipitation with TMPA and land surface characteristics in China. Int. J. Climatol. https://doi.org/10.1002/joc.5148.

Mahadevan, A., Lévy, M., Mémery, L., 2004. Mesoscale variability of sea surface pCO2: what does it respond to? Glob. Biogeochem. Cycles https://doi.org/10.1029/2003gb002102.

Marrec, P., Cariou, T., Macé, E., Morin, P., Salt, L.A., Vernet, M., Taylor, B., Paxman, K., Bozec, Y., 2015. Dynamics of air-sea CO2 fluxes in the northwestern European shelf based on voluntary observing ship and satellite observations. Biogeosciences https://doi.org/10.5194/bg-12-5371-2015.

Mendelssohn, I.A., Byrnes, M.R., Kneib, R.T., Vittor, B.A., 2017. Coastal habitats of the Gulf of Mexico. Habitats and Biota of the Gulf of Mexico: Before the Deepwater Horizon Oil Spill https://doi.org/10.1007/978-1-4939-3447-8_6.

Miller, B.A., Koszinski, S., Wehrhan, M., Sommer, M., 2015. Impact of multi-scale predictor selection for modeling soil properties. Geoderma https://doi.org/10.1016/j.geoderma.2014.09.018.

Millero, F.J., Graham, T.B., Huang, F., Bustos-Serrano, H., Pierrot, D., 2006. Dissociation constants of carbonic acid in seawater as a function of salinity and temperature. Mar. Chem. https://doi.org/10.1016/j.marchem.2005.12.001.

Milliman, J.D., Meade, R.H., 1983. World-wide delivery of sediment to the oceans. J. Geol. https://doi.org/10.1086/628741.

Moussa, H., Benallal, M.A., Goyet, C., Lefèvre, N., 2016. Satellite-derived CO2 fugacity in surface seawater of the tropical Atlantic Ocean using a feedforward neural network. Int. J. Remote Sens. https://doi.org/10.1080/01431161.2015.1131872.

Nakaoka, S., Telszewski, M., Nojiri, Y., Yasunaka, S., Miyazaki, C., Mukai, H., Usui, N., 2013. Estimating temporal and spatial variation of ocean surface pCO2 in the North Pacific using a self-organizing map neural network technique. Biogeosciences https://doi.org/10.5194/bg-10-6093-2013.

Oey, L.Y., Ezer, T., Lee, H.C., 2013. Loop current, rings and related circulation in the Gulf of Mexico: a review of numerical models and future challenges. Circulation in the Gulf of Mexico: Observations and Models https://doi.org/10.1029/161GM04.

Olsen, A., Triñanes, J.A., Wanninkhof, R., 2004. Sea-air flux of CO2 in the Caribbean Sea estimated using in situ and remote sensing data. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2003.10.011.

Ono, T., Saino, T., Kurita, N., Sasaki, K., 2004. Basin-scale extrapolation of shipboard pCO2 data by using satellite SST and Chla. Int. J. Remote Sens. 25, 3803–3815. https://doi.org/10.1080/01431160310001657515.

Peng, J., Biswas, A., Jiang, Q., Zhao, R., Hu, J., Hu, B., Shi, Z., 2019. Estimating soil salinity from remote sensing and terrain data in southern Xinjiang Province, China. Geoderma https://doi.org/10.1016/j.geoderma.2018.08.006.

Pierrot, D.E.L., Wallace, D.W.R., 2006. MS Excel Program Developed for CO2 System Calculations, ORNL/CDIAC-105. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, TN.

Pierrot, D., Neill, C., Sullivan, K., Castle, R., Wanninkhof, R., Lüger, H., Johannessen, T., Olsen, A., Feely, R.A., Cosca, C.E., 2009. Recommendations for autonomous underway pCO2 measuring systems and data-reduction routines. Deep. Res. Part II Top. Stud. Oceanogr. 56, 512–522. https://doi.org/10.1016/j.dsr2.2008.12.005.

Pouladi, N., Møller, A.B., Tabatabai, S., Greve, M.H., 2019. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma https://doi.org/10.1016/j.geoderma.2019.02.019.

Qin, B.Y., Tao, Z., Li, Z.W., Yang, X.F., 2014. Seasonal changes and controlling factors of sea surface pCO2 in the Yellow Sea. IOP Conference Series: Earth and Environmental Science https://doi.org/10.1088/1755-1315/17/1/012025.

Quinlan, J.R., 1992. Learning with continuous classes. Mach. Learn. 92, 343–348 (doi: 10.1.1.34.885).

Quinlan, J.R., 1993. Combining instance-based and model-based learning. Machine Learning Proceedings 1993 https://doi.org/10.1016/b978-1-55860-307-3.50037-x.

Rabalais, N.N., Turner, R.E., Wiseman, W.J., 2002. Gulf of Mexico hypoxia, a.k.a. "The dead zone.". Annu. Rev. Ecol. Syst. https://doi.org/10.1146/annurev.ecolsys.33.010802.150513.

Rudiyanto Minasny, B., Setiawan, B.I., Saptomo, S.K., McBratney, A.B., 2018. Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. Geoderma https://doi.org/10.1016/j.geoderma.2017.10.018.

Salisbury, J.E., Campbell, J.W., Linder, E., David Meeker, L., Müller-Karger, F.E., Vörösmarty, C.J., 2004. On the seasonal correlation of surface particle fields with wind stress and Mississippi discharge in the northern Gulf of Mexico. Deep. Res. Part II Top. Stud. Oceanogr. https://doi.org/10.1016/j.dsr2.2004.03.002.

Sarma, V.V.S.S., Saino, T., Sasaoka, K., Nojiri, Y., Ono, T., Ishii, M., Inoue, H.Y., Matsumoto, K., 2006. Basin-scale pCO2 distribution using satellite sea surface temperature, Chl a, and climatological salinity in the North Pacific in spring and summer. Glob. Biogeochem. Cycles https://doi.org/10.1029/2005GB002594.

Sarmiento, J.L., 1993. OCEAN. Chem. Eng. News 71 (22), 30–43.

Shi, W., Wang, M., 2007. Observations of a Hurricane Katrina-induced phytoplankton bloom in the Gulf of Mexico. Geophys. Res. Lett. 34, 1–5. https://doi.org/10.1029/2007GL029724.

Shim, J.H., Kim, D., Kang, Y.C., Lee, J.H., Jang, S.T., Kim, C.H., 2007. Seasonal variations in pCO2 and its controlling factors in surface seawater of the northern East China Sea. Cont. Shelf Res. https://doi.org/10.1016/j.csr.2007.07.005.

Signorini, Sergio R., 2013. Surface ocean pCO2 seasonality and sea-air CO2 flux estimates for the North American east coast. J. Geophys. Res. Ocean. 118, 5439–5460.

Signorini, S.R., Mannino, A., Najjar, R.G., Friedrichs, M.A.M., Cai, W.J., Salisbury, J., Wang, Z.A., Thomas, H., Shadwick, E., 2013. Surface ocean pCO2 seasonality and sea-air CO2 flux estimates for the North American east coast. J. Geophys. Res. Ocean. 118, 5439–5460. https://doi.org/10.1002/jgrc.20369.

Song, X., Bai, Y., Cai, W.J., Arthur Chen, C.T., Pan, D., He, X., Zhu, Q., 2016. Remote sensing of sea surface pCO2 in the Bering sea in summer based on a mechanistic semi-analytical algorithm (MeSAA). Remote Sens. https://doi.org/10.3390/rs8070558.

Stephens, M.P., Samuels, G., Olson, D.B., Fine, R.A., Takahashi, T., 1995. Sea-air flux of CO2 in the North Pacific using shipboard and satellite data. J. Geophys. Res. 100. https://doi.org/10.1029/95jc00901.

Takahashi, T., Sutherland, S.C., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., Bates, N., Wanninkhof, R., Feely, R.A., Sabine, C., Olafsson, J., Nojiri, Y., 2002. Global sea-air CO2 flux based on climatological surface ocean pCO2, and seasonal biological and temperature effects. Deep. Res. Part II Top. Stud. Oceanogr. https://doi.org/10.1016/S0967-0645(02)00003-6.

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

Takahashi, T., Sutherland, S.C., Chipman, D.W., Goddard, J.G., Ho, C., 2014. Climatological distributions of pH, pCO2, total CO2, alkalinity, and CaCO3 saturation in the global surface ocean, and temporal changes at selected locations. Mar. Chem. 164, 95–125. https://doi.org/10.1016/j.marchem.2014.06.004.

Teague, W.J., Jarosz, E., Carnes, M.R., Mitchell, D.A., Hogan, P.J., 2006. Low-frequency current variability observed at the shelfbreak in the northeastern Gulf of Mexico: May-October, 2004. Cont. Shelf Res. https://doi.org/10.1016/j.csr.2006.08.002.

Telszewski, M., Chazottes, A., Schuster, U., Watson, A.J., Moulin, C., Bakker, D.C.E., González-Dávila, M., Johannessen, T., Körtzinger, A., Lüger, H., Olsen, A., Omar, A., Padin, X.A., Ríos, A.F., Steinhoff, T., Santana-Casiano, M., Wallace, D.W.R., Wanninkhof, R., 2009. Estimating the monthly pCO2 distribution in the north Atlantic using a self-organizing neural network. Biogeosciences https://doi.org/10.5194/bg-6-1405-2009.

Ternon, J.F., Oudot, C., Dessier, A., Diverres, D., 2000. A seasonal tropical sink for atmospheric CO2 in the Atlantic ocean: the role of the Amazon River discharge. Mar. Chem. https://doi.org/10.1016/S0304-4203(99)00077-8.

Thomas, H., Bozec, Y., Elkalay, K., De Baar, H.J.W., 2004. Enhanced open ocean storage of CO2 from Shelf Sea pumping. Science (80-.). doi:https://doi.org/10.1126/science.1095491.

Walker, N.D., Wiseman, W.J., Rouse, L.J., Babin, A., 2005. Effects of river discharge, wind stress, and slope eddies on circulation and the satellite-observed structure of the Mississippi River plume. J. Coast. Res. https://doi.org/10.2112/04-0347.1.

Walsh, J.J., Jolliff, J.K., Darrow, B.P., Lenes, J.M., Milroy, S.P., Remsen, A., Dieterle, D.A., Carder, K.L., Chen, F.R., Vargo, G.A., Weisberg, R.H., Fanning, K.A., Muller-Karger, F.E., Shinn, E., Steidinger, K.A., Heil, C.A., Tomas, C.R., Prospero, J.S., Lee, T.N., Kirkpatrick, G.J., Whitledge, T.E., Stockwell, D.A., Villareal, T.A., Jochens, A.E., Bontempi, P.S., 2006. Red tides in the Gulf of Mexico: where, when, and why? J. Geophys. Res. Ocean. https://doi.org/10.1029/2004JC002813.

Walton, J.T., 2008. Subpixel urban land cover estimation: comparing cubist, random forests, and support vector regression. Photogramm. Eng. Remote Sensing https://doi.org/10.14358/PERS.74.10.1213.

Wawrik, B., Paul, J.H., Campbell, L., Griffin, D., Houchin, L., Fuentes-Ortega, A., Muller-Karger, F., 2003. Vertical structure of the phytoplankton community associated with a coastal plume in the Gulf of Mexico. Mar. Ecol. Prog. Ser. 251, 87–101. https://doi.org/10.3354/meps251087.

Weiss, R.F., 1974. Carbon dioxide in water and seawater: the solubility of a non-ideal gas. Mar. Chem. https://doi.org/10.1016/0304-4203(74)90015-2.

Wiseman, W.J., Rabalais, N.N., Turner, R.E., Dinnel, S.P., Macnaughton, A., 1997. Seasonal and interannual variability within the Louisiana coastal current: stratification and hypoxia. J. Mar. Syst. https://doi.org/10.1016/S0924-7963(96)00100-5.

Xue, Z., He, R., Fennel, K., Cai, W.J., Lohrenz, S., Huang, W.J., Tian, H., Ren, W., Zang, Z., 2016. Modeling pCO2 variability in the Gulf of Mexico. Biogeosciences 13, 4359–4377. https://doi.org/10.5194/bg-13-4359-2016.

Yan, F.P., Shangguan, W., Zhang, J., Hu, B.F., 2020. Depth-to-bedrock map of China at a spatial resolution of 100 meters. Sci. Data 7 (1), 1–13.

Yang, B., Byrne, H.B., 2015. Subannual variability of total alkalinity distributions in the northeastern Gulf of Mexico. J. Geophys. Res. Ocean. 120 (5), 3805–3816. https://doi.org/10.1002/2014JC010387.

Yuan, J., Miller, R.L., Powell, R.T., Dagg, M.J., 2004. Storm-induced injection of the Mississippi River plume into the open Gulf of Mexico. Geophys. Res. Lett. https://doi.org/10.1029/2003GL019335.

Zhu, Y., Shang, S., Zhai, W., Dai, M., 2009. Satellite-derived surface water pCO2 and air-sea CO2 fluxes in the northern South China Sea in summer. Prog. Nat. Sci. https://doi.org/10.1016/j.pnsc.2008.09.004.