Estimating the Carbon Content of Oceans using Satellite Sensor Data

Aadidev Sooknanan (aadidevsooknanan@gmail.com)
University of the West Indies
Patrick Hosein
University of the West Indies

Method Article

Keywords: Remote Sensing, Climate Change, Artificial Intelligence, Carbon Emissions, Ocean Acidification

Posted Date: February 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1374912/v1

Noname manuscript No.
(will be inserted by the editor)

Estimating the Carbon Content of Oceans using Satellite Sensor Data

Aadidev Sooknanan · Patrick Hosein

the date of receipt and acceptance should be inserted later

Abstract The impact of chemical processes in ocean surface waters is far-reaching. Recently, increased significance has been placed on the concentration of Carbon and its compounds and the effects these may have on climate change. Remote-sensing enables near real-time measurement of key sea-surface data which can be used to estimate Carbon levels. We illustrate with the use of hybrid Satellite sensor data. To validate our results we use data collected from cruise ships as the ground truth when training our algorithms. The error rate of our predictor is found to be small and hence the proposed approach can be used to estimate Carbon levels in any ocean. This work improves upon previous research in many ways including the use of sea water salinity as a proxy for Carbon estimates. Binary combinations of typically unary predictor attributes are used for the purposes of predicting the Carbon content of surface water and an inherently non-linear model is used to quantify the relationship.

Keywords Remote Sensing, Climate Change, Artificial Intelligence, Carbon Emissions, Ocean Acidification

1 Introduction

Climate change and ocean acidification are equally critical problems (Cooke and Kim (2019)). The Carbon cycle is important to climate, ecology and overall human livelihood (Liu and Xie (2017)). Moreover, CO_2 is the most significant anthropogenic (human-induced) source of Carbon driving climate change (Heinze et al. (2015)). Atmospheric CO_2 has oscillated between 200 and 280 *parts-per-million (ppm)* for the 400,000 years prior to industrialization. However, current levels now approach 300*ppm* due to mainly anthropogenic sources (Feely et al. (2004)).

Carbon sequestration is the transfer and secure storage of atmospheric CO_2) into other long-lived Carbon pools called *sinks*). The oceans play a significant role in regulating earth's systems (Lekshmi et al. (2019)), specifically in sequestration

Aadidev Sooknanan and Patrick Hosein
The University of the West Indies, St. Augustine, Trinidad
E-mail: aadidevsooknanan@gmail.com, patrick.hosein@sta.uwi.edu

of global carbon dioxide concentrations (Heinze et al. (2015)). The global ocean is the largest of the five global Carbon sinks and its level of Carbon uptake is increasing at a rate of 2.3 *peta-grams* (2.3×10^{15}) of Carbon per year. Regional fluctuations in CO_2 partial pressures can have potential consequences on global trends of carbon-related phenomenon (Jiahui et al. (2019)). The saturation of oceanic surface waters by Carbon Dioxide will result in a net decrease in the rate of carbonic uptake, and is estimated to contribute to a global temperature rise of 30% to 69% (Gattuso et al. (2015)).

The magnitude and rate of anthropogenic Carbon sequestered by the oceans exceeds the extent of variation due to natural sources for the past millennium (Imaoka et al. (2010)) which further indicates oceanic saturation. The longest record for in-situ Carbon measurements begun in 1960, and shows that current rate of increase is as much as 30 times faster than pre-industrial times (Doney et al. (2009)). Fossil fuels and cement production account for approximately 48% of the world's global Carbon emissions (Sabine et al. (2004)). In addition, deforestation, industrialization and land-use-changes have led to the unprecedented increase in Carbon emissions over the past 200 years (Guinotte and Fabry (2008)).

Ocean acidification is the change in ocean chemistry driven by the oceanic uptake of Carbon (Guinotte and Fabry (2008)). It is characterised by a series of chemical reactions initiated when CO_2 is absorbed by seawater. This CO_2 dissolves into Carbonate and Bicarbonate resulting in an increase in hydrogen ion concentration. According to Körtzinger (1999), ocean acidity has increased by 26% since the start of the industrial revolution. Additionally, the atmospheric-oceanic carbon concentration gradient (difference between the Carbon concentration in the atmosphere and in surface layers of the ocean) is likely to further affect climate changes and warming scenarios (Krishna et al. (2020)). Ocean acidification has significant negative impacts on fundamental bio-ecological ocean processes (Kroeker et al. (2010); Sabine et al. (2004)). More alarmingly, past extinction events have been linked to ocean acidification (Guinotte and Fabry (2008)), and the current rate of change in seawater chemistry is unprecedented (Sabine et al. (2004)).

The ability to quickly establish a baseline measurement for the Carbon content of large oceanic bodies is crucial in determining the levels of acidification in addition to assessing the rate of temperature rise in oceanic surface waters. However, there remains significant gaps in our ability to consistently and reliably estimate the carbon content of various sources and sinks (such as oceans)(Le Quéré et al. (2009)). Prediction and detection of carbon sinks are important issues with implications for all of human kind (Imaoka et al. (2010)).

Laboratory measurements are the gold standard for assessing the carbon content of seawater; however, research vessel time is costly and limited in coverage. Land et al. (2015). Remote sensing has the advantages of being fast, effective and near real-time (Paltsyn et al. (2019)). The recent development of wide-band satellite imaging sensors has resulted in large quantities of high-resolution imagery being available (Zhu et al. (2019)). Orbiting platforms additionally have the advantages of large observation range and high observation frequency (Yang (2019)) surpassing that of all alternate techniques (such as in-situ buoys, marine vessels, etc.).

Remote sensing is yet to be fully exploited and has significant potential in providing extensive global measurements. However, there are currently no orbiting platforms capable of directly measuring oceanic Carbon levels. According to

Tollefson (2016), multiple technical and political issues plague the development and launch of additional Carbon-monitoring instruments. This is evidenced by the multiple studies (Dugan et al. (2017); Johnson and Ohlhorst (1981); Ohlhorst (1978); Saba et al. (2018); Johnson (2013); Zui et al. (2016a); Hyde et al. (2011); Bosma et al. (2016)) outlining various approaches at developing proxies and underscoring challenges in oceanic measurements from space.

Surface-water Carbon Dioxide is influenced by thermodynamic and biological factors, and is adequately represented by the partial-pressure of Carbon dioxide pCO_2 at the surface. The most significant source is as a result of physical mixing processes caused by sea surface temperature (SST), sea surface salinity (SSS), chlorophyll-a (Chla), mixed layer depth (MLD), colored dissolved organic matter (CDOM), net primary productivity (NPP), photo-synthetically active radiation (PAR), wind speed and other factors. Although sea-surface measurements may not fully encompass biological processes, observations at the surface are relevant proxies for oceanic Carbon content, since the changes in carbonate chemistry due to atmospheric CO_2 occurs in the surface first (Land et al. (2015)). Thus, remote sensing derived data, via orbiting platforms hold great potential as a tool for monitoring changes in oceanic chemistry.

2 Related Work

In the past decade, a large number of new earth-observation orbiting platforms have been launched. As such, much effort has been placed on utilizing the unique, superior viewpoint of orbiting platforms for observation of both natural and human phenomena. In the following, we focus on works targeting large-scale oceanic measurements. In particular, we investigate works that do not derive oceanic carbon-related parameters (strictly) from first principles. Instead, we analyse works that employ data-driven approaches and statistical techniques to infer or derive patterns from observations.

To examine the oceanic carbon content, Poli et al. (2009) aimed to calculate the constants for the dissociation of carbonic acid. These constants define the tendency of a higher-order molecule (such as Carbon Dioxide) to decompose into its constituents (ions of Hydrogen and various carbonates) and are typically a function of pressure and temperature. The partial pressure of CO_2 in ocean surface waters was then determined from *Dissolved Inorganic Carbon* (DIC) and Total Alkalinity. These were used to validate previous measurements of the constants of dissociation. However, Poli et al. (2009) concluded that the optimal choice for these constants is subject to significant variability. Therefore usage of any one set of previously defined measurements for such derivations at a large scale.were not recommended. This further justifies our usage of a data-driven approach, as the physiochemical relationships vary depending on large numbers of inter-dependent parameters. We can instead measure statistical significance by means of inference, with a higher degree of confidence for a given region than the alternative first-principles derivation.

Bates et al. (2012) modelled seawater carbonate chemistry in the North Atlantic Ocean using in-situ measurements collected from Hydrostation S sites and a number of cruises. These measurements were collected at a minimum once per month since 1983 and were analysed for Dissolved Inorganic Carbon *(DIC)* by a

variety of methods. The resulting DIC values was then used to establish a time-series for the region. However, according to Bates et al. (2012), the data collected was heavily biased towards spring-time conditions owing to a sampling bias. This was evidenced by an apparent decrease in sea-surface temperature in in-situ measurements, which does not agree with other independent studies (Keil et al. (2020); Sutton et al. (2007)). From this work, it was made clear that in-situ measurements must account for at least an annual cycle to overcome seasonal sampling biases.

Zui et al. (2016b) compared models for Dissolved Inorganic Carbon (DIC) at the ocean surface using both satellite and *in-situ* data. Specifically, the Moderate Resolution Imaging Spectro-radiometer (MODIS) array of satellite sensors was used to establish a relationship with DIC measurements at two point locations over the course of 9 years. Zui et al. (2016b) did not derive a purely unknown relationship between oceanic parameters and DIC, but rather compared earlier models' performances with new data. In addition, their work regarded validation using in-situ measurements as the most precise method for confirming relationships found in their observations. This was a key theme in multiple published works, as the ability to directly derive sea-surface, chemical parameters via remote-sensing is not yet possible.

Dixit et al. (2019) analysed the partial pressure of Carbon Dioxide pCO_2 at the air-sea interface. A single autonomous system was deployed at 15°N 90°E in order to collect in-situ measurements of pCO_2. A linear, large-margin separation model was shown to more accurately estimate the relationship between SST and Salinity than a multiple-linear-regression (MLR) model. According to Krishna et al. (2020), the influence of Sea-Surface Salinity (SSS), Sea-Surface Temperature(SST) and Chl-a on pCO_2 varies depending on the domains of SST, SSS and Chl-a. Sabia et al. (2015) therefore justified development of a multi-parametric (bracketed) non-linear model. It was shown that DIC values can be parameterized by Chlorophyll-a, Sea-Surface Salinity and Sea-Surface Temperature. However, the measurement values for for Chlorophyll-A were not verified with in-situ data. Additionally, notwithstanding the use of multiple ranges for input attributes, accuracy was not comparable to previous approaches at regional models for CO_2 (Sabia et al. (2015)).

Fugacity is the surface signature of ocean acidity, dynamics, and bio-geochemistry (Liu W. Timothy (2014)). Liu W. Timothy (2014) developed a statistical model based on a linear Support-Vector Machine using Chlorophyll and Sea-Surface Temperature for predicting surface-level oceanic carbon content. Data was sourced from NASA's MODIS Satellites (Aqua and Terra) at a resolution of 0.25 °. Owing to the large data gaps on these platforms however, significant smoothing/averaging was done in order to extrapolate point measurements from the satellite data.

Similarly, Liu and Xie (2017) modelled Carbon Dioxide partial pressure at the ocean surface. However, according to the authors, their choice of model may have not fully captured the desired relationships owing to a relatively low temporal resolution (250000 over 8 years) and usage of a linear kernel. Additionally, large gaps existed in the satellite mapping which was interpolated heavily to accommodate for lower resolution images (as compared to what is currently available). Therefore their model usage for a specific region requires additional training using recent, region-specific in-situ measurements (Liu and Xie (2017)). Moreover, according to the authors, sufficient salinity measurements were not available at the time of writing (similar to Liu W. Timothy (2014)) at sufficient spatio-temporal

resolutions. This resulted in a lack of trend in model prediction, where predicted levels of Carbon Dioxide at the ocean surface did not follow the general increase as evidenced by in-situ measurements.

Following from the aforementioned, this work seeks to utilize recent advances in remote-sensing and integrated data sources to provide a Carbon model for the Caribbean region. This model exploits the relationship (Takahashi et al. (2002)) between surface-level Carbon content and ocean-surface temperature. Additionally, this work improves upon previous approaches (such as Liu W. Timothy (2014); Liu W. Timothy (2014)) in the use of sea-surface-salinity data and usage of in-situ data of significantly higher spatial density. Finally, this work utilizes an inherently non-linear method to robustly model the aforementioned relationships.

3 Description of Datasets

3.1 HYCOM Data

HYCOM is a data-assimilative hybrid isopycnal-sigma-pressure model and part of a multi-institutional effort sponsored by the National Ocean Partnership Program (NOPP) as part of the US Global Ocean Data Assimilation Experiment (GODAE). Within non-extreme latitudes (non-polar regions between 80.48°S and 80.48°S), HYCOM data are available on a standardized grid.

Data is assimilated from a combination of remote-sensing platforms (GFO Barry et al. (1995), ENVISAT Louet and Bruzzi (1999) and Jason-1 Ménard et al. (2003)) which provide information on space-time variability of surface-wind stress, temperature and specific humidity. Vertical profiles from eXpendable BathyThermographs, Argo floats and Conductivity-Temperature-Depth sensors enhance subsurface variability mapping. However, these profiles are typically too sparse to be used by themselves Chassignet et al. (2007).

The resolution of HYCOM data are 0.08 *arc-degrees*(approximately $9km$ square). The main advantage of the HYCOM data is its indexing of ocean-surface parameters by means of z-coordinates. z-coordinates index ocean depth at standard levels, and allows for a smooth transition from upper-ocean to deep-ocean layers. This results in an ease of comparative computations at differing sea-depths. Additionally, there is no current single orbiting platform with the necessary spatio-temporal density for inference of parameters which define latent relationships between observed variables.

HYCOM data contain a the following combination of oceanic parameters indexed by z-coordinates indexed by latitude, longitude, reference date/time and depth:

- Downward Surface Flux (heat)
- Water Flux into the Ocean
- Surface Temperature Trends
- Surface Salinity Trends
- Ocean Mixed-layer Thickness
- Sea Water Salinity
- Sea Water Velocity
- Sea Water Temperature

Table 1 Sample from HYCOM Data

lat	lon	t_0	s_0	t_2	s_2
18.520	-68.327	26.590	35.784	26.476	35.788
25.148	-77.325	25.252	36.602	25.071	36.608
24.064	-73.898	26.279	36.585	26.183	36.586
22.559	-71.743	26.549	36.584	26.405	36.585

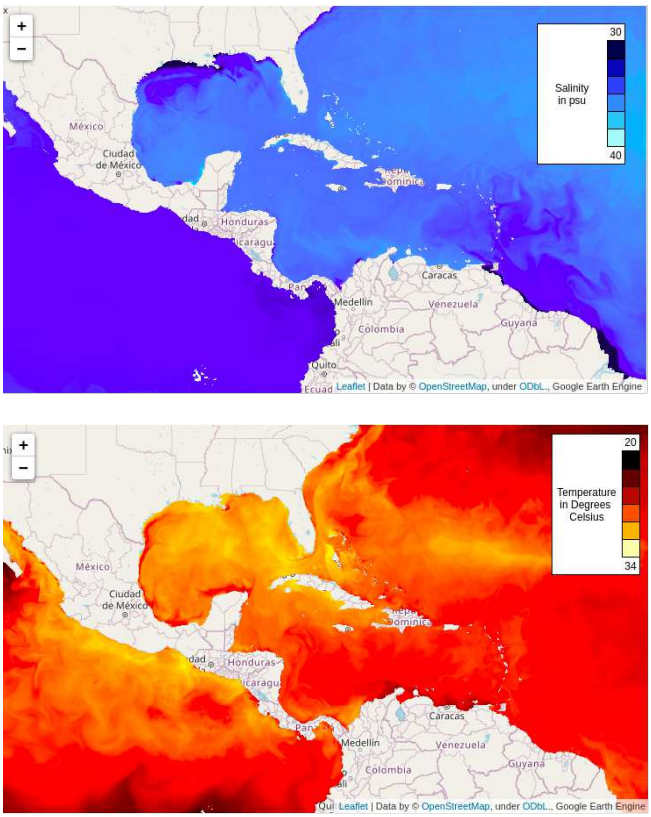


Fig. 1 Mean Salinity (top) and Temperature (bottom) at 0m for 2019-2020

For this work, the z-coordinate (depth), location (latitude and longitude), sea water salinity (s_i) and sea water temperature (t_i) are used where the subscript i indicates depth in meters. Table 1 shows latitude, longitude, temperature (C) and salinity (psu) measurements from the HYCOM data for the uppermost z-levels (0 m and 2 m respectively). In Figure 1 we provide the mean salinity and mean temperature for the Caribbean region at a depth of 0 m.

3.2 NOAA Labelled Data

The Ocean Chemistry and Ecosystems Division (OCED) of The National Oceanic and Atmospheric Administration (NOAA) focuses understanding the ocean's role within the context of the global environment. Automated systems for pCO_2 measurement were installed on cruise ships of Royal Caribbean International Cruises and subsidiaries Wanninkhof et al. (2020) by the NOAA. This system provides measurements of multiple ocean-water parameters, beginning in 2002 and continuing to the present. The instruments consist of equilibrators, a condenser, water flow meter, drying tubes and additional equipment for analysing the output of the equilibrator Pierrot et al. (2009). For this work, data from the *Allure of the Seas* was used for the period 2019-2020 within the Caribbean Region. This data covered the range of latitudes between $11.677°N$ and $26.817°N$ and longitudes $-59.638°$ and $-87.675°$. This resulted in approximately one-million data-tuples indexed spatially, containing the following attributes:

- Mole fraction of CO_2 in the equilibrator headspace (dry) at equilibrator temperature
- Mole fraction of CO_2 measured in dry outside air
- Mole fraction of CO_2 in outside air associated with each water analysis ppm
- Barometric pressure in the equilibrator headspace
- Barometric pressure measured outside
- Water temperature in equilibrator
- Sea surface temperature
- Sea surface salinity
- Fugacity of CO_2 in Sea Water
- Fugacity of CO_2 in Air
- Quality control flags for equilibrator functions

The mean values for the above parameters was found such that the spatial resolution matched HYCOM grids. Two flags in this data indicated as to the nominal operation of the equilibrator. Values out of range (negative) and other anomalous instances were removed by filtering against these quality flags. The latitude-longitude positions were then used to cross-reference the HYCOM data, resulting in attributes above (specifically Fugacity) being indexed by spatial(latitude/longitude) and z(depth) coordinates.

A sample of equilibrator attributes *eq*, atmospheric attributes *atm* and sea attributes (*sea*) are given in Table 2. All pressures p are measured in standard hPa, temperatures t are given in degrees Celsius and Carbon-dioxide/Fugacity f are given in units of micro-atmospheres.

3.3 Carbon Fugacity

In order to measure the air-sea exchange of gases, the partial pressure of gas in the ocean surface is first determined. The concentration of gas in the equilibrator headspace (gaseous) is directly proportional to the concentration in the equivalent volume of seawater (liquid), parameterized by temperature and salinity Johnson (1999), $C_w = \alpha C_e$.

Table 2 Sample from NOAA Data

lat	lon	$CO_{2\mathsf{eq}}$	$CO_{2\mathsf{atm}}$	p_{eq}	p_{atm}	t_{eq}	t_{sea}	sal_{sea}	f_{sea}	f_{atm}	$f_{\mathsf{atm}} - f_{\mathsf{sea}}$
24.4117	-81.8218	391.10	420.16	1025.0	1020.16	25.49	25.4232	36.3921	381.00	408.52	-27.53
24.4088	-81.8188	391.91	420.16	1025.0	1019.86	25.52	25.4422	36.4019	381.58	408.38	-26.80
24.4092	-81.8138	392.60	420.15	1025.0	1019.66	25.53	25.4500	36.3980	382.22	408.28	-26.06
24.4125	-81.8090	392.34	420.15	1025.0	1020.16	25.53	25.4539	36.3911	382.03	408.49	-26.46

The instrumentation installed in the ships of Royal Caribbean measures the mole fraction of CO_2 in dry air, which is converted to fugacity by correcting for the non-ideal gas and the water vapour level Jang et al. (2017). An alternative approach, in determining air-sea flux, is given by the difference in partial pressures between the air and sea CO_2

$$F = k_g(pCO_{2,\mathsf{a}} - pCO_{2,\mathsf{s}}) \tag{1}$$

However, the wind-speed dependent gas exchange coefficient k_g is not precisely known, and there exists a discrepancy in the global mean values as obtained by two different methods (Siegenthaler and Sarmiento (1993)).

4 Methods

4.1 Supervised Learning

Machine learning is the field of Artificial Intelligence whereby statistical rules are derived from data. A supervised learning algorithm models the implicit relationship existing in labelled data by means of a set of equations. In other terms, given a set of labelled data D, a supervised learning algorithm aims to learn the relationship between input attributes X and an output attribute y in order to predict the output $\hat{y}$ given previously unseen X. Supervised learning algorithms receive feedback from a *loss function*, which quantitatively informs how closely the model matches the relationships within the data (Saravanan and Sujatha (2018)).

In this work, the output attribute y is represented by the fugacity of the partial-pressure of Carbon Dioxide, dependent on the input attributes as determined by a feature-selection process. This is modelled using a gradient-boosting regression tree, which provides a continuous-valued output as a nonlinear function of its input attributes.

4.2 Decision-Tree Regression

The supervised gradient-boosting regression tree (GBRT) iteratively defines a series of mappings from a labelled training dataset and progressively refines itself by means of an explicit loss function (Mitchell et al. (2018)). The GBRT learns

an input-output relation by means of a series of conditional/thresholding operations. It is able to break down a complex decision-making process into a collection of simpler decisions thus providing a solution which is often easier to interpret (Safavian and Landgrebe (1991)). In other terms, at each stage in the regression tree, the output domain is refined (subsetted) by means of an information criteria. The stochastic Gradient Boosted Regression Tree is adaptable, easy to interpret and produces highly accurate models (Wen et al. (2018)). The Stochastic GBRT starts with a single decision-tree, and iteratively appends new trees based on their performance with respect to some objective function. The heuristic used to define branching decisions generates splits based on the distribution of input attributes, and greedily selects new decision trees f based on an objective function $\mathcal{Z}$:

$$\mathcal{Z} = \sum_i l(y_i, \hat{y}_i^{t-1} + f_t(\mathbf{x}_i) + \sum_k \Omega(f_k) \tag{2}$$

A regularization term Ω is included to prevent over-fitting (Ye et al. (2009)) by means of penalizing tree coefficients. In order to offset the overhead obtained by Gradient Boosted Decision Tree (GBDT) on large datasets, the NVIDIA RAPIDS (Hricik et al. (2020)) and the XGBoost (Chen and Guestrin (2016)) libraries were implemented on a GPU.

4.3 Loss Function

A loss function is used to quantify the performance of a given learning algorithm. Many loss functions have been proposed for supervised learning, however the Mean-Absolute Error and Root Mean-Squared Error have become popular the field of geosciences (Chai and Draxler (2014)). The Mean Absolute Error (MAE) is given by the following:

$$\mathsf{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

An MAE assigns equal weighting to all errors (known as l_1 optimization). If we denote ground-truth values as y and predicted values as $\hat{y}$, the loss value is directly proportional to the magnitude of divergence in the predicted value $\hat{y}$ from the ground truth label y.

The Root-Mean-Squared-Error (RMSE) attempts to penalize variance in output predictions by squaring the error term from MAE (also known as l_2 optimization):

$$\mathsf{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

This results in unequal weighting of error terms, where larger error values are more heavily penalized than smaller values.

While the MAE and RMSE have been used as standard metrics for model performance for many years, there is no consensus on the most appropriate metric for model errors (Chai and Draxler (2014)). Additionally, measures based on the sum-of-squared calculation do not describe average error alone. The distribution of

error magnitudes become more variable in a non-monotonic fashion with increasing error (Chai and Draxler (2014)).

As a result, the Huber loss is used to optimize the regression tree. The Huber loss behaves quadratically for small residual errors and linearly for large residual errors (Huber (1992)). The Huber loss is given by the following:

$$\mathcal{L}(y, \hat{y}) = \begin{cases} (y - \hat{y})^2 & \forall \ |y - \hat{y}| \leq \alpha \\ |y - \hat{y}| & \forall \ |y - \hat{y}| > \alpha \end{cases} \tag{3}$$

This is the equivalent to minimizing the Kullback-Leibler divergence (Meyer (2021)) and is hence used in robust regression to take advantage of the desirable properties of both l_1 and l_2 penalties. The Huber Loss is regulated by the hyper-parameter $\alpha > 0$. For absolute values smaller than α the corresponding distribution resembles the normal distribution while for values outside this range it resembles the Laplace distribution. This is the equivalent of a data-defined transition from a quadratic to absolute-valued function (Meyer (2021)). This affords the Huber loss a significant advantage in resisting outliers (common in remote sensing data) when compared to MAE and RMSE.

This loss function is used to guide and actively correct the learning of the stochastic GBRT during training, and additionally used to validate the stochastic GBRT during testing. A loss function is used to concisely quantify the performance of a machine-learning model, and its derivative is used to quantify the magnitude of correction required for model parameters. Following training, the model is evaluated using the loss function and updated based on the magnitude and size of the errors made. This process is complete when the model's increase in performance (or change in parameters) do not exceed some pre-defined threshold.

5 Numerical Results

5.1 Data Selection

Water temperature and Sea-Surface-Salinity at depths of 0m, 2m, 4m, 6m and 8m from the HYCOM platform was used for regression (Seabold and Perktold (2010)). The target data are Fugacity values (fCO_2) for ocean surface water from the NOAA data. Binary interaction attributes were derived from the single attribute values by means of multiplicative combination, for a total of 32+5 predictor variables.

In order to determine an optimal set of predictor attributes, statistical p-value testing was used to determine the significance of relationship between predictor attributes and the target data. (A lower p-value indicates a decreased probability of the observed relation occurring by pure chance). Therefore the chances of disproving a non-trivial relationship between predictor and target attributes are directly proportional to p-value. The target size $\alpha = 0.05$ was used as a threshold to filter non-predictive attributes in both unary and binary cases. Stage-wise variable selection was used by means of the method of Ordinary Least Squares Regression resulting in the predictor attributes in Figures 3 and 4.

Table 3 Unary Predictors

Attribute	p-value
salinity_0	0.000110
salinity_2	0.001892
temp_8	0.009946

Table 4 Binary Predictors

Attribute I	Attribute II	p-value
temp_6	temp_8	0.000418
salinity_2	temp_6	0.002439
temp_4	temp_8	0.002573
salinity_2	temp_8	0.002829
salinity_0	temp_8	0.005784
salinity_0	temp_6	0.006055
salinity_4	temp_6	0.011255
temp_4	salinity_4	0.015059
salinity_2	temp_4	0.019423
temp_4	salinity_6	0.026789
temp_6	salinity_6	0.027088
temp_2	temp_8	0.028429
temp_4	temp_6	0.035664
salinity_0	temp_4	0.040986

Table 5 GB-Tree Hyper Parameters

Hyper-parameter	Value
Maximum Tree Depth	7
Minimum Child Weight	5.0
Subsample	0.8
Learning Rate	0.01

5.2 Gradient-Boosting Regression Tree

The GBRT was implemented using the XGBoost library and NVIDIA RAPIDS on a GTX 1060M GPU. Figure **??** shows the test scores for single-parameters using the F-test. Enumerative grid-search was used to tune parameters for the gradient boosted tree. After 1000 iterations for each combination in the parameter subspace the optimal values listed in Figure 5 were obtained.

Fig. 2 Omnibus Test Values for Single Parameters

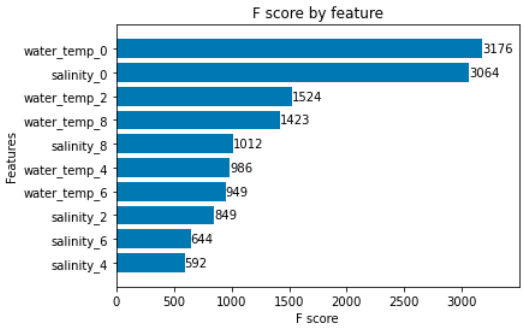


Fig. 3 Fugacity of CO2 in Sea Water Against Temperature at 0m

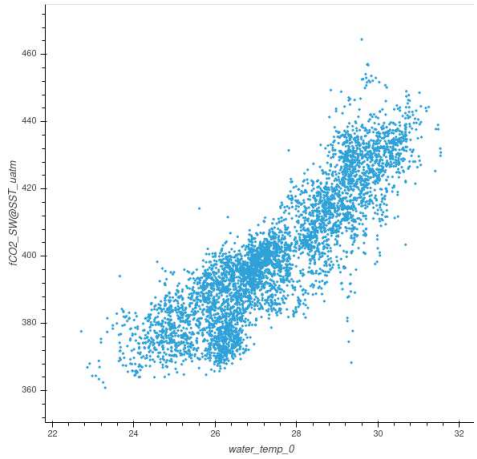


Fig. 4 Fugacity of CO2 in Sea Water Against Temperature at 0m

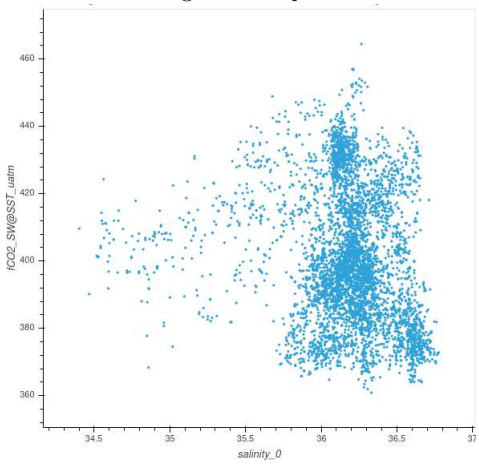
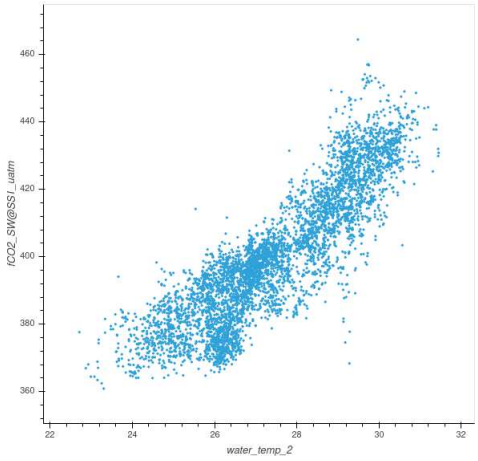


Fig. 5 Fugacity of CO2 in Sea Water Against Temperature at 2m



5.3 Model Validation

Ground-truth in-situ measurements were used to validate our model. This target data consisted fCO_2 in *units of micro-atmospheres* (μatm). The domain of values for fCO_2 as measured by the equilibrator (for all valid measurements as indicated by the use of the quality flag) was [397.68, 492.46] with a mean of 400.73 and a standard-deviation of 19.78. Roughly one-third of the training data was reserved for model validation. The model was trained using cross-validation and evaluated using the Huber Loss function. Using the selected attributes and optimal GBRT, the Huber Loss for fCO_2 was found to be 3.98, or approximately 1% of the mean value for the target attribute.

As shown in Figure 6 (top), usage of binary interaction variables significantly improved the residual distribution of our model prediction. As shown in Figure 6 (bottom), the residual (error) terms are zero-centred with little spread. This is desirable, particularly when compared against the top plot. This inconsistent spread of high variance (an undesirable property) leads to inconsistent model behaviour. In contrast the symmetric, narrow residual distribution for our model with interaction variables is indicative of consistent model performance (low variance).

6 Discussion

6.1 Scientific Implications

Traditional remote-sensing (satellite-based) approaches do not measure Ocean Carbon levels directly. In this work, we investigate and exploit the impact sea-surface salinity and sea-surface temperature on the fugacity of Carbon-Dioxide, $fC0_2$, at the surface layer of the ocean. This work confirms that SST at $0m$ is the primary unary predictor on $fC0_2$ (Liu and Xie (2017);Takahashi et al. (2002)).

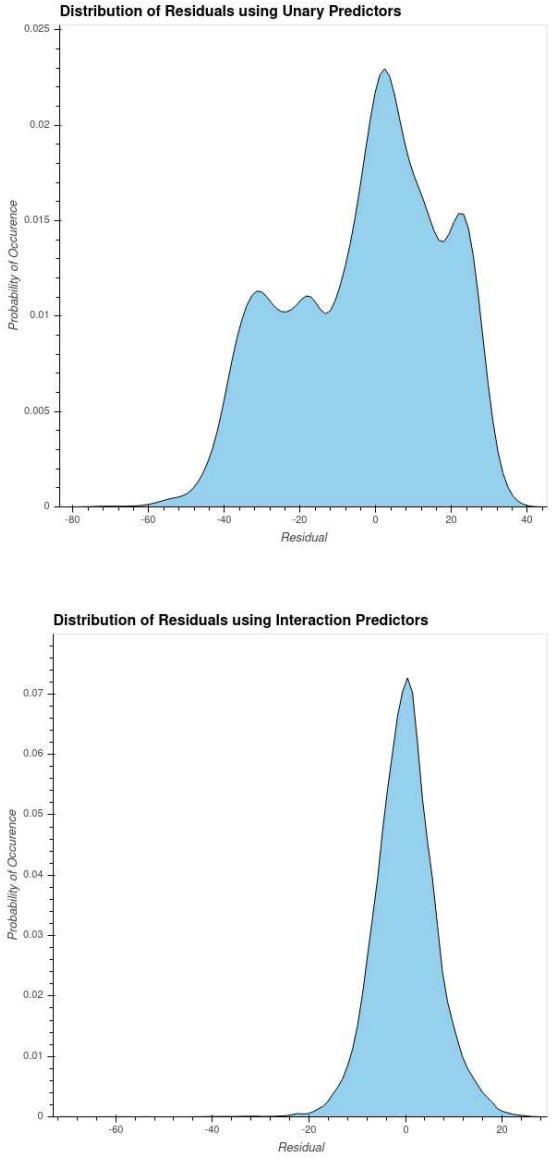


Fig. 6 Distribution of Residuals for Model Using Unary Predictors Only (top) and Interaction Variables (bottom)

This work does not rely on multiple stationary in-surface installations for the purpose of making predictions. Therefore, the model developed in this work can be applied to sea-surface areas where in-situ data is not currently available. Moreover, the periodic nature of satellite observations enables our model to be the basis for spatial and temporal analysis. Similarly, our work may be used for the discovery and quantification of both Carbon sources and sinks in the open ocean. In this way, the model developed may be used to find novel Carbon sinks in the open ocean,

and quantify the rate of sequestration over time, without the need for earth-based, surface-level measurements.

The increased longitudinal spread of the in-situ data used greatly decreases the influence of coastal anomalies on our model's derivation. This work describes the methods by which remote sensing data can be used to indirectly estimate surface oceanic carbon content. Moreover, the ability of remotely-sensed data sources to operate in times of anomalous weather conditions on a regular, periodic basis establishes a non-trivial advantage when compared to earth-borne surface methods.

As of writing, the relationship between SST, SSS and fCO_2 are not unanimously agreed upon. In order to account for annual weather events, this data period captured an entire annual cycle, thereby controlling against seasonal oceanic events. This enables our model to be used year-round in non-extreme latitudes, notwithstanding the non-availability of in-situ cruise ship measurements.

The low variance and error terms observed from our model is validated using in-situ surface data. The validation described in Section 5.3 above shows accuracy greater than that which was achieved by a similar global model Liu W. Timothy (2014) and Krishna et al. (2020). Our work improves upon both of these approaches by incorporating SSS for all data. In addition, this work improves upon previous approaches by consideration of binary factors of influence (see Table 4) and application of an inherently non-linear model. In further confirmation of our model, our model was run on previously unseen data following initial cross-validation, with the same results as initially found.

6.2 Limitations and Future Work

The target area for this work was the Caribbean region and, as a result, is limited to non-extreme (non-polar latitudes). Additionally, the large-scale, inter-sea mixing of waters is left for future investigation.

Several studies have investigated the relationship between Chlorophyll-A surface concentrations and its impact on surface-level Carbon content. Moreover, the impact of wind speed and direction has been shown to influence CO_2 flux at the sea-air boundary. Notwithstanding, the relationship between SST and SSS holds as validated in 5.3. However, this relationship between surface winds and fCO_2 at the ocean surface may explain outliers observed during the validation process (Liu W. Timothy (2014)).

The Orbiting Carbon Observatory-2 (OCO-2) and OCO-3 are NASA's first Earth-orbiting satellites dedicated to measuring atmospheric Carbon Dioxide. At the time of writing, these instruments are dedicated strictly to observing near-infrared CO_2 and A-band molecular oxygen. However, future orbiting platforms may investigate the relationships between surface-level atmospheric concentration and the rate of transfer between air and ocean (Liu and Xie (2017)).

7 Conclusion

In this paper, a non-linear model is implemented for satellite data and validated using in-situ measurements in the Caribbean region. This model exploited the relationship between temperature, salinity and sea-surface Carbon concentration

to remotely predict surface-water Carbon content. We found that limiting scope of observations to non-extreme latitudes yielded lower losses when compared to global approaches.

We also found that the usage of binary predictor attributes significantly reduced prediction errors when compared to previous approaches. Additionally, the usage of an inherently non-linear model with the Huber loss-function improved upon previous approaches that used linear penalised models. Finally, our results were validated using a rich source of in-situ measurements made available via the NOAA. Future work may seek to incorporate Chl-a, surface wind-velocity as well as other anthropogenic factors into prediction models.

8 Declarations

8.1 Funding

Not applicable

8.2 Ethics Approval

Not applicable

8.3 Consent for Publication

Not applicable (all data is freely available from public online sources)

8.4 Authors' contributions

This publication was written by Aadidev Sooknanan and writtenedited by Patrick Hosein. All authors reviewed the manuscript.

8.5 Competing / Conflicts of Interest

Not applicable

8.6 Availability of data and material

Data used for this work is openly available via Google Earth Engine and the NOAA. Earth Engine Satellite Data was specifically published according to Cummings and Smedstad (2013), and in-situ data is available via the NOAA's website. Standard software packages, namely *NVIDIA RAPIDS*, *Scikit-Learn* and *XGBoost* was used in this work. Additionally, a code-repository of Python code is available at https://github.com/aadi350/water-salinity.

8.7 Acknowledgements

References

Barry R, Finkelstein J, Kilgus C, Mooers CNK, Needham B, Crawford M (1995) Geosat follow-on satellite to supply ocean sciences data. Eos, Transactions American Geophysical Union 76(4):33–36, DOI https://doi.org/10.1029/EO076i004p00033-02, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/EO076i004p00033-02, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/EO076i004p00033-02

Bates N, Best M, Neely K, Garley R, Dickson A, Johnson R (2012) Detecting anthropogenic carbon dioxide uptake and ocean acidification in the north atlantic ocean. Biogeosciences 9(7):2509–2522

Bosma J, Izett R, Izett R (2016) Challenges with collecting data for measured ph and dissolved inorganic carbon (dic) in coastal waters. In: OCEANS 2016 MTS/IEEE Monterey, pp 1–5, DOI 10.1109/OCEANS.2016.7761421

Chai T, Draxler RR (2014) Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. Geoscientific model development 7(3):1247–1250

Chassignet EP, Hurlburt HE, Smedstad OM, Halliwell GR, Hogan PJ, Wallcraft AJ, Baraille R, Bleck R (2007) The hycom (hybrid coordinate ocean model) data assimilative system. Journal of Marine Systems 65(1-4):60–83

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

Cooke SL, Kim SC (2019) Exploring the "evil twin of global warming": public understanding of ocean acidification in the united states. Science Communication 41(1):66–89

Cummings JA, Smedstad OM (2013) Variational data assimilation for the global ocean. In: Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II), Springer, pp 303–343

Dixit A, Lekshmi K, Bharti R, Mahanta C (2019) Net sea–air co _{2} fluxes and modeled partial pressure of co _{2} in open ocean of bay of bengal. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12(7):2462–2469

Doney SC, Balch WM, Fabry VJ, Feely RA (2009) Ocean acidification: a critical emerging problem for the ocean sciences. Oceanography 22(4):16–25

Dugan D, Janzen C, McCammon M, Evans W, Bidlack A (2017) The evolution of ocean acidification observing efforts in alaska and the development of an alaska ocean acidification network. In: OCEANS 2017-Anchorage, IEEE, pp 1–6

Feely RA, Sabine CL, Lee K, Berelson W, Kleypas J, Fabry VJ, Millero FJ (2004) Impact of anthropogenic co2 on the caco3 system in the oceans. Science 305(5682):362–366

Gattuso JP, Magnan A, Billé R, Cheung WW, Howes EL, Joos F, Allemand D, Bopp L, Cooley SR, Eakin CM, et al. (2015) Contrasting futures for ocean and society from different anthropogenic co2 emissions scenarios. Science 349(6243)

Guinotte JM, Fabry VJ (2008) Ocean acidification and its potential effects on marine ecosystems. Annals of the New York Academy of Sciences 1134(1):320–

342
Heinze C, Meyer S, Goris N, Anderson L, Steinfeldt R, Chang N, Le Quéré C, Bakker DC (2015) The ocean carbon sink–impacts, vulnerabilities and challenges. Earth System Dynamics 6(1):327–358
Hricik T, Bader D, Green O (2020) Using rapids ai to accelerate graph data science workflows. In: 2020 IEEE High Performance Extreme Computing Conference (HPEC), IEEE, pp 1–4
Huber PJ (1992) Robust estimation of a location parameter. In: Breakthroughs in statistics, Springer, pp 492–518
Hyde A, Vandemark D, Shellito S, Salisbury J, Irish J, DeGrandpre M (2011) A multiyear assessment of biological pertubations of co¡inf¿2¡/inf¿ in the northeast channel of the gulf of maine. In: OCEANS'11 MTS/IEEE KONA, pp 1–5, DOI 10.23919/OCEANS.2011.6107017
Imaoka K, Kachi M, Fujii H, Murakami H, Hori M, Ono A, Igarashi T, Nakagawa K, Oki T, Honda Y, et al. (2010) Global change observation mission (gcom) for monitoring carbon, water cycles, and climate change. Proceedings of the IEEE 98(5):717–734
Jang E, Im J, Park GH, Park YG (2017) Estimation of fugacity of carbon dioxide in the east sea using in situ measurements and geostationary ocean color imager satellite data. Remote Sensing 9(8), DOI 10.3390/rs9080821, URL https://www.mdpi.com/2072-4292/9/8/821
Jiahui W, Liang L, Han L, Chunyang C, Ting H, Di G (2019) Interpretation of the report on temporal dynamics and spatial distribution of global carbon source and sink. In: 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), pp 1–4, DOI 10.1109/Agro-Geoinformatics.2019.8820487
Johnson JE (1999) Evaluation of a seawater equilibrator for shipboard analysis of dissolved oceanic trace gases. Analytica chimica acta 395(1-2):119–132
Johnson KS (2013) Bioargo: A global scale chemical sensor network to observe carbon, oxygen, and nitrogen cycles in the ocean. In: SENSORS, 2013 IEEE, pp 1–1, DOI 10.1109/ICSENS.2013.6688480
Johnson RW, Ohlhorst CW (1981) Application of remote sensing to monitoring and studying dispersion in ocean dumping. In: Ocean Dumping of Industrial Wastes, Springer, pp 175–191
Keil P, Mauritsen T, Jungclaus J, Hedemann C, Olonscheck D, Ghosh R (2020) Multiple drivers of the north atlantic warming hole. Nature Climate Change 10(7):667–671
Körtzinger A (1999) Determination of carbon dioxide partial pressure (p (co2)). Methods of Seawater Analysis pp 149–158
Krishna KV, Shanmugam P, Nagamani PV (2020) A multiparametric nonlinear regression approach for the estimation of global surface ocean pco2 using satellite oceanographic data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13:6220–6235, DOI 10.1109/JSTARS.2020.3026363
Kroeker KJ, Kordas RL, Crim RN, Singh GG (2010) Meta-analysis reveals negative yet variable effects of ocean acidification on marine organisms. Ecology letters 13(11):1419–1434
Land PE, Shutler JD, Findlay HS, Girard-Ardhuin F, Sabia R, Reul N, Piolle JF, Chapron B, Quilfen Y, Salisbury J, et al. (2015) Salinity from space unlocks satellite-based assessment of ocean acidification

570 Le Quéré C, Raupach MR, Canadell JG, Marland G, Bopp L, Ciais P, Conway
571 TJ, Doney SC, Feely RA, Foster P, et al. (2009) Trends in the sources and sinks
572 of carbon dioxide. Nature geoscience 2(12):831–836
573 Lekshmi K, Bharti R, Mahanta C (2019) Spatio-temporal distribution of carbon
574 dioxide partial pressure in the bay of bengal. In: IGARSS 2019-2019 IEEE In-
575 ternational Geoscience and Remote Sensing Symposium, IEEE, pp 8165–8168
576 Liu WT, Xie X (2017) Space observation of carbon dioxide partial pressure at
577 ocean surface. IEEE Journal of Selected Topics in Applied Earth Observations
578 and Remote Sensing 10(12):5472–5484
579 Liu W Timothy XX (2014) Ocean surface carbon dioxide fugacity observed from
580 space. Report, National Aeronautics and Space Administration, USA, URL
581 https://archimer.ifremer.fr/doc/00651/76344/
582 Louet J, Bruzzi S (1999) Envisat mission and system. In: IEEE 1999 Inter-
583 national Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.
584 99CH36293), IEEE, vol 3, pp 1680–1682
585 Ménard Y, Fu LL, Escudier P, Parisot F, Perbos J, Vincent P, Desai S, Haines
586 B, Kunstmann G (2003) The jason-1 mission special issue: Jason-1 calibra-
587 tion/validation. Marine Geodesy 26(3-4):131–146
588 Meyer GP (2021) An alternative probabilistic interpretation of the huber loss. In:
589 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
590 Recognition, pp 5261–5269
591 Mitchell R, Adinets A, Rao T, Frank E (2018) Xgboost: Scalable gpu accelerated
592 learning. arXiv preprint arXiv:180611248
593 Ohlhorst CW (1978) Quantitative mapping by remote sensing of an ocean acid-
594 waste dump, vol 1275. National Aeronautics and Space Administration
595 Paltsyn MY, Gibbs JP, Mountrakis G (2019) Integrating traditional ecological
596 knowledge and remote sensing for monitoring rangeland dynamics in the altai
597 mountain region. Environmental management 64(1):40–51
598 Pierrot D, Neill C, Sullivan K, Castle R, Wanninkhof R, Lüger H, Johannessen T,
599 Olsen A, Feely RA, Cosca CE (2009) Recommendations for autonomous under-
600 way pco2 measuring systems and data-reduction routines. Deep Sea Research
601 Part II: Topical Studies in Oceanography 56(8-10):512–522
602 Poli S, Franzolin E, Fumagalli P, Crottini A (2009) The transport of carbon and
603 hydrogen in subducted oceanic crust: An experimental study to 5 gpa. Earth
604 and Planetary Science Letters 278(3-4):350–360
605 Saba GK, Wright-Fairbanks E, Miles TN, Chen B, Cai WJ, Wang K, Barnard
606 AH, Branham CW, Jones CP (2018) Developing a profiling glider ph sensor
607 for high resolution coastal ocean acidification monitoring. In: OCEANS 2018
608 MTS/IEEE Charleston, IEEE, pp 1–8
609 Sabia R, Fernández-Prieto D, Shutler J, Donlon C, Land P, Reul N (2015) Re-
610 mote sensing of surface ocean ph exploiting sea surface salinity satellite observa-
611 tions. In: 2015 IEEE International Geoscience and Remote Sensing Symposium
612 (IGARSS), pp 106–109, DOI 10.1109/IGARSS.2015.7325709
613 Sabine CL, Feely RA, Gruber N, Key RM, Lee K, Bullister JL, Wanninkhof R,
614 Wong C, Wallace DW, Tilbrook B, et al. (2004) The oceanic sink for anthro-
615 pogenic co2. science 305(5682):367–371
616 Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology.
617 IEEE transactions on systems, man, and cybernetics 21(3):660–674

618 Saravanan R, Sujatha P (2018) A state of art techniques on machine learning al-
619 gorithms: a perspective of supervised learning approaches in data classification.
620 In: 2018 Second International Conference on Intelligent Computing and Control
621 Systems (ICICCS), IEEE, pp 945–949
622 Seabold S, Perktold J (2010) Statsmodels: Econometric and statistical modeling
623 with python. In: Proceedings of the 9th Python in Science Conference, Austin,
624 TX, vol 57, p 61
625 Siegenthaler U, Sarmiento JL (1993) Atmospheric carbon dioxide and the ocean.
626 Nature 365(6442):119–125
627 Sutton RT, Dong B, Gregory JM (2007) Land/sea warming ratio in response
628 to climate change: Ipcc ar4 model results and comparison with observations.
629 Geophysical research letters 34(2)
630 Takahashi T, Sutherland SC, Sweeney C, Poisson A, Metzl N, Tilbrook B, Bates
631 N, Wanninkhof R, Feely RA, Sabine C, et al. (2002) Global sea–air co2 flux
632 based on climatological surface ocean pco2, and seasonal biological and tem-
633 perature effects. Deep Sea Research Part II: Topical Studies in Oceanography
634 49(9-10):1601–1622
635 Tollefson J (2016) Carbon-sensing satellite system faces high hurdles: space agen-
636 cies plan an advanced fleet, but technical and political challenges abound. Nature
637 533(7604):446–448
638 Wanninkhof R, Pierrot D, Sullivan K, Barbero L, Triñanes J (2020) A 17-year
639 dataset of surface water fugacity of co 2 along with calculated ph, aragonite
640 saturation state and air–sea co 2 fluxes in the northern caribbean sea. Earth
641 System Science Data 12(3):1489–1509
642 Wen Z, He B, Kotagiri R, Lu S, Shi J (2018) Efficient gradient boosted decision
643 tree training on gpus. In: 2018 IEEE International Parallel and Distributed
644 Processing Symposium (IPDPS), IEEE, pp 234–243
645 Yang K (2019) Progress, challenge and prospect for remote sensing monitoring of
646 flood and drought disasters in china. In: IGARSS 2019-2019 IEEE International
647 Geoscience and Remote Sensing Symposium, IEEE, pp 4280–4283
648 Ye J, Chow JH, Chen J, Zheng Z (2009) Stochastic gradient boosted distributed
649 decision trees. In: Proceedings of the 18th ACM conference on Information and
650 knowledge management, pp 2061–2064
651 Zhu Q, Sun X, Zhong Y, Zhang L (2019) High-resolution remote sensing image
652 scene understanding: A review. In: IGARSS 2019-2019 IEEE International Geo-
653 science and Remote Sensing Symposium, IEEE, pp 3061–3064
654 Zui T, Tingting L, Xiang Z, Sheng M, Xiangbing K (2016a) Monitoring of sinking
655 flux of ocean particulate organic carbon using remote sensing methods. In: 2016
656 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp
657 3788–3791, DOI 10.1109/IGARSS.2016.7729982
658 Zui T, Tingting L, Xiang Z, Sheng M, Xiangbing K (2016b) Monitoring of sinking
659 flux of ocean particulate organic carbon using remote sensing methods. In: 2016
660 IEEE International Geoscience and Remote Sensing Symposium (IGARSS),
661 IEEE, pp 3788–3791