Journal of Big Data

METHODOLOGY

Open Access

Estimating the carbon content of oceans using satellite sensor data

Aadidev Sooknanan* and Patrick Hosein

*Correspondence:
aadidevsooknanan@gmail.com

The University of the West Indies, St. Augustine, Trinidad

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

Introduction

Climate change and ocean acidification are equally critical problems [7]. The Carbon cycle is important to climate, ecology and overall human livelihood [35]. Moreover, CO_2 is the most significant anthropogenic (human-induced) source of Carbon driving climate change [16]. 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 [12].

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 [34], specifically in sequestration of global carbon dioxide concentrations [16]. 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 [23]. 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% [13].

The magnitude and rate of anthropogenic Carbon sequestered by the oceans exceeds the extent of variation due to natural sources for the past millennium [21] 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 [10]. Fossil fuels and cement production account for approximately 48% of the world's global Carbon emissions [52]. In addition, deforestation, industrialization and land-use-changes have led to the unprecedented increase in Carbon emissions over the past 200 years [15].

Ocean acidification is the change in ocean chemistry driven by the oceanic uptake of Carbon [15]. 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 [28], 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 [29]. Ocean acidification has significant negative impacts on fundamental bio-ecological ocean processes [30, 52]. More alarmingly, past extinction events have been linked to ocean acidification [15], and the current rate of change in seawater chemistry is unprecedented [52].

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 [32]. Prediction and detection of carbon sinks are important issues with implications for all of human kind [21].

Laboratory measurements are the gold standard for assessing the carbon content of seawater, however research vessel time is costly and limited in coverage [31]. Remote sensing has the advantages of being fast, effective and near real-time [46]. The recent development of wide-band satellite imaging sensors has resulted in large quantities of high-resolution imagery being available [64]. Orbiting platforms additionally have the advantages of large observation range and high observation frequency [62] surpassing that of all alternate techniques such as in-situ buoys and marine vessels.

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 [59], multiple technical and political issues plague the development and launch of additional Carbon-monitoring instruments. This is evidenced by the multiple studies [3, 11, 20, 25, 26, 45, 50, 65] 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 [31]. Thus, remote sensing derived data via orbiting platforms hold great potential as a tool for monitoring changes in oceanic chemistry.

The remainder of this paper is organized as follows: In Section II, we review the related research in the field of earth observation and previous approaches at quantifying carbon levels, with special focus to how machine-learning methods facilitate the modeling of carbon levels. A description of the datasets used and the choice of carbon proxy is introduced in Section III. Section IV highlights the model-development process with particular focus to the choice of loss function. Section V details the results of model-development and training, and presents comparisons using permutations of predictor variables across three different models. Section VI underscores the scientific implications for our work, in addition to few limitations and potential for future work. Finally, our conclusions are discussed in Section VII.

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. [48] 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. [48] 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 several 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. [2] 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. [2], 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 [27, 57]. 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. [66] compared models for Dissolved Inorganic Carbon 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. [66] 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. [9] 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. [29], 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. [51] 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 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 [51].

Fugacity is the surface signature of ocean acidity, dynamics, and bio-geochemistry [36]. Liu W. Timothy [36] 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 and averaging was done in order to extrapolate point measurements from the satellite data.

Similarly, Liu and Xie [35] modelled Carbon Dioxide partial pressure at the ocean surface. However, according to the authors, their choice of model may not have 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 [35]. Moreover, according to the authors, sufficient salinity measurements were not available at the time of writing (similar to Liu W. Timothy [36]) 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 [58] between surface-level Carbon content and ocean-surface temperature.

Additionally, this work improves upon previous approaches (such as Liu W. Timothy [36]) 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 (Table 1).

Description of datasets

The HYCOM dataset

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°N and 80.48°S), HYCOM data are available on a standardized grid.

Data is assimilated from a combination of remote-sensing platforms GFO [1], ENVISAT [37] and Jason-1 [39] 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 [5].

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

Table 1 Table summarizing previous related work

Author	Main Idea	Limitation
Bates et al. [2]	Establish time-series for oceanic carbon levels using in-situ measurements	Data used was biased to spring-season, and relies on seaborne vessels to take measurements
Zui et al. [66]	Sea surface temperature and Chlorophyll were used to derived POC fluxes	This method did not incorporate salinity data, and the validation of results was done using point in-situ sources
Dixit et al. [9]	MLR and SVM models were compared at estimating pCO_2	In-situ verification of chlorophyll measurements were not performed, and the accuracy was not comparable to previous approaches due to limited longitude
Liu W. Timothy [36]	An SVM model using SST and SSS to predict sea-surface Carbon fugacity	Large data gaps from MODIS sensors resulted in the authors using significant smoothing to the input data, possibly reducing the model's ability to capture smaller variations
Liu and Xie [35]	Carbon dioxide partial pressure was modelled using a linear kernel	The temporal resolution of input data on a yearly-basis was objectively low, with large gaps in the satellite mapping data. Additionally, the authors highlighted the lack of dense, in-situ salinity measurements for model verification

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 [14, 42, 43]. The closest comparable earth-observing platform is NASA's Orbiting Carbon Observatories (OCO-2 and OCO-3). However these platforms are targeted to measuring atmospheric carbon levels and do not target the intended, ocean-based carbon.

HYCOM data contain the following combination of oceanic parameters indexed by z-coordinates, 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

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 $i \in \{0, 2\}$ indicates depth in meters. Table 2 shows latitude, longitude, temperature (C) and salinity (psu) measurements from the HYCOM data for the uppermost z-levels (0m and 2m respectively). In Fig. 1 we provide the mean salinity and mean temperature for the Caribbean region at a depth of 0m.

NOAA labelled data

The Ocean Chemistry and Ecosystems Division (OCED) of The National Oceanic and Atmospheric Administration (NOAA) focuses on 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. [60] 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 [47]. For this study 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

Table 2 Samples of the 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

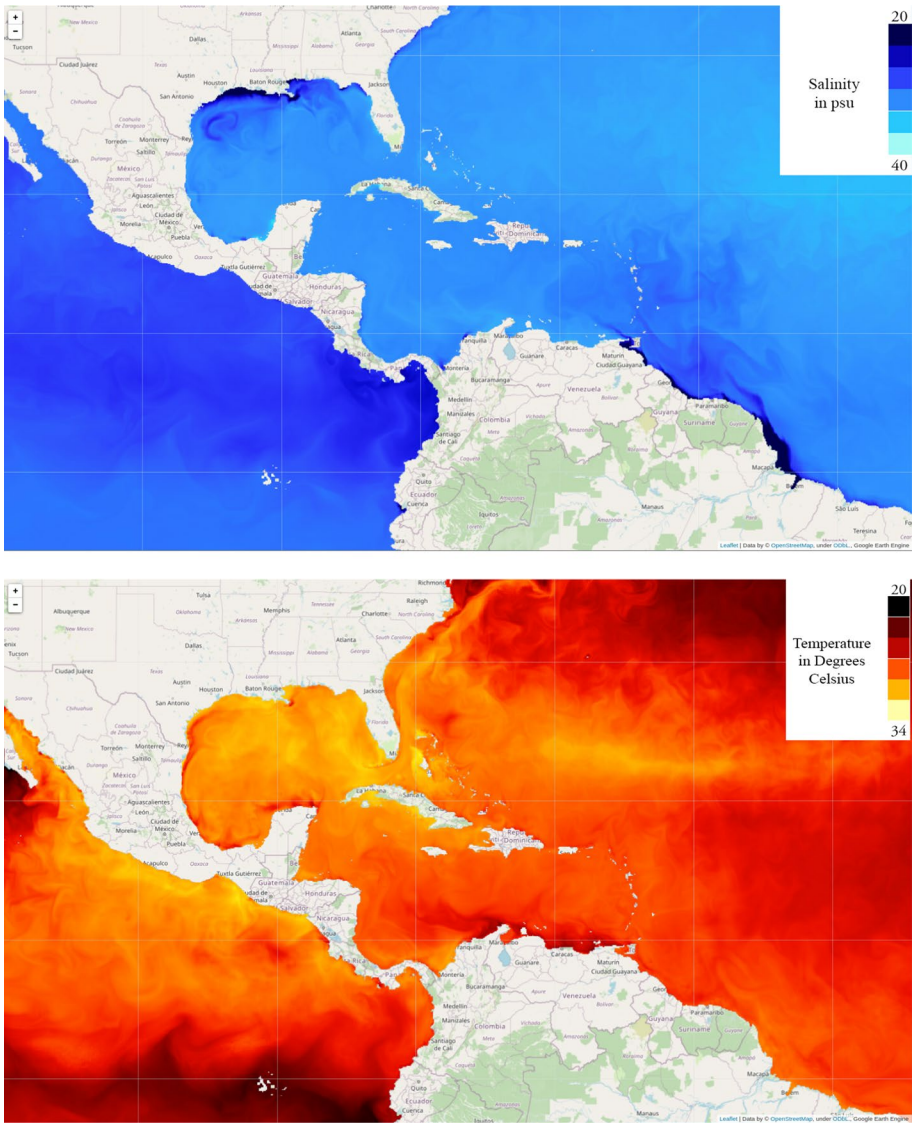

Fig. 1 Mean salinity (top) and mean temperature (bottom) at 0m for 2019-2020

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 were found such that the spatial resolution matched the HYCOM grids. Two flags in this data indicated 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 3. 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.

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 [24] and so $C_w = \alpha C_e$.

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 [22]. An alternative approach, in determining air-sea flux, is given by the difference in partial pressures between the air and sea CO_2

$$F = k_g \left(pCO_2(\mathsf{air}) - pCO_{2(\mathsf{sea})} \right) \tag{1}$$

However, the wind-speed dependent gas exchange coefficient k_g is not precisely known and there exists a discrepancy in the global mean values obtained by the two different methods [56]. In Fig. 2 we show the relationship between Carbon Fugacity and temperature (left) and salinity (right) at 0m.

Methods

A supervised learning algorithm models the implicit relationship existing in labelled data by means of a set of equations. 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 [54]. In our case the output attribute y is the fugacity of the partial-pressure of Carbon Dioxide which is dependent on the input attributes as determined by a feature-selection process. We use a gradient-boosting regression tree which provides a continuous-valued output as a nonlinear function of its input attributes.

Table 3 Samples of the NOAA data

lat	lon	CO_{2eq}	CO_{2atm}	p_{eq}	p_{atm}	t_{eq}	t_{sea}	sal_{sea}	f_{sea}	f_{atm}	$f_{atm} - f_{sea}$
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

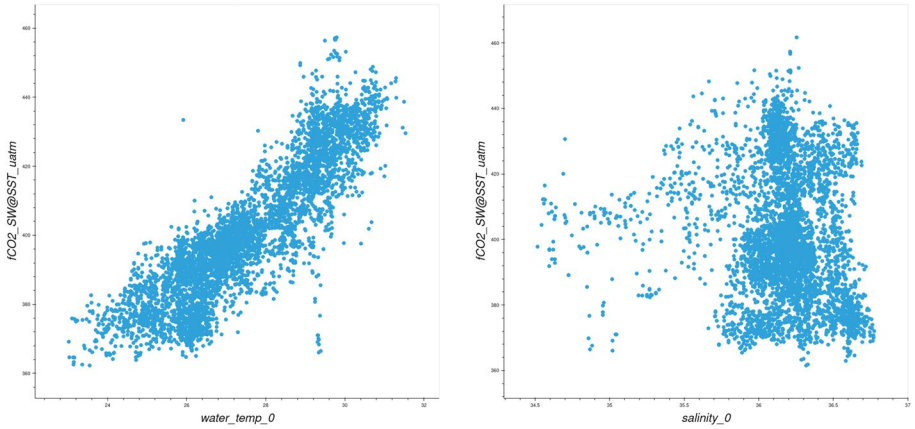

Fig. 2 Fugacity of CO_2 in sea water versus temperature and salinity at 0m

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 [41]. 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 [53]. In other terms, at each stage in the regression tree, the output domain is refined by means of an information criteria. The stochastic Gradient Boosted Regression Tree is adaptable, easy to interpret and produces highly accurate models [61]. 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 [63], by means of penalizing tree coefficients. In order to offset the overhead of the Gradient Boosted Decision Tree (GBDT) approach on large datasets, the NVIDIA RAPIDS [18] and the XGBoost [6] libraries were implemented on a GPU.

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 in the field of geosciences [4]. The MAE metric 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.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3}$$

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):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4}$$

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 modeling errors [4]. 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 [4].

As a result, the Huber loss is typically used to optimize the regression tree. The Huber loss behaves quadratically for small residual errors and linearly for large residual errors [19]. The Huber loss is given by the following:

$$\mathcal{L}(y,\hat{y}) = \begin{cases} (y-\hat{y})^2 & \forall \quad |y-\hat{y}| \leq \alpha \\ |y-\hat{y}| & \forall \quad |y-\hat{y}| > \alpha \end{cases} \tag{5}$$

This is equivalent to minimizing the Kullback-Leibler divergence [40] 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 [40]. This affords the Huber loss a significant advantage in managing 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 of the errors made. This process is complete when the model's increase in performance (or change in parameters) do not exceed some predefined threshold.

Numerical results

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 [55]. The target data are Fugacity values (fCO_2) for ocean surface water from the NOAA data. Binary interaction attributes were

Table 4 Unary predictors

Attribute	*p*-value
$salinity_0$	0.000110
$salinity_2$	0.001892
$temp_8$	0.009946

Table 5 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

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 the *p*-value. Both unary and binary predictor attributes were explored for the prediction task. 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 Tables 4 and 5, with feature importance for unary predictors being displayed in Fig. 3.

Model selection

Note that we initially investigated three models, Linear Regression, Gradient-Boosted Regression Tree and a Deep Neural Network consisting of 15 fully-connected layers. These models were trained using GPU libraries and evaluated with the Huber loss function. Since the GBRT model performed the best we focused solely on that approach. However, in order to demonstrate the differences we provide MSE values for the three approaches in Table 6.

Table 6 Comparing the error values for regression models

Model	Huber Loss	Mean-Squared Error
Linear regression	6.13	71.93
GBRT	4.39	42.06
Deep Neural Network	5.38	60.21

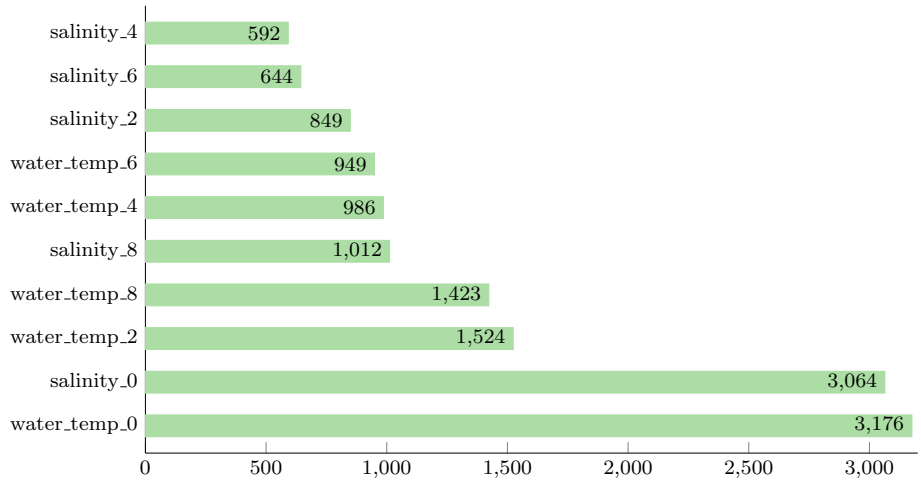

Fig. 3 Omnibus test values for single parameters

Table 7 Optimal GBRT hyper parameters

Hyper-parameter	Value
Maximum tree depth	7
Minimum child weight	5.0
Subsample	0.8
Learning rate	0.01

The GBRT was implemented using the XGBoost library and NVIDIA RAPIDS on a GTX 1060M GPU. 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 Fig. 7 were obtained.

Model validation

Ground-truth in-situ measurements were used to validate our model. This target data consisted fCO_2 in units of micro-atmospheres (μ-atm). 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

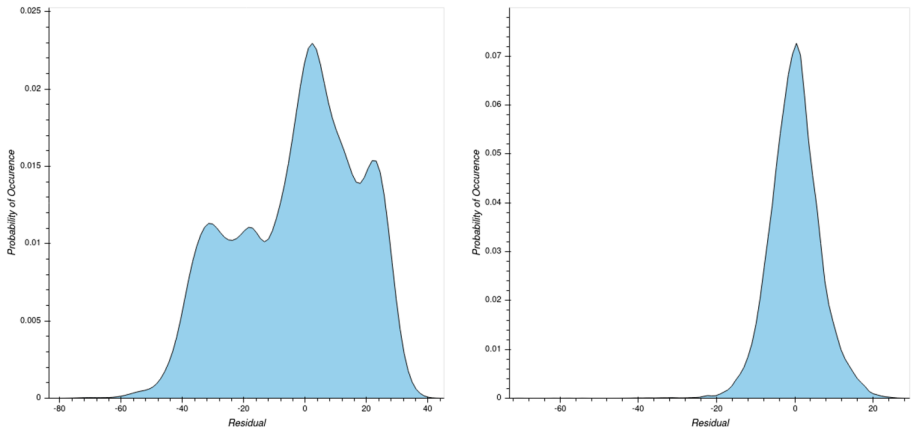

Fig. 4 Distribution of residuals using only unary predictors (left) and only interaction variables (right)

be 3.98, or approximately 1% of the mean value for the target attribute which is quite accurate.

As shown in Fig. 4, usage of binary interaction variables significantly improved the residual distribution of our model prediction. As shown in Fig. 4, the residual (error) terms for binary interaction variables are zero-centred with little spread. This is desirable, particularly when compared against the case of unitary predictors. 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).

Discussion

Scientific implications

Traditional remote-sensing (satellite-based) approaches do not measure Ocean Carbon levels directly. In this work, we investigate and the impact of 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 $fC0_2$ [35, 58].

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, there is no consensus agreement on the relationship between SST, SSS and fCO_2. 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 [36] and Krishna et al. [29]. 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 5) and application of an inherently non-linear model.

Some limitations and potential 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. Future work may potentially explore the usage of multi-modal approaches [44], particularly in the field of deep-learning [33], which has proven to be useful in handling heterogeneous inputs across multiple fields, including remote-sensing [17, 38, 49] 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 [36].

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 [35].

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.

Abbreviations

SST	Sea surface temperature
SSS	Sea surface salinity
MLD	Mixed layer depth
CDOM	Coloured dissolved organic matter
NPP	Net primary productivity
PAR	Photo-synthetically active radiation
DIC	Dissolved inorganic carbon
MODIS	Moderate resolution imaging spectro-radiometer
MLR	Multiple linear-regression model
MLR	Multiple linear-regression model
HYCOM	Hybrid coordinate ocean model
NOPP	National Ocean Partnership Program
NOAA	National Oceanic and Atmospheric Administration
GODAE	Global Ocean Data Assimilation Experiment
OCED	Ocean Chemistry and Ecosystems Division
GBRT	Gradient-boosting regression tree
RMSE	Root-Mean-Squared error
MAE	Mean absolute error
NASA	National Aeronautics and Space Administration
OCO	Orbiting carbon observatory

Acknowledgements
We wish to thank NVIDIA for their generous donation of the hardware resources (GPU) used for this research.

Author Contributions
This research and publication were pursued by AS and PH. All authors reviewed the manuscript.

Funding
Not applicable.

Availability of data and materials
Data used for this work is openly available via Google Earth Engine and the NOAA. Earth Engine Satellite Data was specifically published according to [8], and in-situ data is available via the NOAA's website. Standard software packages, namely ***NVIDIA RAPIDS***, ***Scikit-Learn*** and ***XGBoost*** were used in this work. Additionally, a code-repository of Python code is available at https://github.com/aadi350/water-salinity.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable (all data is freely available from public online sources).

Competing interests
Not applicable.

Received: 8 December 2021 Accepted: 27 June 2022
Published online: 15 July 2022

References

1. Barry R, Finkelstein J, Kilgus C, Mooers CNK, Needham B, Crawford M. Geosat follow-on satellite to supply ocean sciences data. Eos Trans Am Geophys Union. 1995;76(4):33–6.
2. Bates N, Best M, Neely K, Garley R, Dickson A, Johnson R. Detecting anthropogenic carbon dioxide uptake and ocean acidification in the north Atlantic ocean. Biogeosciences. 2012;9(7):2509–22.
3. Bosma J, Izett R, Izett R. Challenges with collecting data for measured ph and dissolved inorganic carbon (dic) in coastal waters. In: OCEANS 2016 MTS/IEEE Monterey. 2016. p. 1–5. https://doi.org/10.1109/OCEANS.2016.7761421
4. Chai T, Draxler RR. Root mean square error (rmse) or mean absolute error (mae)?-arguments against avoiding rmse in the literature. Geoscie Model Dev. 2014;7(3):1247–50.
5. Chassignet EP, Hurlburt HE, Smedstad OM, Halliwell GR, Hogan PJ, Wallcraft AJ, Baraille R, Bleck R. The hycom (hybrid coordinate ocean model) data assimilative system. J Mar Syst. 2007;65(1–4):60–83.
6. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. p. 785–794
7. Cooke SL, Kim SC. Exploring the evil twin of global warming: public understanding of ocean acidification in the united states. Sci Commun. 2019;41(1):66–89.
8. Cummings JA, Smedstad OM. Variational data assimilation for the global ocean. In: Data assimilation for atmospheric, oceanic and hydrologic applications, vol. II. Springer; 2013. p. 303–43.

9. Dixit A, Lekshmi K, Bharti R, Mahanta C. Net sea-air co_2 fluxes and modeled partial pressure of co co_2 in open ocean of bay of bengal. IEEE J Sel Top Appl Earth Obs Remote Sens. 2019;12(7):2462–9.
10. Doney SC, Balch WM, Fabry VJ, Feely RA. Ocean acidification: a critical emerging problem for the ocean sciences. Oceanography. 2009;22(4):16–25.
11. Dugan D, Janzen C, McCammon M, Evans W, Bidlack A. The evolution of ocean acidification observing efforts in alaska and the development of an alaska ocean acidification network. In: OCEANS 2017-Anchorage, IEEE; 2017. p. 1–6
12. Feely RA, Sabine CL, Lee K, Berelson W, Kleypas J, Fabry VJ, Millero FJ. Impact of anthropogenic CO_2 on the caco3 system in the oceans. Science. 2004;305(5682):362–6.
13. Gattuso JP, Magnan A, Billé R, Cheung WW, Howes EL, Joos F, Allemand D, Bopp L, Cooley SR, Eakin CM, et al. Contrasting futures for ocean and society from different anthropogenic CO_2 emissions scenarios. Science 2015;349(6243)
14. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ. 2017. https://doi.org/10.1016/j.rse.2017.06.031.
15. Guinotte JM, Fabry VJ. Ocean acidification and its potential effects on marine ecosystems. Ann N Y Acad Sci. 2008;1134(1):320–42.
16. Heinze C, Meyer S, Goris N, Anderson L, Steinfeldt R, Chang N, Le Quéré C, Bakker DC. The ocean carbon sink-impacts, vulnerabilities and challenges. Earth Syst Dyn. 2015;6(1):327–58.
17. Hong D, Gao L, Yokoya N, Yao J, Chanussot J, Du Q, Zhang B. More diverse means better: multimodal deep learning meets remote-sensing imagery classification. IEEE Trans Geosci Remote Sens. 2020;59(5):4340–54.
18. Hricik T, Bader D, Green O. Using rapids ai to accelerate graph data science workflows. In: 2020 IEEE high performance extreme computing conference (HPEC). IEEE; 2020. p. 1–4
19. Huber PJ. Robust estimation of a location parameter. In: Breakthroughs in statistics. Springer; 1992. p. 492–518
20. Hyde A, Vandemark D, Shellito S, Salisbury J, Irish J, DeGrandpre M. 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. 2011. p. 1–5, https://doi.org/10.23919/OCEANS.2011.6107017
21. Imaoka K, Kachi M, Fujii H, Murakami H, Hori M, Ono A, Igarashi T, Nakagawa K, Oki T, Honda Y, et al. Global change observation mission (gcom) for monitoring carbon, water cycles, and climate change. Proc IEEE. 2010;98(5):717–34.
22. Jang E, Im J, Park GH, Park YG. Estimation of fugacity of carbon dioxide in the east sea using in situ measurements and geostationary ocean color imager satellite data. Remote Sens 2017;9(8). https://doi.org/10.3390/rs9080821, https://www.mdpi.com/2072-4292/9/8/821
23. Jiahui W, Liang L, Han L, Chunyang C, Ting H, Di G. 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). 2019. p. 1–4. https://doi.org/10.1109/Agro-Geoinformatics.2019.8820487
24. Johnson JE. Evaluation of a seawater equilibrator for shipboard analysis of dissolved oceanic trace gases. Anal Chim Acta. 1999;395(1–2):119–32.
25. Johnson KS. Bioargo: A global scale chemical sensor network to observe carbon, oxygen, and nitrogen cycles in the ocean. In: SENSORS, 2013 IEEE; 2013. p. 1. https://doi.org/10.1109/ICSENS.2013.6688480
26. Johnson RW, Ohlhorst CW. Application of remote sensing to monitoring and studying dispersion in ocean dumping. In: Ocean dumping of industrial wastes. Springer; 1981. p. 175–191
27. Keil P, Mauritsen T, Jungclaus J, Hedemann C, Olonscheck D, Ghosh R. Multiple drivers of the north Atlantic warming hole. Nat Clim Chang. 2020;10(7):667–71.
28. Körtzinger A. Determination of carbon dioxide partial pressure (p (CO_2)). Methods of seawater analysis. 1999. p. 149–158
29. Krishna KV, Shanmugam P, Nagamani PV. A multiparametric nonlinear regression approach for the estimation of global surface ocean pco_2 using satellite oceanographic data. IEEE J Select Top Appl Earth Obs Remote Sens. 2020;13:6220–35. https://doi.org/10.1109/JSTARS.2020.3026363.
30. Kroeker KJ, Kordas RL, Crim RN, Singh GG. Meta-analysis reveals negative yet variable effects of ocean acidification on marine organisms. Ecol Lett. 2010;13(11):1419–34.
31. Land PE, Shutler JD, Findlay HS, Girard-Ardhuin F, Sabia R, Reul N, Piolle JF, Chapron B, Quilfen Y, Salisbury J, et al. Salinity from space unlocks satellite-based assessment of ocean acidification 2015.
32. Le Quéré C, Raupach MR, Canadell JG, Marland G, Bopp L, Ciais P, Conway TJ, Doney SC, Feely RA, Foster P, et al. Trends in the sources and sinks of carbon dioxide. Nat Geosci. 2009;2(12):831–6.
33. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
34. Lekshmi K, Bharti R, Mahanta C. Spatio-temporal distribution of carbon dioxide partial pressure in the bay of bengal. In: IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium. IEEE; 2019. p. 8165–8168
35. Liu WT, Xie X. Space observation of carbon dioxide partial pressure at ocean surface. IEEE J Sel Top Appl Earth Obs Remote Sens. 2017;10(12):5472–84.
36. Liu W Timothy XX. Ocean surface carbon dioxide fugacity observed from space. Report, National Aeronautics and Space Administration, USA, 2014. https://archimer.ifremer.fr/doc/00651/76344/
37. Louet J, Bruzzi S. Envisat mission and system. In: IEEE 1999 international geoscience and remote sensing symposium. IGARSS'99 (Cat. No. 99CH36293), vol 3. IEEE; 1999. p. 1680–1682
38. Maimaitijiang M, Sagan V, Sidike P, Hartling S, Esposito F, Fritschi FB. Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sens Environ. 2020;237: 111599.
39. Ménard Y, Fu LL, Escudier P, Parisot F, Perbos J, Vincent P, Desai S, Haines B, Kunstmann G. The Jason-1 mission special issue: Jason-1 calibration/validation. Mar Geodesy. 2003;26(3–4):131–46.
40. Meyer GP. An alternative probabilistic interpretation of the huber loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021; p. 5261–5269
41. Mitchell R, Adinets A, Rao T, Frank E. Xgboost: Scalable gpu accelerated learning. 2018. arXiv preprint arXiv:1806.11248
42. NASA (2022a) Ocean biology daac (ob.daac). https://earthdata.nasa.gov/eosdis/daacs/obdaac

43. NASA (2022b) Physical oceanography distributed active archive center (po.daac) home page. https://podaac.jpl.nasa.gov/
44. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: ICML 2011.
45. Ohlhorst CW. Quantitative mapping by remote sensing of an ocean acid-waste dump, vol 1275. National Aeronautics and Space Administration; 1978.
46. Paltsyn MY, Gibbs JP, Mountrakis G. Integrating traditional ecological knowledge and remote sensing for monitoring rangeland dynamics in the altai mountain region. Environ Manage. 2019;64(1):40–51.
47. Pierrot D, Neill C, Sullivan K, Castle R, Wanninkhof R, Lüger H, Johannessen T, Olsen A, Feely RA, Cosca CE. Recommendations for autonomous underway pco_2 measuring systems and data-reduction routines. Deep Sea Res Part II. 2009;56(8–10):512–22.
48. Poli S, Franzolin E, Fumagalli P, Crottini A. The transport of carbon and hydrogen in subducted oceanic crust: an experimental study to 5 gpa. Earth Planet Sci Lett. 2009;278(3–4):350–60.
49. Ramachandram D, Taylor GW. Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process Mag. 2017;34(6):96–108.
50. Saba GK, Wright-Fairbanks E, Miles TN, Chen B, Cai WJ, Wang K, Barnard AH, Branham CW, Jones CP. Developing a profiling glider ph sensor for high resolution coastal ocean acidification monitoring. In: OCEANS 2018 MTS/IEEE Charleston. IEEE; 2018. p. 1–8
51. Sabia R, Fernéndez-Prieto D, Shutler J, Donlon C, Land P, Reul N. Remote sensing of surface ocean ph exploiting sea surface salinity satellite observations. In: 2015 IEEE international geoscience and remote sensing symposium (IGARSS). 2015; p. 106–109. https://doi.org/10.1109/IGARSS.2015.7325709
52. Sabine CL, Feely RA, Gruber N, Key RM, Lee K, Bullister JL, Wanninkhof R, Wong C, Wallace DW, Tilbrook B, et al. The oceanic sink for anthropogenic CO_2. science. 2004;305(5682):367–71.
53. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern. 1991;21(3):660–74.
54. Saravanan R, Sujatha P. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In: 2018 second international conference on intelligent computing and control systems (ICICCS). IEEE; 2018. p. 945–949
55. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: Proceedings of the 9th python in science conference, Austin, TX, vol 57. 2010. p. 61
56. Siegenthaler U, Sarmiento JL. Atmospheric carbon dioxide and the ocean. Nature. 1993;365(6442):119–25.
57. Sutton RT, Dong B, Gregory JM. Land/sea warming ratio in response to climate change: Ipcc ar4 model results and comparison with observations. Geophys Res Lett 2007;34(2)
58. Takahashi T, Sutherland SC, Sweeney C, Poisson A, Metzl N, Tilbrook B, Bates N, Wanninkhof R, Feely RA, Sabine C, et al. Global sea-air CO_2 flux based on climatological surface ocean pco_2, and seasonal biological and temperature effects. Deep Sea Res Part II. 2002;49(9–10):1601–22.
59. Tollefson J. Carbon-sensing satellite system faces high hurdles: space agencies plan an advanced fleet, but technical and political challenges abound. Nature. 2016;533(7604):446–8.
60. Wanninkhof R, Pierrot D, Sullivan K, Barbero L, Triñanes J. A 17-year dataset of surface water fugacity of co 2 along with calculated ph, aragonite saturation state and air-sea co 2 fluxes in the northern caribbean sea. Earth Syst Sci Data. 2020;12(3):1489–509.
61. Wen Z, He B, Kotagiri R, Lu S, Shi J. Efficient gradient boosted decision tree training on gpus. In: 2018 IEEE international parallel and distributed processing symposium (IPDPS). IEEE; 2018. p. 234–243
62. Yang K. Progress, challenge and prospect for remote sensing monitoring of flood and drought disasters in china. In: IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium. IEEE; 2019. p. 4280–4283
63. Ye J, Chow JH, Chen J, Zheng Z. Stochastic gradient boosted distributed decision trees. In: Proceedings of the 18th ACM conference on Information and knowledge management. 2009. p. 2061–2064
64. Zhu Q, Sun X, Zhong Y, Zhang L. High-resolution remote sensing image scene understanding: A review. In: IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium. IEEE; 2019. p. 3061–3064
65. : Zui T, Tingting L, Xiang Z, Sheng M, Xiangbing K. Monitoring of sinking flux of ocean particulate organic carbon using remote sensing methods. In: 2016 IEEE international geoscience and remote sensing symposium (IGARSS). 2016a. p. 3788–3791. https://doi.org/10.1109/IGARSS.2016.7729982
66. Zui T, Tingting L, Xiang Z, Sheng M, Xiangbing K. Monitoring of sinking flux of ocean particulate organic carbon using remote sensing methods. In: 2016 IEEE international geoscience and remote sensing symposium (IGARSS). IEEE. 2016b. p. 3788–3791

Publisher's Note