# Estimating historical air-sea CO2 fluxes: Incorporating physical knowledge within a data-only approach

Val Bennington[1], Tomislav Galjanic[2], and Galen A McKinley[3]

[1]Lamont Doherty Earth Institute of Columbia University
[2]Data Science Institute at Columbia University
[3]Lamont Doherty Earth Observatory of Columbia University

November 25, 2022

**Abstract**

The ocean plays a critical role in reducing human impact on the global climate by absorbing and sequestering CO2 from the atmosphere. To quantify the ocean's role in the global carbon budget, we need surface ocean pCO2 across space and time, but only sparse observations exist. The typical approach to reconstructing pCO2 is to train a machine learning approach on a subset of the pCO2 data and available physical and biogeochemical observations. Though the variables are all related to the pCO2, these approaches are often perceived as black boxes, as it is unclear how inputs are physically linked to pCO2 outputs. Here, we add physics by incorporating our knowledge of the direct effect of temperature on surface ocean pCO2. We use the machine learning algorithm XGBoost to develop a function between satellite and in-situ observations and the difference between observed pCO2 and the pCO2 that would exist if temperature variations were the only driver of variability. We show the resulting model is physically consistent, and performs at least as well as other data approaches. Uncertainty in the reconstructed pCO2 and its impact on the estimated CO2 fluxes are quantified. Uncertainty in piston velocity drives flux uncertainties. The historical reconstructed CO2 fluxes show larger interannual variability than the smoother neural network approaches, but a lesser trend since 2005. We estimate an air-sea flux of -2.3 +/- 0.5 PgC/yr for 1990-2018, agreeing with other data products and the Global Ocean Carbon Budget models of 2021 estimate of -2.3 +/- 0.4 PgC/yr.

# Estimating historical air-sea $CO_2$ fluxes: Incorporating physical knowledge within a data-only approach

**Val Bennington[1], Tomislav Galjanic[2], Galen A McKinley[1]**

[1]Lamont-Doherty Earth Institute, Columbia University

[2]Data Science Institute, Columbia University

**Key Points:**

- New machine learning approach incorporates physical knowledge of ocean carbon system into algorithms to extrapolate to global coverage.
- Reconstructed $pCO_2$ agree with independent data.
- Estimated ocean carbon uptake has a trend since 2005 that is on the lower end of previous observation-based estimates.

Corresponding author: Val Bennington, vbennington@ldeo.columbia.edu

## Abstract

The ocean plays a critical role in reducing human impact on the global climate by absorbing and sequestering $CO_2$ from the atmosphere. To quantify the ocean's role in the global carbon budget, we need surface ocean $pCO_2$ across space and time, but only sparse observations exist. The typical approach to reconstructing $pCO_2$ is to train a machine learning approach on a subset of the $pCO_2$ data and available physical and biogeochemical observations. Though the variables are all related to the $pCO_2$, these approaches are often perceived as black boxes, as it is unclear how inputs are physically linked to $pCO_2$ outputs. Here, we add physics by incorporating our knowledge of the direct effect of temperature on surface ocean $pCO_2$. We use the machine learning algorithm XGBoost to develop a function between satellite and in-situ observations and the difference between observed $pCO_2$ and the $pCO_2$ that would exist if temperature variations were the only driver of variability. We show the resulting model is physically consistent, and performs at least as well as other data approaches. Uncertainty in the reconstructed $pCO_2$ and its impact on the estimated $CO_2$ fluxes are quantified. Uncertainty in piston velocity drives flux uncertainties. The historical reconstructed $CO_2$ fluxes show larger interannual variability than the smoother neural network approaches, but a lesser trend since 2005. We estimate an air-sea flux of -2.3 $\pm$ 0.5 PgC/yr for 1990-2018, agreeing with other data products and the Global Ocean Carbon Budget models of 2021 estimate of -2.3 $\pm$ 0.4 PgC/yr. (Friedlingstein et al., 2021).



## Plain Language Summary

The ocean absorbs carbon dioxide from the atmosphere, moderating the human impact on the world's climate. To quantify how much carbon dioxide is removed from the atmosphere by the ocean each year, we must know how much gas is exchanged at each location across the ocean over time. The observations necessary to quantify this gas exchange are very sparse and require gap-filling in both space and time. Because of the heterogeneity of this gas exchange, complex relationship between the ocean observations with near global coverage and ocean carbon are determined using machine learning algorithms. These techniques are often perceived as black boxes, where inputs are converted to outputs without much explanation. Here, we develop a novel machine learning approach that explicitly incorporates physical knowledge of the ocean carbon cycle into the reconstruction approach. We show that our technique results in a physically consistent model and

estimates the ocean carbon sink to be in similar magnitude as ocean carbon biogeochemical models.

## 1 Introduction

The ocean plays a significant role in reducing human impact on the climate by absorbing and sequestering approximately one quarter of anthropogenic carbon dioxide ($CO_2$) emissions each year (McKinley et al., 2016; Khatiwala et al., 2013; Sabine et al., 2004; Friedlingstein et al., 2021). Since the beginning of the Industrial Revolution, the ocean has absorbed about a third of the total anthropogenic emissions. The mean state of the ocean carbon cycle is well defined (Takahashi et al., 2009; Gloege, Yan, et al., 2021). Yet, the quantification of year-to-year variability and long-term changes in this carbon sink remains a challenge. This quantification is necessary for climate policies worldwide in order to separate the impact of any mitigation policies from interannual variability in the ocean carbon sink (Peters et al., 2017).

To quantify the variability and trend in the ocean carbon cycle, both global ocean biogeochemical models (GOBMs) and statistical approaches are used. The degree to which these methods agree builds confidence in estimates of the ocean carbon sink, and its variability. GOBMs are mechanistic models which incorporate our knowledge of the processes that control the ocean carbon cycle and the resulting air-sea fluxes of carbon dioxide. While the models can be compared to observations to assess performance (Hauck et al., 2020), they do not directly incorporate observations of the partial pressure of carbon dioxide in the surface ocean ($pCO_2$). The nine models included in the Global Carbon Budget have some significant biases and often poor correlations with independent data sets (Fay & McKinley, 2021). Observation-based data products reconstruct the surface ocean $pCO_2$ across the global ocean in both space and time from sparse measurements using statistical techniques. Then, air-sea fluxes of carbon dioxide are calculated from the resulting air-sea difference ($\Delta pCO_2 = pCO_2^{ocean} - pCO_2^{atm}$). These data products typically use machine learning algorithms to develop a nonlinear function between observations of surface ocean $pCO_2$ and related variables that can be observed with greater spatio-temporal coverage. The resulting function is then used to extrapolate $pCO_2$ across the global ocean in both space and time. While the resulting observation-based products show higher correlations and smaller RMSE against observations than do models (Hauck et

al., 2020), the algorithms are often viewed as black boxes due to the purely statistical machine learning techniques used to extrapolate.

Reichstein et al. (2019) state that while machine learning approaches may fit observations well, they may be unrealistic or not be interpretable. One method to improve plausibility and confidence in the data-based approach of estimating global air-sea fluxes is to incorporate well-accepted physical knowledge into the novel machine learning approaches. Incorporating physical knowledge within a machine learning approach is relatively new to the geosciences, but has typically been implemented using a modified cost function that penalizes unphysical results. Typical machine learning algorithms develop a function by minimizing a cost function. This cost function is usually a sum of the Mean Squared Error (MSE) between the predicted $pCO_2$ and the observed $pCO_2$ plus a regularization term. This regularization term is used to penalize complexity in the resulting algorithm, so the algorithm will generalize better. Read et al. (2019), use a neural network approach to predict lake temperature profiles in Lake Mendota and Sparkling Lake. Root mean squared error (RMSE) was smaller compared to predictions form a process based model. However, a standard neural network approach resulted in unphysical conditions at times. To improve upon the standard neural network approach, Read et al. (2019) modify their cost function to include a penalty for model predictions that result in non-physical conditions: denser water on top of lighter water. Their final model further reduces RMSE such that the final neural network provides the best prediction of lake temperature profiles.

For Lake Mendota and Sparkling Lake temperature profiles, there is a physical condition that can easily be penalized. When reconstructing heterogeneous surface ocean $pCO_2$, there is no obvious way to penalize the model for a given $pCO_2$ prediction based upon neighboring predictions. So how can we incorporate the physical mechanisms we know control the ocean carbon cycle within a machine learning approach? Previous machine learning approaches to reconstructing surface ocean $pCO_2$ rely on the algorithm to decipher the ways in which atmospheric $CO_2$, sea surface temperature, chlorophyll-a, mixed layer depth climatology, sea surface salinity, winds, geographic location, and time of year impact the resulting surface ocean $pCO_2$. Each of these features impacts $pCO_2$. Chlorophyll-a provides a measure of the biological production that removes dissolved inorganic carbon (DIC) from the surface ocean, thereby reducing surface ocean $pCO_2$. Mixed layer depth is a proxy for ocean stratification. During highly stratified times,

the phytoplankton are held within the lit surface ocean, setting up production. During periods of deeper mixing, DIC from depth is brought to the surface, and an increase in surface ocean $pCO_2$ occurs. However, mixed layer depths are also an indicator of temperature. Temperature has both direct and indirect effects on surface ocean $pCO_2$. Increasing (decreasing) temperatures directly result in an increase (decrease) of $pCO_2$ (Takahashi et al., 2002). However, temperature variations also set up biological production via stratification and wintertime vertical mixing, processes that result in opposing $pCO_2$ changes compared to the direct temperature effect on $pCO_2$.

Previous approaches rely on a machine learning algorithm to create a single function that disentangles the competing effects of temperature variations by relying on other variables such as Chlorophyll-a (Chl-a) and mixed layer depth (MLD). At the same time, we do know well the direct temperature impact on $pCO_2$ from the empirical work of Takahashi et al. (2002). Here, we develop a hybrid modeling approach that removes the well-known direct effect of temperature on $pCO_2$ from our regression, and asks the machine learning algorithm to learn only the indirect effects of temperature on $pCO_2$, supported by the information from other input variables. We introduce the $pCO_2$-Residual approach and show that the resulting model does in fact capture the gross physical processes we know to be true. Additionally, it performs as well as the best other data-based approaches when compared to observations. The resulting model is used to estimate the air-sea $CO_2$ fluxes for 1985-2019, and uncertainties are quantified.

## 2 Methods

### 2.1 $pCO_2$-Residual

To incorporate physical knowledge of the system, we calculate a residual ($pCO_2$-Residual), the difference between observed $pCO_2$ and the purely temperature driven component of $pCO_2$ ($pCO_2$-T). We use a machine learning algorithm, eXtreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016), to develop a function between observations and the $pCO_2$-Residual, to reconstruct the residual across all space and time. For the final reconstruction of surface ocean $pCO_2$, we add $pCO_2$-T back to our residual. $CO_2$ fluxes are then calculated using the reconstructed $pCO_2$.

#### *2.1.1 Pre-processing SOCAT observations*

We calculate surface ocean $pCO_2$ from the SOCAT v2021 monthly gridded $fCO_2$ product (Bakker et al., 2016). This is a quality-controlled dataset containing observations of the fugacity of carbon dioxide ($fCO_2$) in the surface ocean that is converted to surface ocean $pCO_2$ according to Equation 1,

$$pCO_2 = fCO_2 \cdot exp\left(P_{atm} \cdot \frac{B + 2\delta}{R \cdot T}\right)^{-1} \quad (1)$$

where $P_{atm}$ is the atmospheric pressure at sea level from ERA5, T is the sea surface temperature (SST) in Kelvin from the National Oceanic and Atmospheric Administration (NOAA) optimally interpolated SST version 2 (OISSTv2), B and $\delta$ are virial coefficients from Weiss (1974), and R is the gas constant (Dickson et al., 2007). The data are sparse in both space and time, with significant coverage gaps throughout the southern hemisphere, particularly during winter. See Gregor et al. (2019) and Gloege et al. (2021) for details of data coverage.

### 2.2 Initial $pCO_2$ Reconstruction

Utilizing processed SOCAT $pCO_2$ and the XGBoost algorithm, we do an initial global reconstruction of $pCO_2$ for 1982-2019 utilizing the observations and data products in Table 1. This reconstruction is only used to determine the mean $pCO_2$ at all locations over the period 1985-2019 that is required for calculation of the $pCO_2$-Residual (see Section 2.2.1).

#### *2.2.1 Calculating $pCO_2$-Residual*

We calculate the temperature driven component of $pCO_2$ ($pCO_2$-T) via Equation 2 (Takahashi et al., 2002),

$$pCO_2T = \overline{pCO_2} \cdot exp(0.0423 \cdot (SST - \overline{SST})) \quad (2)$$

where $\overline{pCO_2}$ is mean surface ocean $pCO_2$ from the initial $pCO_2$ reconstruction, SST is temperature in Celsius from NOAA OISSTv2, and $\overline{SST}$ is the local long term mean in SST in Celsius from NOAA OISSTv2. The residual ($pCO_2$-Residual) is calculated as the difference between observed $pCO_2$ and $pCO_2$T for all observations, and this process is shown in detail in Figure 1.

$$pCO_2^{Residual} = pCO_2 - pCO_2T \tag{3}$$

We examine the properties of the residual in Figure 2. In regions such as the subtropics, where $pCO_2$ is primarily driven by the direct effects of temperature, mean absolute value of the residual is small (Figure 2a). Regions where the seasonal cycle of $pCO_2$ is not dominantly controlled by temperature, such as the subpolar regions, have larger residuals. Thus, the subtropical regions have residuals on the order of 10 $\mu$atm, while subpolar regions may have residuals on the order of 100 $\mu$atm. Looking at the seasonality of the residual in Figure 2c and 2d, we see that during local winter, the residual is large and positive in the subpolar regions where vertical mixing returns DIC to the surface waters and $pCO_2$ is increased even though temperatures are low. During local summer, the subpolar regions have negative residuals, where biological drawdown of DIC reduces the increase in $pCO_2$ expected from the increases in temperature. The seasonal residual is small in magnitude in the subtropical regions where temperature is primary driver of surface ocean $pCO_2$. The $pCO_2$-Residual in the observations is approximately normally distributed (Figure 2b), with a small positive mean. This non-zero mean is due to the increasing rate of sampling, with more observations occurring when the $pCO_2$-Residual is larger in magnitude.

We experimented with this approach using ensembles of four Earth System Models, a technique developed by Gloege et al. (2021), and confirmed its ability to reconstruct surface ocean $pCO_2$, providing confidence in our approach (Section S1). We found significance increases in the ability to reconstruct of $pCO_2$ across the global ocean, particularly in the poorly sampled regions in the southern hemisphere where temperature is a primary driver of surface ocean $pCO_2$, as compared to when reconstructing $pCO_2$ without the knowledge of $pCO_2$-T.

### 2.3 XGBoost

The machine learning algorithm XGBoost is used to reconstruct the $pCO_2$-Residual across the global surface ocean for 1982-2019. XGBoost is a supervised machine learning algorithm that utilizes Extreme Gradient Boosting (Chen & Guestrin, 2016) to predict a target variable (y), the $pCO_2$-Residual from multiple features, (X) such as SST, SSS, chlorophyll-a, and mixed layer depth. The algorithm estimates a non-linear function such that $f(X) \approx y$. The algorithm begins with a single initial guess of the $pCO_2$-

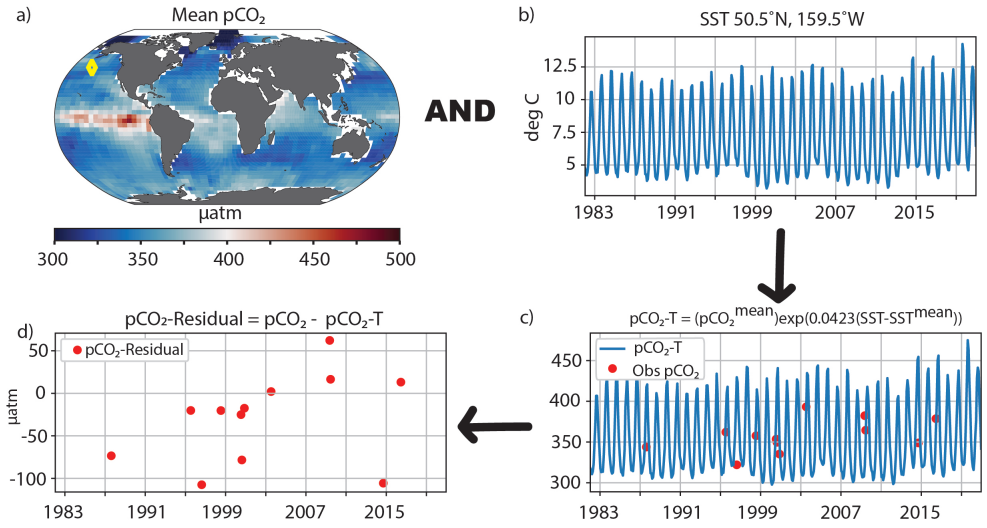


**Figure 1.** a) Mean surface ocean $pCO_2$ from the initial run. b) Observed satellite SST time series from location of yellow diamond in subplot (a). c) Calculated $pCO_2T$ (blue) and observed $pCO_2$ (red dots) at yellow diamond located in (a). d) The calculated $pCO_2$Residual, or the difference between observed $pCO_2$ and calculated $pCO_2T$ at location specified in (a).

Residual (one value for the entire globe at all times.) Then, decision trees made up of the features are added one by one, which adjust the initial guess to reduce the loss, or difference between the $pCO_2$-Residual in the training data and the prediction. The process of adding trees is continued until the maximum number of trees permitted is reached, or when adding an additional tree does not improve the calculated cost function. Here, the loss function is the mean squared error (MSE) between the training data and the predictions. The final prediction of $pCO_2$-Residual is the sum of the initial guess and the result of all the decision trees.

The features and associated $pCO_2$-Residuals are split into validation, training, and testing sets. The validation set is used to optimize the hyperparameters of the algorithm, namely, the number of trees used and maximum depth of each tree. Our final XGBoost algorithm uses 1000 decision trees with a maximum depth of 7 levels. The training set is used to build the function between the features and the residual; i.e., the training set builds the decision trees. The testing set is withheld to test how well the function generalizes. Once the hyperparameters are determined, we separate the training data from the test data by month. Four months are used for training, and then the next month for testing, similar to Gregor et al. (2019), who shift years. This is repeated throughout the

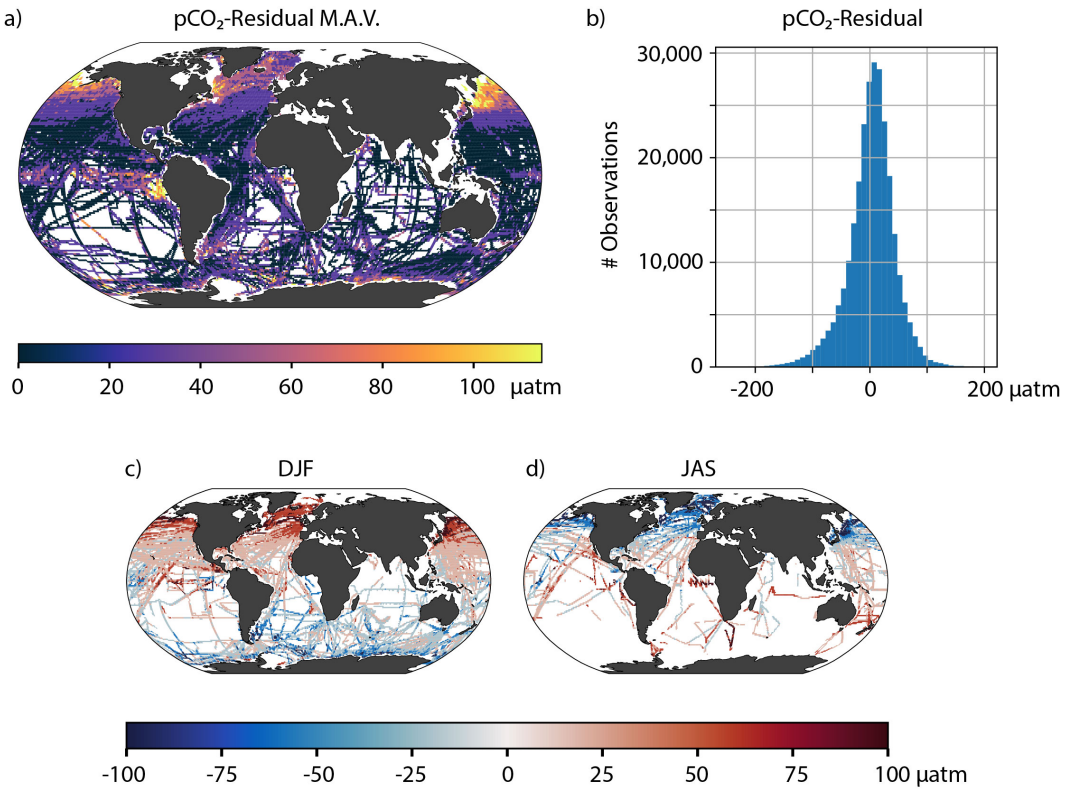


**Figure 2.** (a) Mean of the absolute value (M.A.V.) of the $pCO_2$-Residual calculated from all observations in the SOCAT database. (b) Histogram of the calculated $pCO_2$-Residual from SOCAT observations. (c) Mean $pCO_2$-Residual calculated for all observations during the northern hemisphere winter (DJF). (d) Same as (c) but for southern hemisphere winter (JAS).

dataset. This is done to reduce the number of individual cruises seen in both the training and test data, but to train on observations from all years. We develop five models by shifting our initial month of testing data, and our final estimate of the residual is the ensemble mean of the five predictions.

### 2.4 Features

In order to reconstruct the residual across both space and time, datasets with approximately full global coverage are used (Table 1): Sea Surface Temperature (SST) and Chlorophyll-a (Chl-a) from satellite; Sea Surface Salinity (SSS) from in situ data (Good et al., 2013); Mixed Layer Depth (MLD) climatology from Argo floats (de Boyer Montégut et al., 2004); and the mixing ratio of atmospheric $CO_2$ from global stations (Masarie, 2012). Additional interannual anomalies are derived for SST, SSS, and Chl-a by subtracting the monthly climatology of the feature from a given month's observation. Geographic location and time of year are incorporated using an N-vector transformation of latitude and longitude and a time transformation of day of year. We tested using self organizing maps to separate the ocean according to their feature properties into 5, 10, and 15 biomes, but improvement was negligible, so we maintain the simpler model (Supplementary).

We tested the sensitivity of the reconstruction to the source of mean $pCO_2$ ($\overline{pCO_2}$) used in the calculation of $pCO_2T$ with Equation 2, which is then input to the $pCO_2$-Residual calculation in Equation 3. Reconstructions using LDEO $pCO_2$ (Takahashi et al., 2009) and the mean $pCO_2$ of the SeaFlux data products (Fay et al., 2021) as the $pCO_2$ mean. The alternative sources of $pCO_2$ mean did not significantly impact the reconstructed $pCO_2$ or resulting air-sea $CO_2$ exchange, so we maintain the internally consistent method of the initial reconstruction of $pCO_2$.

#### *2.4.1 Chlorophyll-a*

We utilize satellite Chlorophyll-a of GlobColour (Maritorena et al., 2010) for 1998-2019. We fill the missing winter months at the poles by linearly interpolating between the last month observed prior to the winter and the first month observed after winter. This results in lower chlorophyll values during winter than if we had used annual means to fill in the gaps. This same technique is used when any month is missing observations

**Table 1.** Summary of the products, variables, and processing steps used for feature and target datasets.

| **Product** | **Variable** | **Abbreviation** | **Processing** |
|---|---|---|---|
| NOAA OISSTv2[1] | Sea Surface Temperature | SST | - |
| | SST anomaly | SST' | SST - monthly clim |
| Met Office: EN4[2] | Salinity | SSS | - |
| | SSS anomaly | SSS' | SSS - monthly clim |
| NOAA: GLOBALVIEW[3] | Atmospheric $CO_2$ | $xCO_2$ | - |
| ESA GlobColour[4] | Chl a | Chl a | $Log_{10}$(Chla) |
| | Chl a anomaly | Chl a' | Chl a - monthly clim |
| deBoyer Montegut[5] | Mixed Layer Depth | MLD | $Log_{10}$(MLD) |
| $pCO_2$ | Mean $pCO_2$ | $pCO_2$ clim | Equation 2 |
| SOCATv2020[6] | Partial pressure of $CO_2$ | $pCO_2$ | Equations 1,3 |
| | Geographic Location | A | $\sin(\lambda)$ |
| - | | B | $\sin(\mu)\cos(\lambda)$ |
| | | C | $-\cos(\mu)\cos(\lambda)$ |
| | Time of Year | $T_0$ | $sin\left(\frac{j*2\pi}{365}\right)$ |
| - | | $T_1$ | $cos\left(\frac{j*2\pi}{365}\right)$ |

[1] Source: https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html, (Reynolds et al., 2002)

[2] Source: https://www.metoffice.gov.uk/hadobs/en4/, (Good et al., 2013)

[3] Source: https://www.esrl.noaa.gov/gmd/ccgg/globalview/co2/co2 intro.html, (Masarie, 2012)

[4] Source: http://www.globcolour.info/, (Maritorena et al., 2010)

[5] Source: http://www.ifremer.fr/cerweb/deboyer/mld/home.php, (de Boyer Montégut et al., 2004)

[6] Source: https://www.socat.info/, (Bakker et al., 2016)

outside of the poles. Since no full year of satellite observations are available prior to 1998, we use the climatology of Chlorophyll-a calculated from 1998-2019 observations at all locations and months prior to 1998. Within the Large Ensemble Testbed, utilizing climatological chlorophyll prior to 1998 introduced a mean uncertainty of 0.1 Pg C / yr to the global air-sea $CO_2$ exchange (Section S2).

### 2.5 Feature Importance

One of the benefits of the XGBoost algorithm is the ability to determine relative contributions by each of the features to the final estimate of $pCO_2$-Residual. This is called feature importance. This tells us the relationships between $pCO_2$-Residual and the input features that have been identified through model training. This supports assessment of the degree to which known physical and biogeochemical mechanisms are embodied in the reconstruction. In other words, this allows us to physically interpret our algorithm. Here we utilize SHapley Additive exPlanations (SHAP) (Shapley, 1953) calculated using the SHAP module in Python (Lundberg et al., 2018), to examine both local and global interpretability of the resulting model.

SHAP computes the contribution of each feature to the final prediction, and solves the game theory problem of relative contributions of players, and therefore fairly distributed payouts, amongst players in cooperative games. In our case, SHAP calculates the importance of each predictor (feature) by starting with the mean values of all features, and the expected value of the $pCO_2$-Residual. For a given month's reconstruction of the $pCO_2$-Residual in a single grid cell, each feature is adjusted one-by-one to the observed value from its mean. As the features are adjusted, the change in the expected value of the $pCO_2$-Residual is calculated, and the difference from the previous expected value is determined. This difference is the feature importance. Since the ordering of the features matters, SHAP computes these attributions for every permutation of feature ordering, and final feature importance is the mean contribution by a given feature to the final reconstruction of the $pCO_2$-Residual, across all ordering permutations.

### 2.6 Independent Datasets

Due to the fact that 80% of the observations contained within the SOCAT database are used to construct the method, and only 20% of the observations remain for testing,

we also examine how well the reconstruction method performs against independent observations not contained within the SOCAT database. We utilize two ocean time series locations: Bermuda Atlantic Time-Series Study (BATS) and Hawaii Ocean Time-Series (HOT). We also examine how well the reconstructed $pCO_2$ compares to observations contained only in the Lamont-Doherty Earth Observatory (LDEO) dataset (data already in SOCAT are removed) and the GLobal Ocean Data Analysis Project version 2 (GLODAPv2). For LDEO, $pCO_2$ is directly measured. For the other datasets, $pCO_2$ is calculated from observations of Total Alkalinity (TA), Dissolved Inorganic Carbon (DIC) and temperature using the PyCO2SYS package in Python (Humphreys et al., 2021). Uncertainties for both directly measured $pCO_2$ and indirectly calculated $pCO_2$ are given in Table 3 of Gloege et al. (2021), and range from 2.5 $\mu$atm in LDEO (directly measured) to >12 $\mu$atm in GLODAPv2 (calculated). Given the known larger biases in some of the other data-based products in the 1980s, we compare to observations within the time frame 1990-2019.

### 2.7 Regression Skill

To compare predicted $pCO_2$ (P) to the observations (O), we examine the correlation (r), bias, root mean squared error (RMSE), mean absolute error (Mean AE), and median absolute error (Median AE). Bias, RMSE, Mean AE, and Median AE measure the size of the error in the predicted $pCO_2$. Bias is calculated as the Mean Prediction - Mean Observation (bias = $\overline{P}-\overline{O}$), and simply indicates whether the regression tends to over- or under-estimate $pCO_2$. A large positive (negative) bias indicates a tendency to overestimate (underestimate) $pCO_2$. However, a bias of small magnitude may be due to large, compensating biases. RMSE measures magnitude of the predicted error, but penalizes larger errors and outliers. It is calculated as the square root of the mean of the squared errors $\sqrt{\overline{(P-O)^2}}$. The Mean AE simply determines the average of the absolute value of the error, treating each error equally. The Median AE is the most common value of the absolute error. The Pearson correlation coefficient (r) measures how much the observations and reconstruction tend to vary together, with values near +1 (-1) indicating a high tendency to vary together (opposite). It is calculated as the covariance between the predictions and the observations, divided by the product of their individual standard deviations.

### 2.8 Arctic and Coastal Zones

The $pCO_2$-Residual product does not reconstruct coastal or Arctic Ocean $pCO_2$, and thus only covers 89.6% of the global ocean. Before air-sea fluxes are calculated, coastal and Arctic regions not reconstructed by the data products must be filled. For consistent comparisons, these coastal areas are filled with the scaled coastal $pCO_2$ climatology (Landschützer et al., 2020) according to Fay et al. (2021) for all data products.

### 2.9 $CO_2$ Flux Calculations

The bulk air-sea CO2 flux (FCO2) is calculated as:

$$FCO_2 = K_w \cdot K0 \cdot (1 - ice_{fraction}) \cdot (pCO_2^{sea} - pCO_2^{atm}) \quad (4)$$

where $K_w$ is the gas-transfer velocity calculated from wind speeds, scaled to the 16.5 cm/hr 14C bomb flux estimate according to Wanninkhof (1992); K0 is the solubility calculated using EN4 salinity and OISST temperatures (Weiss, 1974); icefraction is from the OISST product; $pCO_2^{atm}$ is calculated from NOAA's marine boundary layer product, corrected for water vapor pressure using ERA5 mean sea level pressure; and $pCO_2^{sea}$ is the reconstructed surface ocean $pCO_2$ for a given product. For a consistent comparison $K_w$, K0, ice fraction, and $pCO_2^{atm}$ from SeaFlux are used (Fay et al., 2021). The SeaFlux dataset (Gregor & Fay, 2021) includes $K_w$ for 3 wind speed products: CCMPv2, ERA5, and JRA55. Fluxes presented are the mean flux across the three wind products.

#### *2.9.1 Other Observational-based products*

We compare our reconstruction error statistics and air-sea carbon dioxide flux estimates to those of five other observation-based data products that use machine-learning or statistical modeling (Table 2). The harmonized $pCO_2$ data products and resulting fluxes were obtained from SeaFlux (Gregor & Fay, 2021).

#### *2.9.2 Anthropogenic Carbon Flux*

Data products which incorporate observations of surface ocean $pCO_2$ include both natural and anthropogenic carbon in the resulting $pCO_2$ and $CO_2$ flux product. This is the net $CO_2$ flux ($F_{net} = F_{natural} + F_{ant}$). Global ocean biogeochemical models exclude the natural outgassing of riverine carbon, the dominant driver of the anthropogenic

| Product | Reference |
|---|---|
| CSIR ML6 | Gregor et al. (2019) |
| CMEMS | Denvil-Sommer et al. (2019) |
| HPD | Gloege et al. (2021) |
| MLS | Rodenbeck et al. (2013) |
| MPI-SOMFFN | Landschutzer et al. (2014); Landschutzer, Gruber, and Bakker (2020) |

**Table 2.** Observational data products used for comparison (Gregor & Fay, 2021; Fay et al., 2021)

.

air-sea $CO_2$ flux (Aumont et al., 2001). To quantify the anthropogenic air-sea $CO_2$ flux, the riverine efflux of carbon dioxide must be subtracted from our net flux. Quantifying the global air-sea $CO_2$ flux due to decomposition and outgassing of riverine carbon is itself a complex scientific problem, one that is still being worked on. Here, as in Gloege et al. (2021), we use an average of three estimates: Jacbobson et al. (2007): (0.45 +/- 0.18 PgC/yr), Resplandy et al. (2018): (0.78 +/- 0.41 PgC/yr), and Lacroix et al. (2020): (0.23 Pg C / yr). The combined estimated efflux due to riverine carbon is 0.49 +/- 0.26 Pg C/yr, and we remove the efflux of 0.49 PgC/yr from the estimated annual air-sea $CO_2$ fluxes calculated using the Residual and other data products' $pCO_2$ .

## 3 Results

### 3.1 Model Skill

The $pCO_2$-Residual approach is an ensemble of five reconstructions. The test statistics for $pCO_2$ for each of the five reconstructions and their mean are shown in Table 2. We have a mean test RMSE of 16.33$\mu$atm, lower than the recent data product of Gregor et al. (2019) (17.16 $\mu$atm). Each run has a relatively small bias and is highly correlated with the test observations. The Mean Absolute Error (Mean AE) is near 11 $\mu$atm, and the Median Absolute Error (Median AE) is less than 8 $\mu$atm. For the ensemble, RMSE is lowest (below 10 $\mu$atm) in the subtropical regions as we would expect, and higher in the equatorial Pacific, Southern Ocean, and subpolar North Atlantic and subpolar North

| | **Run 1** | **Run 2** | **Run 3** | **Run 4** | **Run 5** | **Mean** |
|---|---|---|---|---|---|---|
| **RMSE ($\mu$atm)** | 16.13 | 16.02 | 16.76 | 16.51 | 16.25 | 16.33 |
| **Bias ($\mu$atm)** | 0.28 | 0.50 | -0.21 | 0.61 | -0.30 | 0.18 |
| **Correlation** | 0.89 | 0.90 | 0.88 | 0.89 | 0.89 | 0.89 |
| **Mean AE ($\mu$atm)** | 10.88 | 10.87 | 11.20 | 11.13 | 10.92 | 11.00 |
| **Median AE ($\mu$atm)** | 7.41 | 7.49 | 7.57 | 7.68 | 7.46 | 7.52 |

**Table 3.** $pCO_2$ Test statistics for each of the five ensemble members, and their mean values.

Pacific (not shown). The ensemble model bias and RMSE are stable over time, with no clear trends. Previous techniques have exhibited a higher bias in the 1980s (Gregor et al., 2019).

The technique was also examined within the Large Ensemble Testbed (Gloege, McKinley, et al., 2021) and showed a decrease in RMSE as compared to reconstructing $pCO_2$ without the knowledge of direct temperature effects, both for test data and in extrapolation to where we have no observations for comparison outside of the model world (Section S1).

### 3.2 Evaluation against Independent Data

We examine the approach's ability to reconstruct surface ocean $pCO_2$ in data sets not contained within the SOCAT data. At the ocean timeseries sites Hawaii (HOT) and Bermuda (BATS), reconstructed surface ocean $pCO_2$ is highly correlated with observations (Figure 3). This is true for all data products shown, as seasonality is well captured in these subtropical regions (Rödenbeck et al., 2015; Gloege, Yan, et al., 2021). The $pCO_2$-Residual technique is amongst the most highly correlated at both stations and is also amongst the best three at capturing the variability (Figure 3a,b).

GLODAP and LDEO are observations taken along ship transects traveled irregularly. As the data are not located at repeat stations, the correlations are lower, because they represent the spatial patterns of observations as well as temporal variability and change. Again, the $pCO_2$-Residual technique is amongst the top performing observation-based data products, with high correlations. It underestimates the amplitude of observed

variability, as do all techniques except JENA-MLS (Figure 3c,d). Compared to the other observation-based data products, the unbiased RMSE is approximately equal to that of the LDEO-HPD technique, which is the best-performing gap-filling technique compared to these data (Gloege, Yan, et al., 2021).

### 3.3 Physical Mechanisms

Machine learning algorithms are often thought of as black boxes, but the XGB algorithm allows us to dig into that "black box" and examine the relative contributions of features to the model prediction. The first column in Figure 4 shows the mean (1982-2019) importance of mixed layer depth; geographic location and day of year; SST; and Chl-a to the model's prediction of the $pCO_2$-Residual. These are the dominant controls of the seasonal cycle of the $pCO_2$-Residual within the algorithm. Here, we sum the importance of geographic location and day of year (D.O.Y.), because there is no seasonal cycle in location, but there is geographic variation of the impact of day of year on the $pCO_2$-Residual. The second column of Figure 4 examines the mean seasonal cycles of feature importance for each of these predictors for four biomes of Fay and McKinley (2014). The third column of Figure 4 shows the contributions of interannually varying predictors to the reconstructed $pCO_2$-Residual.

The seasonal cycle of the $pCO_2$-Residual is largely controlled by mixed layer depth, which has large mean feature importance (Figure 4a), but also large seasonal variations away from the equator (Figure 4e-h). Deep winter mixing brings up dissolved inorganic carbon (DIC) and increases $pCO_2$, whereas shallower mixed layer depths set up biological production and a decrease in surface DIC. During northern hemisphere winter (DJF), the algorithm's estimate of the $pCO_2$-Residual is significantly increased (decreased) by mixed layer depth in the northern (southern) hemisphere as expected. There is a small seasonal cycle in the feature importance of MLD along the equator. The geographic location and day of year significantly increases the $pCO_2$-Residual on the mean in the equatorial zones and decreases the $pCO_2$-Residual in the Southern Ocean (Figure 4b). We see the small mean impact of these combined features in the subpolar northern regions is the balanced effect of significant seasonal variations in its importance to the reconstructed $pCO_2$-Residual (Figure 4d-f).

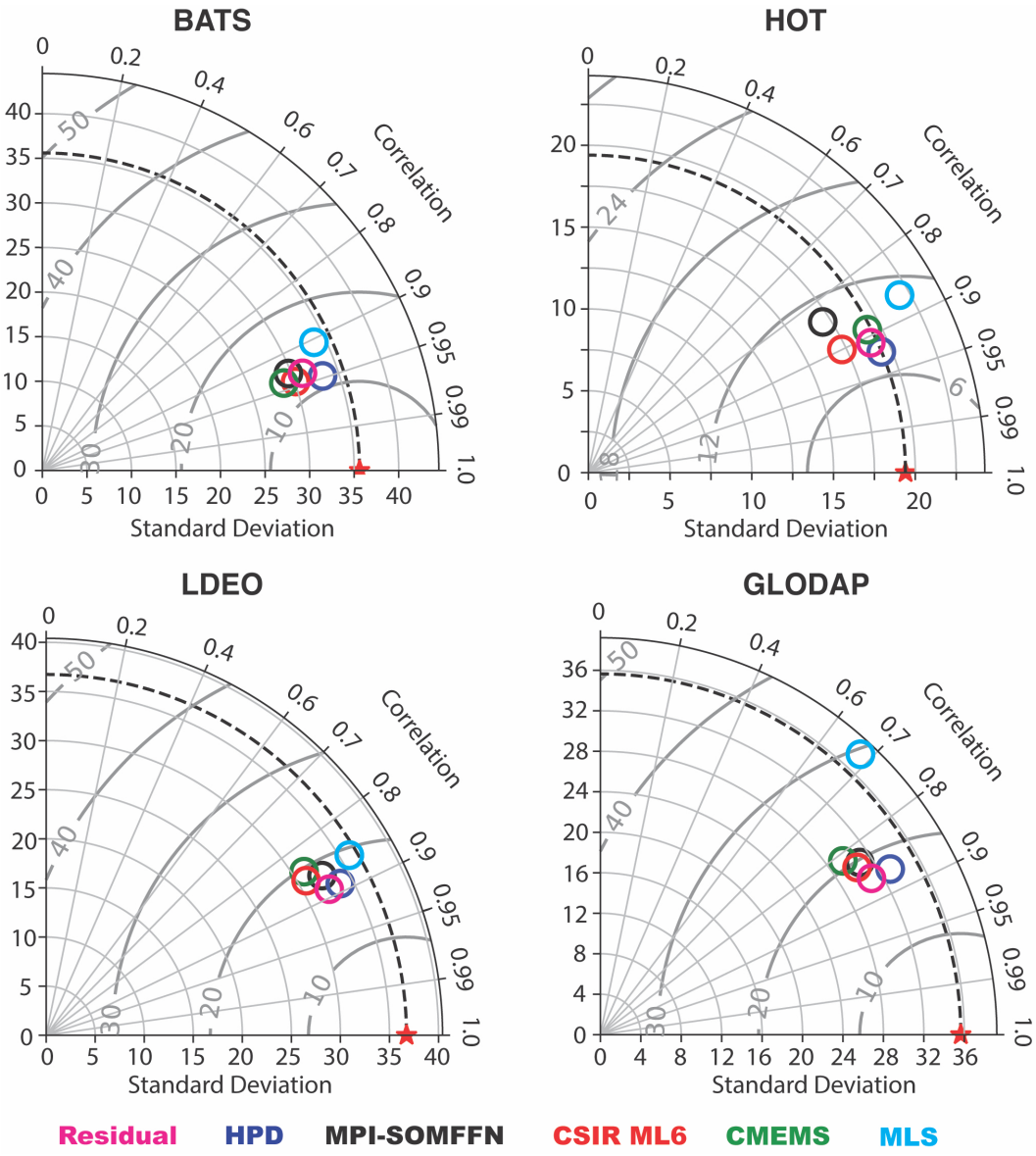


**Figure 3.** Taylor diagrams (Taylor, 2001) of Correlation (along circumference), Standard Deviation (along radii), and RMSE (grey arcs centered at red star) of 5 previous data-based approaches (HPD in blue, MPI-SOMFFN in black, CSIR ML6 in red, CMEMS in green, and MLS in cyan) and the new $pCO_2$-Residual technique (magenta). BATS is shown at the top left, HOT at the top right, LDEO at bottom left, and GLODAP at bottom right.

While the direct impacts of SST are contained within the technique itself, Figure 4g-i show the importance of SST to the reconstructed residual. Summertime stratification can set up biological production, and wintertime deep mixing can bring up older, remineralized dissolved inorganic carbon (DIC). We see that on the mean, the algorithm decreases its estimate of the residual in the warm equatorial regions and along the Gulf Stream, and increases its estimate in colder zones (Figure 4c). Small seasonal variations around this mean impact exist, with decreases in the residual seen in the summer hemisphere (Figure 4e,g,h). Examining the feature importance of chlorophyll on the mean, we see that Chl-a has a small impact on the model's prediction (Figure 4d); in other words, the magnitudes of the adjustments made because of Chl-a values are smaller than other features. We do see, however, a negative adjustment to the residual estimate at times and regions of strong biological production (Figure 4g) and smaller positive adjustments made in less productive regions (Figure 4e,f,h), or outside of the summer season.

The third column of Figure 4 examines the year-to-year variations in feature importance for those features with year-to-year changes (SST anomalies, $xCO_2$, and Chl-a anomalies). Interannual anomalies in Chl-a do not cause significant adjustments to the residual for that year in any of the biomes (Figure 4i-l). However, interannual anomalies in SST do cause significant adjustments to the predicted $pCO_2$-Residual, particularly in the eastern equatorial Pacific. Additionally, if we examine how $xCO_2$ is used to adjust the initial guess of the $pCO_2$-Residual in our algorithm, we see that low $xCO_2$ during the early years of the reconstruction translates to a negative adjustment (decrease) in the $pCO_2$-Residual (Figure 4i-l). As the years progress, this contribution increases and becomes positive and large by the later years of the reconstruction. This is expected, as the ocean $pCO_2$ increases following atmospheric $pCO_2$. $pCO_2$-T does not account for the long term trend in $pCO_2$ since this is caused by the accumulation of DIC. The algorithm must learn why there is an increase in the $pCO_2$-Residual over time, and as shown here, it correctly attributes this increase to $xCO_2$. Within the algorithm, interannual variability in the reconstructed $pCO_2$ Residual is largely controlled by interannual anomalies in SST in all regions. The contribution of the atmospheric $CO_2$ mixing ratio ($xCO_2$) in the $pCO_2$-Residual prediction is homogenous in space (not shown), which distinguishes it from the spatially variable impacts of SST, MLD, and Chl-a (Figure 4). This is as expected because a single global-mean atmospheric $xCO_2$ timeseries is used as a feature for all spatial points.

This analysis demonstrates that the XGBoost algorithm allows an additional layer of understanding to our $pCO_2$ reconstruction. Mixed layer depths, geographic location, time of year, and to a lesser extent, SST and Chl-a control the seasonal cycle of the $pCO_2$-Residual within the algorithm. The long-term $pCO_2$ trend is due to the trend in atmospheric $CO_2$, and year-to-year variations are dominantly driven by SST.

### 3.4 Uncertainty

To quantify uncertainty in our $pCO_2$ reconstruction, a quantile loss function is employed within the XGBoost regression. To do this, a custom evaluation function and loss function are provided to XGBoost as parameters. Random noise is added to the smoothed gradient to improve the performance of XGBoost with quantile loss (Descamps, 2020). Most machine learning loss functions aim to reduce the mean absolute error between the predicted value and the observation. The quantile loss function, however, is used to predict a specified quantile of the prediction, and the loss function is minimized when the reconstruction resides at a given quantile. A quantile is a value below which a fraction of observations lies. Thus, the 90% quantile for $pCO_2$ will over-estimate the observed $pCO_2$ 90% of the time. We reconstruct the 5% quantile and the 95% quantile such that we are confident the true surface ocean $pCO_2$ value lies between these reconstructions approximately 90% of the time. Thus, for a given point in space and time, the reconstructed $pCO_2$ can be quantified with 90% confidence as:

$$pCO_2\ 90\%\ CI = pCO_2 \pm \frac{(pCO_2^{95th} - pCO_2^{5th})}{2} \tag{5}$$

Figure 5 displays the mean value (1985-2019) of the second half of Equation 5, the value added and subtracted from the $pCO_2$ reconstruction to create confidence bounds. We show the magnitude of uncertainty for both the 90% (Figure 5a) and 67% (Figure 5b) confidence bounds. Confidence is highest, with lowest uncertainties within the subtropical oceans (+/- less than 10 $\mu$atm at 67% confidence). Uncertainties become larger within the subpolar regions, and largest within the Southern Ocean and within the equatorial Pacific. The algorithm cannot identify whether the uncertainty arises because of a lack of measurements of surface ocean $pCO_2$ or from noise in the observations. However, uncertainty is largest in regions that are biologically productive, which could be substantial impacted by uncertainty of 30% for Chl-a observations, and highly dynamic regions such as eastern upwelling zones. Uncertainty also increases where there are few obser-

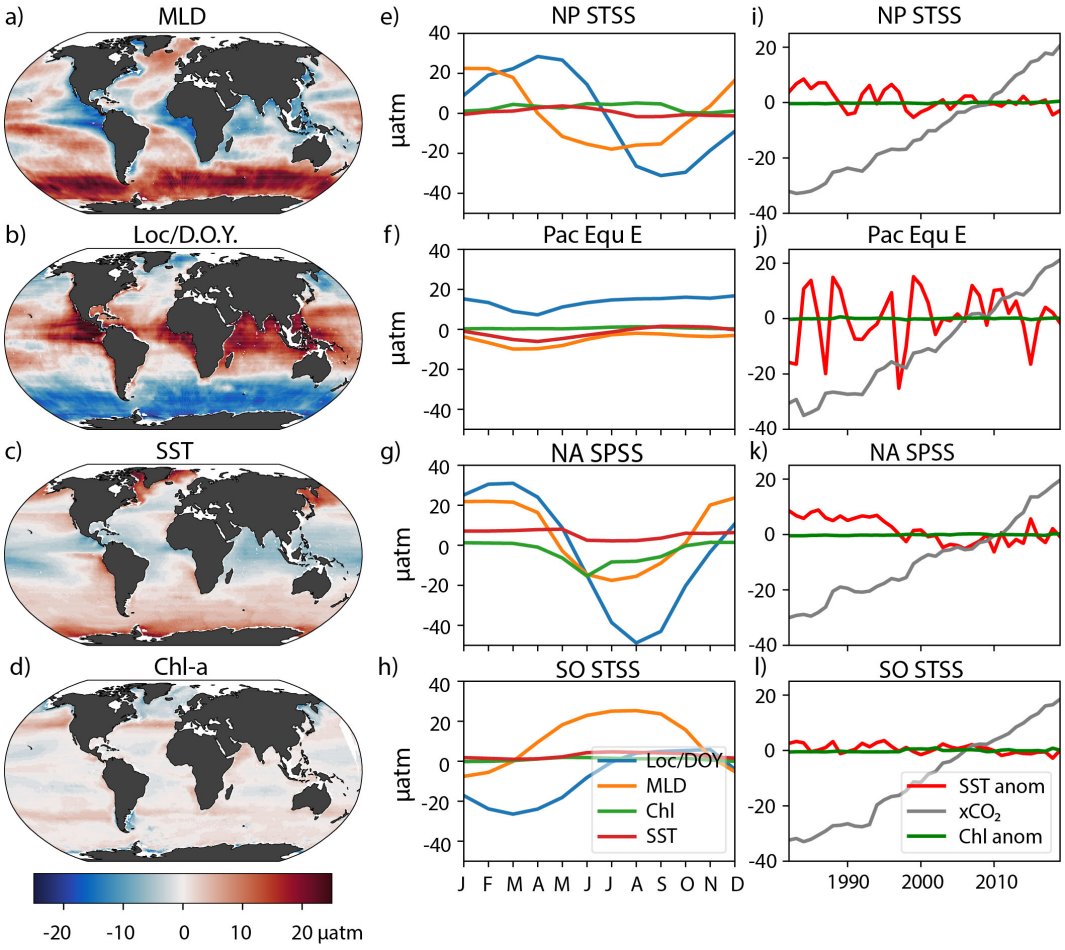


**Figure 4.** Mean feature importance and seasonal cycles and interannual variations in feature importance in sample biomes ($\mu$atm). (a) Mean feature importance of mixed layer depth. (b) Mean feature importance of Location and Day of Year (D.O.Y.). (c) Mean feature importance of SST. (d) Mean feature importance of chlorophyll-a. (e) Mean seasonal cycles of feature importance of Location/D.O.Y., MLD, Chl, and SST in the NP STSS (North Pacific Subtropical Seasonally Stratified) biome. (f) Same as in (e) except for the Pac Equ E biome. (g) Same as in (e) except for the NA SPSS (North Atlantic Subpolar Seasonally Stratified) biome. (h) Same as in (e) except for the SO STSS biome (Southern Ocean Subtropical Seasonally Stratified). (i) Interannual variations in feature importance for SST, chlorophyll-a, and $xCO_2$ within the NP STSS biome. (j) Same as in (i) except for within the Pac Equ E (eastern Equatorial Pacific) biome. (k) Same as in (i) except for within the NA SPSS biome. (l) Same as in (i) except for within the SO STSS biome.

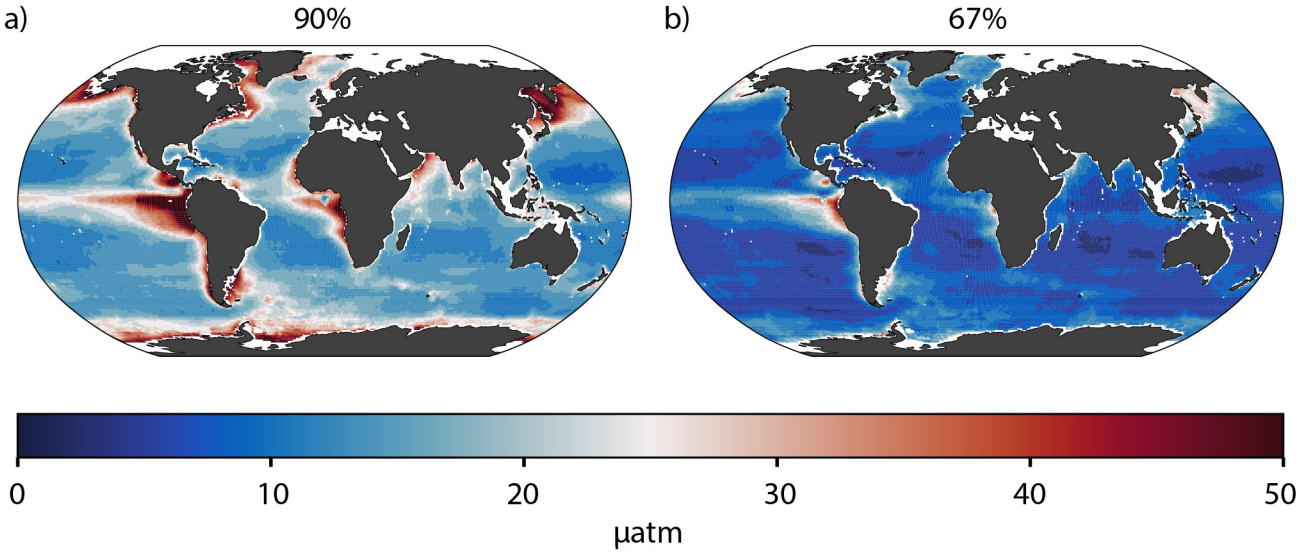


**Figure 5.** Mean $pCO_2$ uncertainty within the 90% (a) and 67% (b) confidence bounds. At a given location, the shading represents the mean value that would be added and subtracted to form the confidence interval of reconstructed $pCO_2$.

vations (off the southwestern tip of South America and within the Indian Ocean, for instance).

### 3.5 $CO_2$ Fluxes

The mean air-sea $CO_2$ fluxes reconstructed using the $pCO_2$-Residual technique for 1985-2019 exhibit known features (Figure 6a). The subpolar North Atlantic is a strong carbon sink, while the equatorial regions efflux carbon dioxide to the atmosphere. Subtropical regions are smaller carbon sinks, and the high latitude Southern Ocean and North Pacific are sources of carbon to the atmosphere. The globally-integrated anthropogenic air-sea $CO_2$ flux has become increasingly more negative, as atmospheric $CO_2$ concentrations have increased. Using the same coastal filling and river correction for all products, we find that the $CO_2$ sink reconstructed by the $pCO_2$-Residual approach is consistent with the other data products (Figure 6b). Year-to-year variability in the air-sea $CO_2$ flux is largest in the reconstructions using the JENA MLS and $pCO_2$-Residual approaches.

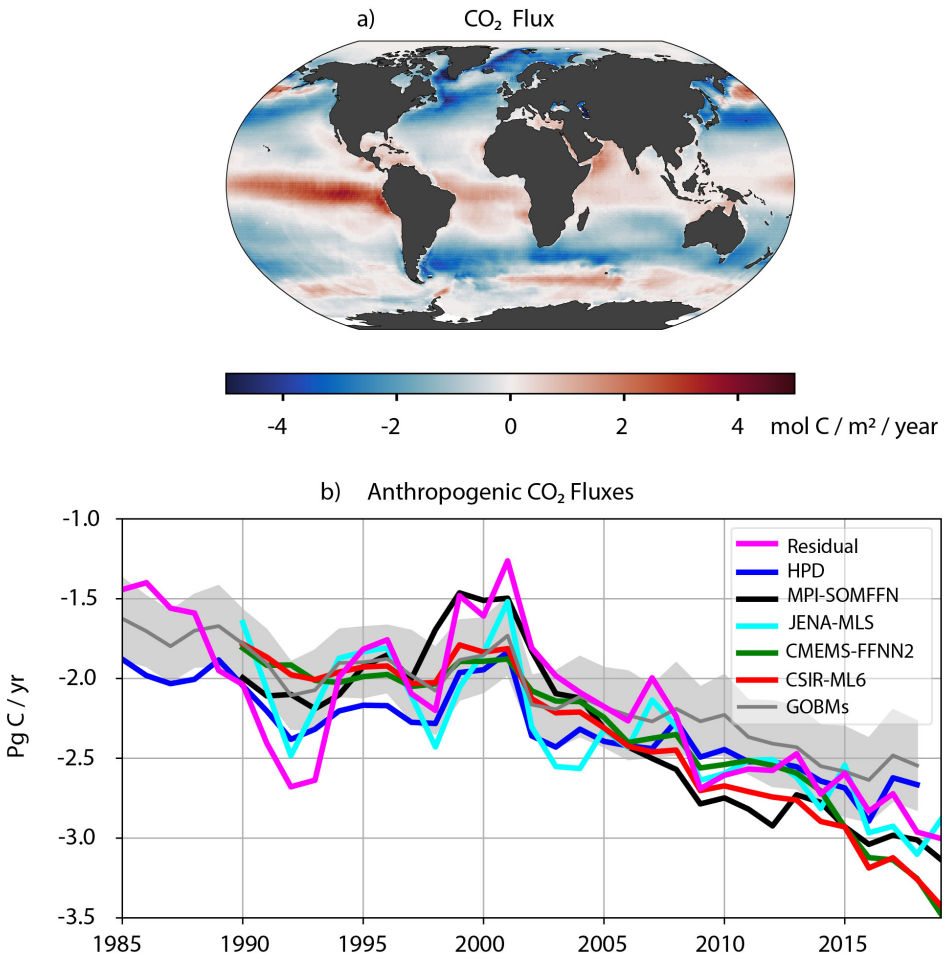


**Figure 6.** (a) Map of mean (1985-2019) anthropogenic air-sea $CO_2$ Flux reconstructed by the $pCO_2$-Residual Technique. (b) Annual mean (1985-2019) air-sea $CO_2$ fluxes estimated by the $pCO_2$-Residual (magenta), HPD (blue), MPI-SOMFFN (black), CSIR ML6 (red), CMEMS (green), and MLS (cyan) data products. Mean of the 9 GOBMs and one standard deviation shading in grey. Harmonized observation-based data products begin in 1990 (Gregor & Fay, 2021).

### 3.6 Uncertainty in $CO_2$ Fluxes

In order to determine the uncertainty in $CO_2$ flux caused by our uncertainty in surface ocean $pCO_2$, we assume zero bias in the reconstruction. This assumption is supported by the analysis of (Gloege, Yan, et al., 2021) and our own analysis with the Large Ensemble Testbed (Section S1). Using a Monte Carlo approach, we randomly sample $pCO_2$ from a normal distribution with mean values equal to our locally reconstructed $pCO_2$ and standard deviation provided by the quantile loss reconstruction. We randomly sample every $1^o$ by $1^o$ grid box 500 times for every month and wind product, and then calculate the local and global air-sea fluxes. Figure 7a shows the resulting mean annual standard deviation of the air-sea flux from the Monte Carlo approach. While the pattern of flux uncertainty grossly mimics the $pCO_2$ uncertainty pattern, there are differences. The largest flux uncertainties are not seen where the $pCO_2$ uncertainties are largest, such as the equatorial Pacific. Instead, the largest flux uncertainties are seen where there are moderate $pCO_2$ uncertainties (Figure 5) and significant piston velocities (Figure 7b). Here, even moderate uncertainties in $pCO_2$ translate into larger air-sea flux uncertainties than the equatorial Pacific, where large $pCO_2$ uncertainties are dampened by much smaller piston velocities.

Figure 7c shows the mean of the zonally integrated $CO_2$ flux for the three wind products CCMP2, ERA5, and JRA55 (blue, orange, and green, respectively) as compared to the zonally integrated uncertainty, as one standard deviation of the zonally integrated flux from the Monte Carlo simulations (CCMP2, ERA5, and JRA55 as blue, orange, and green, respectively). While local standard deviations are a significant portion of the mean flux in some regions (e.g. subtropical North Atlantic), without a bias in the reconstruction, the reconstructed global air-sea flux has very small uncertainties caused by the uncertainty in $pCO_2$ ( 0.01 PgC/yr). However, uncertainties in the piston velocities estimated by different wind products cause a standard deviation of annual fluxes of 0.04 - 0.10 Pg C/yr (not shown). Therefore, we estimate a total uncertainty of 0.11 Pg C / yr, one standard deviation, for the 67% confidence interval.

## 4 Discussion

We show that a physically realistic algorithm results when we incorporate physical knowledge into a data based machine learning approach. By reconstructing the dif-

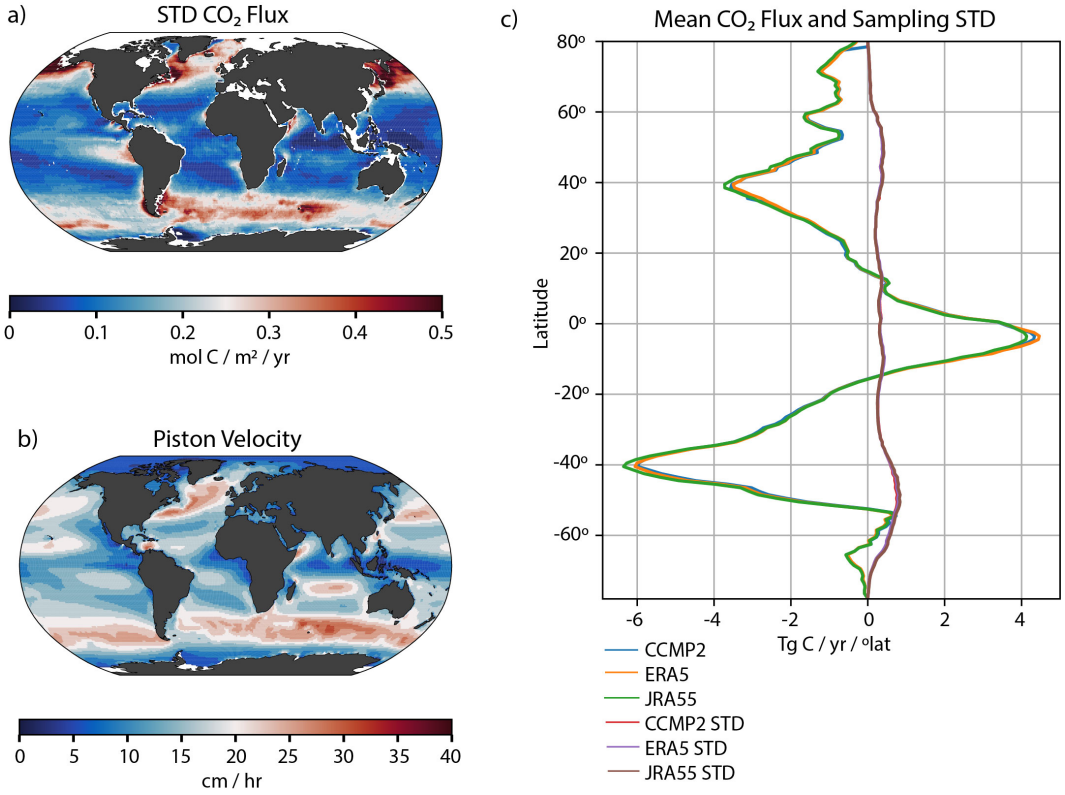


**Figure 7.** Uncertainty in $CO_2$ fluxes. (a) Standard deviation of annual $CO_2$ flux from Monte Carlo simulations (mol C / $m^2$ / yr). (b) Mean piston velocity, average of CCMP2, ERA5 and JRA55 (Kw: cm/hr). (c) Mean $CO_2$ flux by latitude band (Tg C / yr) and wind product, and standard deviation of the mean flux caused by random sampling of $pCO_2$ for each wind product.

ference between observed $pCO_2$ and the $pCO_2$ that would result if only the direct effect of temperature altered surface ocean $pCO_2$ (Figures 1,2), the $pCO_2$ Residual approach requires the machine learning algorithm to learn only the indirect effects of temperature on $pCO_2$. This residual is small within the temperature-controlled subtropical regions and larger in more dynamic ocean areas (Figure 2). This approach tackles two of the five major barriers to adoption of machine learning approaches within the geosciences proposed by Reichstein et al. (2019): interpretability and physical consistency. Within the resulting model, mixed layer depth, location, season, SST, Chl-a, and $xCO_2$ impact the $pCO_2$-Residual as we would expect (Figures 4,5), building confidence in the approach. MLD, location, and time of year strongly control the seasonal cycles of reconstructed $pCO_2$, while atmospheric $CO_2$ concentrations and interannual variations in SST control the interannual variations in reconstructed $pCO_2$. We find that year-to-year variations in chlorophyll-a are not found to drive variability in reconstructed $pCO_2$. This may be, in part, due to the small interannual variations in observed Chl-a, as observed in the North Atlantic (Bennington et al., 2009), but is likely also due to noise in the observations.

The LDEO-HPD approach incorporates physical knowledge of the system by using GOBMs as a first guess. XGBoost is used to reconstruct the model-observation discrepancy to reconstruct the full $pCO_2$ field. That approach must rely upon both observations and models to create a reconstruction. Any alterations to the model output would require the development of a new regression.

This technique's reconstructed $pCO_2$ has small RMSE and high correlations when compared to independent observations, and is one of the best performing observation-based approaches based on comparison to four independent datasets (Figure 3). Uncertainties in reconstructed surface ocean $pCO_2$ due to the algorithm are smallest in the subtropical ocean regions and largest in the equatorial Pacific and subpolar regions (Figure 6), as would be expected in the technique. The pattern of uncertainty in $pCO_2$ is the same as the pattern of test RMSE (not shown), and the magnitude of the global mean test RMSE (16.33 $\mu$atm) lies between the global mean uncertainty magnitude at the 67% confidence interval (9.8 $\mu$atm) and the 90% confidence interval (19.71 $\mu$atm). The resulting air-sea $CO_2$ fluxes are in agreement with previous data-based approaches (Figure 6), and exhibit high interannual variability, similar to MLS inversion approach (Rödenbeck et al., 2013). This may be due to the use of the tree-based XGBoost algorithm, as op-

posed to a neural network in which non-linearities are controlled by the activation function (Baughman & Liu, 1995).

Uncertainty in the resulting air-sea $CO_2$ fluxes, as determined by the Monte Carlo approach, are largest where both piston velocities and $pCO_2$ uncertainty are larger (Figure 7). Although there are regions of significant local uncertainty in the flux, the uncertainty in the globally-integrated air-sea $CO_2$ flux due to random error in $pCO_2$ remains minimal (0.01 Pg C / yr). However, uncertainty due to uncertainty in the piston velocity is larger, with annual flux uncertainty ranging from 0.04 to 0.1 PgC/yr, for a total uncertainty bound of 0.11 PgC/yr. This finding is in agreement with previous work that suggest the largest uncertainties may due to amplification of $pCO_2$ uncertainties by winds (Landschützer et al., 2016; Gregor et al., 2019). Here, we have assumed no bias in the observations (Gloege, Yan, et al., 2021; Fay et al., 2021). However, if observational bias exists in any region with moderate to high piston velocities, the uncertainty would be significantly larger, as regional $pCO_2$ would change in concert. The importance of systematic bias will be explored in future work.

While (Gregor et al., 2019) suggest ocean surface $pCO_2$ reconstructions may have "hit a wall", here we illustrate there are more techniques to consider. The LDEO-HPD (Gloege, Yan, et al., 2021) and the $pCO_2$-Residual technique, both which include physical knowledge, are the two best performers compared to 3 of the 4 independent datasets (Figure 3), suggesting incremental improvements are still possible, even without a significant increase in observations. Here we show that it is possible to incorporate physical knowledge within a data-only technique. We have confidence in the technique not only from comparisons to independent observations, but also from a testbed based in Earth System Models (Section S1). Such interpretable techniques should allow for better integration across differing approaches: numerical modeling, observations, and machine learning (Reichstein et al., 2019).

The Global Ocean Carbon Budget 2021 (Friedlingstein et al., 2021) estimates an anthropogenic ocean carbon sink of -2.5 $\pm$ 0.4 PgC/yr for the period 2000-2020. The Residual technique suggests a similar flux of -2.35 $\pm$ 0.5 PgC/yr.

## 5 Conclusions

We develop a new machine learning approach to reconstruct global ocean $pCO_2$, an approach that incorporates physical knowledge of the ocean carbonate system within a purely data based approach. The $pCO_2$ Residual approach improves upon previous machine learning approaches by removing the direct effect of temperature from the algorithm. The resulting model created using an XGBoost algorithm exhibits realistic physical processes and suggests an air-sea exchange of carbon dioxide within the range of previous data-based approaches and in agreement with the Global Carbon Budget 2021 (Friedlingstein et al., 2021). The approach will be used to further examine reconstruction uncertainties.

## 6 Data Availability

NOAA High Resolution SST data provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their Web site at `https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.highres.html`. Python scripts are made available at `https://github.com/valbennington/JAMES_pub_2022`.

**Acknowledgments**

The authors acknowledge support from NOAA (NA20OAR4310340) and the Data Science Institute of Columbia University. We that all data providers and quality controllers who work tirelessly to create the SOCAT database.

## References

Aumont, O., Orr, J. C., Monfray, P., Ludwig, W., Amiotte-Suchet, P., & Probst, J.-L. (2001). Riverine-driven interhemispheric transport of carbon. *Global Biogeochemical Cycles*, *15*(2), 393-405. doi: https://doi.org/10.1029/1999GB001238

Bakker, D. C. E., Pfeil, B., Landa, C. S., Metzl, N., O'Brien, K. M., Olsen, A., ... Xu, S. (2016). A multi-decade record of high-quality $fco_2$ data in version 3 of the surface ocean $co_2$ atlas (socat). *Earth System Science Data*, *8*(2), 383–413. doi: 10.5194/essd-8-383-2016

Baughman, D., & Liu, Y. (1995). 2 - fundamental and practical aspects of neural computing. In D. Baughman & Y. Liu (Eds.), *Neural networks in bioprocessing and chemical engineering* (p. 21-109). Boston: Academic Press. doi: https://doi.org/10.1016/B978-0-12-083030-5.50008-4

Bennington, V., McKinley, G. A., Dutkiewicz, S., & Ullman, D. (2009). What does chlorophyll variability tell us about export and air-sea co2 flux variability in the north atlantic? *Global Biogeochemical Cycles*, *23*(3).

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, *abs/1603.02754*.

de Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A., & Iudicone, D. (2004). Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. *Journal of Geophysical Research: Oceans*, *109*(C12). doi: https://doi.org/10.1029/2004JC002378

Denvil-Sommer, A., Gehlen, M., Vrac, M., & Mejia, C. (2019). Lsce-ffnn-v1: a

two-step neural network model for the reconstruction of surface ocean pco2 over the global ocean. *Geoscientific Model Development*, *12*, 2091–2105. doi: https://doi.org/10.5194/gmd-12-2091-2019

Descamps, B. (2020, Aug). *Regression prediction intervals with xgboost.* Retrieved from `https://towardsdatascience.com/regression-prediction-intervals-with-xgboost-428e0a018b`

Dickson, A. G., Sabine, C. L., & Christian, J. R. (2007). *Guide to best practices for ocean co2 measurements.* North Pacific Marine Science Organization.

Fay, A. R., Gregor, L., Landschützer, P., McKinley, G. A., Gruber, N., Gehlen, M., ... Zeng, J. (2021). Seaflux: harmonization of air–sea $co_2$ fluxes from surface $pco_2$ data products using a standardized approach. *Earth System Science Data*, *13*(10), 4693–4710. doi: 10.5194/essd-13-4693-2021

Fay, A. R., & McKinley, G. A. (2014). Global open-ocean biomes: mean and temporal variability. *Earth System Science Data*, *6*, 273–284. doi: doi:10.5194/essd-6-273-2014

Fay, A. R., & McKinley, G. A. (2021). Observed regional fluxes to constrain modeled estimates of the ocean carbon sink. *Geophysical Research Letters*, e2021GL095325.

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C., Hauck, J., ... others (2021). Global carbon budget 2021. *Earth System Science Data Discussions*, 1–191.

Gloege, L., McKinley, G. A., Landschützer, P., Fay, A. R., Frölicher, T. L., Fyfe, J. C., ... Takano, Y. (2021). Quantifying errors in observationally based estimates of ocean carbon sink variability. *Global Biogeochemical Cycles*, *35*(4), e2020GB006788. (e2020GB006788 2020GB006788) doi: https://doi.org/10.1029/2020GB006788

Gloege, L., Yan, M., Zheng, T., & McKinley, G. A. (2021). Improved quantification of ocean carbon uptake by using machine learning to merge global models and pco2 data. *Journal of Advances in Modeling Earth Systems*, *in press*.

Good, S. A., Martin, M. J., & Rayner, N. A. (2013). En4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans*, *118*(12), 6704-6716. doi: https://doi.org/10.1002/2013JC009067

Gregor, L., & Fay, A. (2021, Jul). *Seaflux: harmonised sea-air co2 fluxes from surface pco2 data products using a standardised approach.* Zenodo. doi: 10.5281/zenodo.5482547

Gregor, L., Lebehot, A. D., Kok, S., & Monteiro, P. M. S. (2019). A comparative assessment of the uncertainties of global surface ocean co2 estimates using a machine-learning ensemble (csir-ml6 version 2019a) – have we hit the wall? *Geoscientific Model Development*, *12*, 5113–5136. doi: https://doi.org/10.5194/gmd-12-5113-2019

Hauck, J., Zeising, M., Le Quéré, C., Gruber, N., Bakker, D. C. E., Bopp, L., ... Séférian, R. (2020, Oct 27). Consistency and challenges in the ocean carbon sink estimate for the global carbon budget. *Frontiers in Marine Science*.

Humphreys, M. P., Lewis, E. R., Sharp, J. D., & Pierrot, D. (2021). Pyco2sys v1.7: marine carbonate system calculations in python. *Geoscientific Model Development Discussions*, *2021*, 1–45. Retrieved from `https://gmd.copernicus.org/preprints/gmd-2021-159/` doi: 10.5194/gmd-2021-159

Jacobson, A. R., Mikaloff Fletcher, S. E., Gruber, N., Sarmiento, J. L., & Gloor, M. (2007). A joint atmosphere-ocean inversion for surface fluxes of carbon dioxide: 1. methods and global-scale fluxes. *Global Biogeochemical Cycles*, *21*(1). doi: https://doi.org/10.1029/2005GB002556

Khatiwala, S., Tanhua, T., Fletcher, S. M., Gerber, M., Doney, S. C., Graven, H. D., ... Sabine, C. L. (2013). Global ocean storage of anthropogenic carbon. *Biogeosciences*, *10*, 2169.

Lacroix, F., Ilyina, T., & Hartmann, J. (2020). Oceanic co2 outgassing and biological production hotspots induced by pre-industrial river loads of nutrients and carbon in a global modeling approach. *Biogeosciences*, *17*(1), 55-88.

Landschützer, P., Gruber, N., & Bakker, D. C. E. (2016). Decadal variations and trends of the global ocean carbon sink. *Global Biogeochemical Cycles*, *30*(10), 1396-1417. doi: https://doi.org/10.1002/2015GB005359

Landschützer, P., Gruber, N., Bakker, D. C. E., & Schuster, U. (2014). Recent variability of the global ocean carbon sink. *Global Biogeochemical Cycles*, *28*(9), 927-949. doi: https://doi.org/10.1002/2014GB004853

Landschützer, P., Laruelle, G. G., Roobaert, A., & Regnier, P. (2020). A uniform pco2 climatology combining open and coastal oceans. *Earth System Science*

*Data*, *12*(4), 2537-2553.

Lundberg, S. M., Erion, G. G., & Lee, S. (2018). Consistent individualized feature attribution for tree ensembles. *CoRR*, *abs/1802.03888*. Retrieved from `http://arxiv.org/abs/1802.03888`

Maritorena, S., d'Andon, O. H. F., Mangin, A., & Siegel, D. A. (2010). Merged satellite ocean color data products using a bio-optical model: Characteristics, benefits and issues. *Remote Sensing of Environment*, *114*(8), 1791-1804. doi: https://doi.org/10.1016/j.rse.2010.04.002

Masarie, K. A. (2012). *Islscp ii globalview: Atmospheric co2 concentrations.* ORNL Distributed Active Archive Center. doi: 10.3334/ORNLDAAC/1111

McKinley, G. A., Pilcher, D. J., Fay, A. R., Lindsay, K., Long, M. C., & Lovenduski, N. S. (2016). Timescales for detection of trends in the ocean carbon sink. *Nature*, *530*, 469+. doi: doi:10.1038/nature16958

Peters, G. P., Le Quéré, C., Andrew, R. M., Canadell, J. G., Friedlingstein, P., Ilyina, T., ... others (2017). Towards real-time verification of co 2 emissions. *Nature Climate Change*, *7*(12), 848–850.

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., ... Kumar, V. (2019). Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, *55*(11), 9173-9190. doi: https://doi.org/10.1029/2019WR024922

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, *566*(7743), 195–204.

Resplandy, L., Keeling, R. F., Rödenbeck, C., Stephens, B. B., Khatiwala, S., Rodgers, K. B., ... Tans, P. P. (2018, 07). Revision of global carbon fluxes based on a reassessment of oceanic and riverine carbon transport. *Nature Geoscience*, *11*(7), 504-509.

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., & Wang, W. (2002). An improved in situ and satellite sst analysis for climate. *Journal of Climate*, *15*(13), 1609–1625.

Rödenbeck, C., Bakker, D. C. E., Gruber, N., Iida, Y., Jacobson, A. R., Jones, S., ... Zeng, J. (2015). Data-based estimates of the ocean carbon sink variability – first results of the surface ocean pco2 mapping intercomparison (socom).

*Biogeosciences*, *12*(23), 7251-7278.

Rödenbeck, C., Keeling, R. F., Bakker, D. C. E., Metzl, N., Olsen, A., Sabine, C., & Heimann, M. (2013). Global surface-ocean pco2 and sea-air co2 flux variability from an observation-driven ocean mixed-layer scheme. *Ocean Science*, *9*(2), 193 - 216.

Sabine, C. L., Feely, R. A., Gruber, N., Key, R. M., Lee, K., Bullister, J. L., ... Rios, A. F. (2004). The oceanic sink for anthropogenic co2. *Science*, *305*(5682), 367–371.

Shapley, L. (1953). A value fo n-person games. *Ann. Math. Study28, Contributions to the Theory of Games, ed. by HW Kuhn, and AW Tucker*, 307–317.

Takahashi, T., Sutherland, S. C., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., ... Nojiri, Y. (2002). Global sea–air co2 flux based on climatological surface ocean pco2, and seasonal biological and temperature effects. *Deep Sea Research Part II: Topical Studies in Oceanography*, *49*(9), 1601-1622. (The Southern Ocean I: Climatic Changes in the Cycle of Carbon in the Southern Ocean) doi: https://doi.org/10.1016/S0967-0645(02)00003-6

Takahashi, T., Sutherland, S. C., Wanninkhof, R., Sweeney, C., Feely, R. A., Chipman, D. W., ... de Baar, H. J. (2009). Climatological mean and decadal change in surface ocean pco2, and net sea–air co2 flux over the global oceans. *Deep Sea Research Part II: Topical Studies in Oceanography*, *56*(8), 554-577. (Surface Ocean CO2 Variability and Vulnerabilities) doi: https://doi.org/10.1016/j.dsr2.2008.12.009

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183-7192. doi: https://doi.org/10.1029/2000JD900719

Wanninkhof, R. (1992). Relationship between wind speed and gas exchange over the ocean. *Journal of Geophysical Research: Oceans*, *97*(C5), 7373-7382. doi: https://doi.org/10.1029/92JC00188

Weiss, R. (1974). Carbon dioxide in water and seawater: the solubility of a non-ideal gas. *Marine Chemistry*, *2*(3), 203-215. doi: https://doi.org/10.1016/0304-4203(74)90015-2

# Supporting Information for "Estimating historical air-sea $CO_2$ fluxes: Incorporating physical knowledge within a data-only approach"

Val Bennington[1], Galen A McKinley[1], Tomislav Galjanic[2]

[1]Lamont-Doherty Earth Institute, Columbia University

[2]Data Science Institute, Columbia University

**Contents of this file**

1. Text S1 to S4

2. Figures S1 to S4

## S1. Large Ensemble Testbed Findings

Utilizing the Large Ensemble Testbed (Gloege, McKinley, et al., 2021), we analyzed how RMSE was impacted by reconstructing the $pCO_2$-Residual instead of $pCO_2$. The Large Ensemble Testbed consists of 25 ensemble members each from 4 Earth System Models. Within this model setting, we sample model features and $pCO_2$ at the same times and locations as we have actual SOCAT observations, in every ensemble member. Just as done with actual observations, an XGBoost algorithm is trained on the subset of features and $pCO_2$ from the models. We then reconstruct $pCO_2$ everywhere using the resulting functions and compare the reconstructed $pCO_2$ to the model "truth". Thus, the reconstructed $pCO_2$ can be evaluated at all times and locations simulated within the models, not just where we have SOCAT observations. Figure S1 shows that in addition to reducing the RMSE of the test data for each reconstruction ("test data"), RMSE across the globe, where the model has never been sampled ("unseen data"), is reduced using the $pCO_2$-Residual approach. Note also that against both test and unseen data, the high extreme RMSE is reduced by at least 3 $\mu$atm.

## S2. Uncertainty Due to Chlorophyll Climatology

Within the Large Ensemble Testbed, we use XGBoost to reconstruct $pCO_2$ using time-varying chlorophyll-a (every month has modeled chlorophyll-a) and compare to when the monthly climatology of model chlorophyll (1998 onward) is used for prior to 1998. As we do not have satellite observations of chlorophyll-a prior to 1997, this techniques estimates uncertainties caused by using a climatology of chlorophyll-a for the years prior to satellite observations. The calculated air-sea $CO_2$ flux differs significantly prior to the mid-1990s

and decreases to approximately 0.05 PgC/yr by 2005. There is variation across the models, with the largest mean impact on the reconstruction seen within the MPI model. The mean difference across the ESMs and time is less than 0.1 PgC/yr.

## S3. RMSE, Bias, MAE in $pCO_2$-Residual approach

The map of mean RMSE against all SOCAT observations using the $pCO_2$-Residual algorithm is shown in Figure S3. We see lowest RMSE in temperature-controlled regions, with values less than 10 $\mu$atm, as expected, and higher RMSE outside of these regions.

## S4. Test of Clustering with Self-Organizing Maps

To examine whether the regression would be improved by dividing the global ocean into biomes, we utilized the self-organizing map package SOMPY (Moosavi et al., 2014) (https://github. com/sevamoo/SOMPY). The global ocean was divided into 5, 10, and 15 clusters using maximum annual ice fraction, mean $pCO_2$, mean annual sea surface temperature, mixed layer depth, and spring mean chlorophyll (Fay & McKinley, 2014). On the global scale, there was no added skill, quantified based on RMSE and comparisons to independent data at BATS, HOT, LDEO, or GLODAP. We therefore maintain the simpler model.

## References

Fay, A. R., & McKinley, G. A. (2014). Global open-ocean biomes: mean and temporal variability. *Earth System Science Data*, *6*, 273–284. doi: doi:10.5194/essd-6-273-2014

Gloege, L., McKinley, G. A., Landschützer, P., Fay, A. R., Frölicher, T. L., Fyfe, J. C.,

... Takano, Y. (2021). Quantifying errors in observationally based estimates of ocean carbon sink variability. *Global Biogeochemical Cycles*, *35*(4), e2020GB006788. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GB006788 (e2020GB006788 2020GB006788) doi: https://doi.org/10.1029/2020GB006788

Gloege, L., Yan, M., Zheng, T., & McKinley, G. A. (2021). Improved quantification of ocean carbon uptake by using machine learning to merge global models and pco2 data. *Journal of Advances in Modeling Earth Systems*, *in press*.

Moosavi, V., Packmann, S., & Vallés, I. (2014). *Sompy: A python library for self organizing map (som).* (GitHub.[Online]. Available: https://github.com/sevamoo/SOMPY)

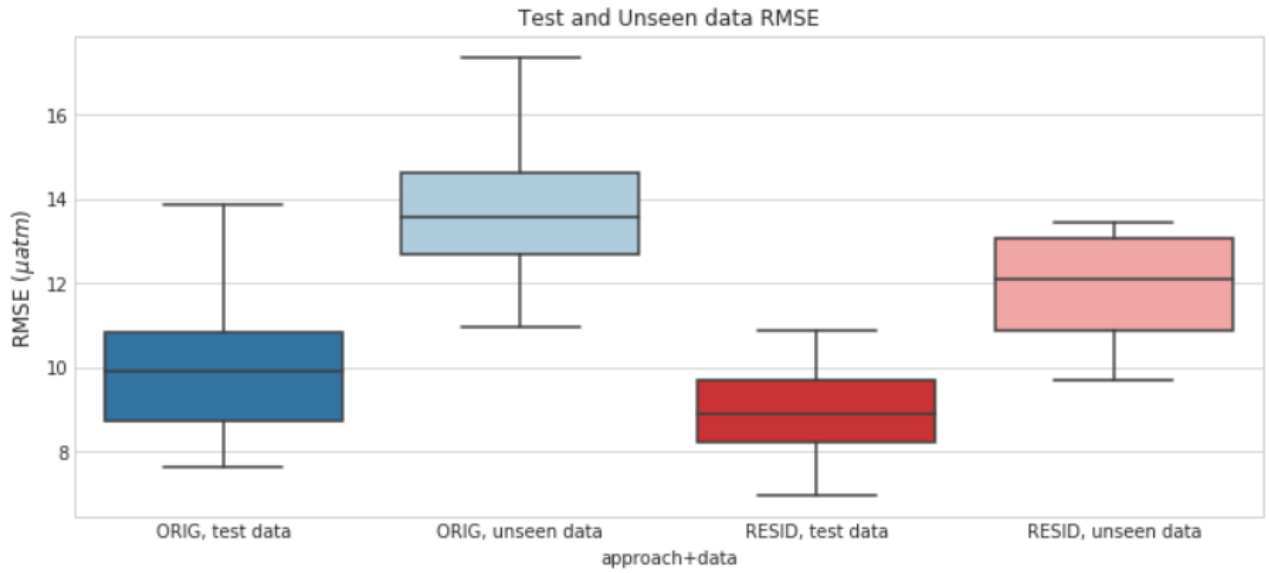


**Figure S1.** Test RMSE for $pCO_2$ reconstruction (ORIG, test data); RMSE at locations not sampled by SOCAT for $pCO_2$ reconstruction (ORIG, unseen data); Test RMSE for $pCO_2$-Residual approach (RESID, test data); RMSE at locations not sampled for $pCO_2$-Residual approach (RESID, unseen data). Each boxplot contains the 100 ensemble members, 25 from each Earth System Model of the Large Ensemble Testbed (Gloege, Yan, et al., 2021).

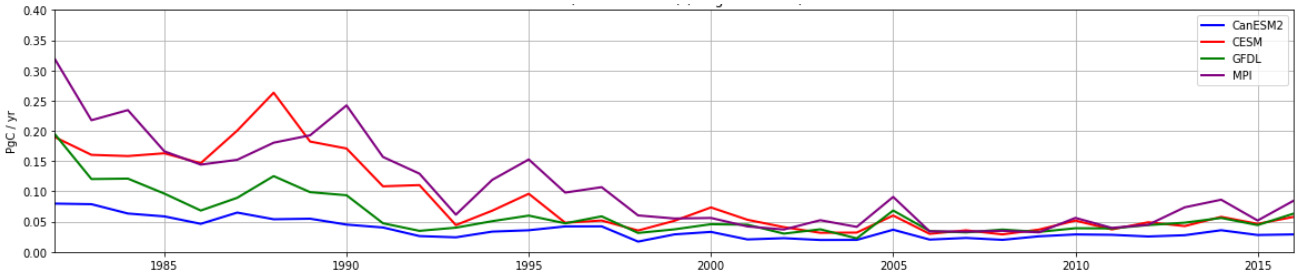


**Figure S2.** Absolute value of difference in globally-integrated reconstructed $CO_2$ flux (PgC/yr) when using a climatology of chlorophyll-a prior to 1998. Different colors represent the four different ESMs.

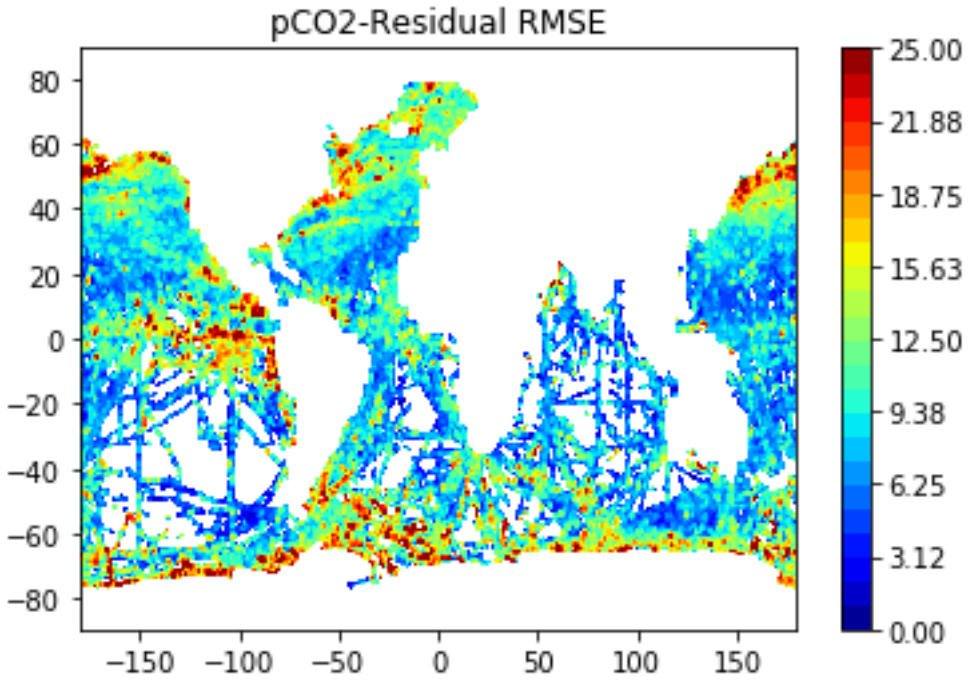


**Figure S3.** Mean RMSE ($\mu$atm) across the global ocean using the $pCO_2$-Residual approach.