JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2021MS002960

Key Points:

- A new approach for pCO_2 reconstruction applies pre-processing to remove the direct effect of temperature, simplifying the target variable for machine learning
- Reconstructed pCO_2 captures independent data more closely than most existing products
- Estimated ocean carbon uptake has a trend since 2005 (-0.05 PgC/yr^2) that is on the lower end of previous observation-based estimates

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

V. Bennington,
valerie.bennington@makai.com

Citation:

Bennington, V., Galjanic, T., & McKinley, G. A. (2022). Explicit physical knowledge in machine learning for ocean carbon flux reconstruction: The pCO_2-Residual method. *Journal of Advances in Modeling Earth Systems*, *14*, e2021MS002960. https://doi.org/10.1029/2021MS002960

Received 5 JAN 2022
Accepted 22 SEP 2022

Author Contributions:

Formal analysis: Tomislav Galjanic
Investigation: Tomislav Galjanic
Methodology: Tomislav Galjanic

Explicit Physical Knowledge in Machine Learning for Ocean Carbon Flux Reconstruction: The pCO_2-Residual Method

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

[1]Columbia University and Lamont-Doherty Earth Observatory, Palisades, NY, USA, [2]Makai Ocean Engineering, Waimanalo, HI, USA, [3]Data Science Institute, Columbia University, New York, NY, USA

Abstract The ocean reduces human impacts on global climate by absorbing and sequestering CO_2 from the atmosphere. To quantify global, time-resolved air-sea CO_2 fluxes, surface ocean pCO_2 is needed. A common approach for estimating full-coverage pCO_2 is to train a machine learning algorithm on sparse in situ pCO_2 data and associated physical and biogeochemical observations. Though these associated variables have understood relationships to pCO_2, it is often unclear how they drive pCO_2 outputs. Here, we make two advances that enhance connections between physical understanding and reconstructed pCO_2. First, we apply pre-processing to the pCO_2 data to remove the direct effect of temperature. This enhances the biogeochemical/physical component of pCO_2 in the target variable and reduces the complexity that the machine learning must disentangle. Second, we demonstrate that the resulting algorithm has physically understandable connections between input data and the output biogeochemical/physical component of pCO_2. The final pCO_2 reconstruction agrees modestly better with independent data than most other approaches. Uncertainties in the reconstructed pCO_2 and impacts on the estimated CO_2 fluxes are quantified. Uncertainty in piston velocity drives substantial flux uncertainties in some regions, but does not increase globally integrated estimates of uncertainty in CO_2 fluxes from observation-based products. Our reconstructed CO_2 fluxes show larger interannual variability than smoother neural network approaches, but a lesser trend since 2005. We estimate an air-sea flux of -1.8 PgC/yr (anthropogenic flux of -2.3 ± 0.5 PgC/yr) for 1990–2019, agreeing with other data products and the Global Carbon Budget 2020 (-2.3 ± 0.4 PgC/yr).

Plain Language Summary The ocean absorbs carbon dioxide from the atmosphere, moderating the human impact on Earth's climate. To quantify how much carbon dioxide is removed from the atmosphere 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 relationships between the ocean observations with near global coverage and ocean carbon are determined using machine learning algorithms and other statistical techniques. A concern is that these statistical algorithms do not require inputs to be linked to outputs in a manner consistent with ocean carbon cycle process understanding. Here, we develop a novel machine learning approach that starts by removing known physical signals from the data to create a cleaner signal for the computer algorithm to learn. Additional analysis demonstrates appropriate mechanistic links between algorithm inputs and outputs.

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 since the 1960s (Friedlingstein et al., 2021). Since the beginning of the Industrial Revolution, the ocean has absorbed about a third of the total anthropogenic emissions (Khatiwala et al., 2013; Sabine et al., 2004). The processes governing the large-scale distribution of ocean pCO_2 and the drivers of seasonality are well understood (Crisp et al., 2022; Takahashi et al., 2002, 2009). Yet, the quantification of year-to-year variability and long-term changes in this carbon sink remains a challenge (Crisp et al., 2022; Hauck et al., 2020; McKinley et al., 2016, 2017). 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 (McKinley et al., 2020). 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 tend to have significant mean and seasonal flux biases over basin-scale regions (Fay & McKinley, 2021).

Most observation-based products reconstruct the surface ocean pCO_2 across the global ocean in both space and time from sparse measurements using statistical techniques. These observation-based products typically use machine learning to develop a nonlinear function between observations of surface ocean pCO_2 and related variables that are 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. The Jena MLS approach reconstructs pCO_2 using a diagnostic model of mixed layer fluxes and statistical fits. In all products, air-sea fluxes of carbon dioxide are calculated from the resulting air-sea difference ($\Delta pCO_2 = pCO_2^{ocean} - pCO_2^{atm}$). While the resulting observation-based products show higher correlations and smaller root mean squared error (RMSE) against observations than do models (Bennington et al., 2022; Hauck et al., 2020), the limited physical interpretability of these statistical algorithms is a concern (Toms et al., 2020).

One way that has been proposed to improve the plausibility and confidence in machine learning based algorithms for geosciences is to explicitly incorporate physical knowledge of the system into the algorithm development workflow (Reichstein et al., 2019). This relatively new to the geosciences, and has typically been implemented using a modified cost function that penalizes unphysical results. Machine learning algorithms are trained by minimizing a cost function. This cost function is usually a sum of the Mean Squared Error (MSE) between the predicted output and the observed training data, 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. RMSE was smaller compared to predictions form a process based model. However, this standard neural network approach sometimes resulted in unphysical conditions. To improve upon the standard neural network approach, Read et al. (2019) modify their cost function to include a penalty for model predictions that cause excessive deviation from energy conservation across timesteps, and also implement algorithm pre-training using output from a one-dimensional physical model. Their final neural network algorithm further reduces RMSE and provides the best prediction of lake temperature profiles.

For our problem of reconstructing full-coverage ocean pCO_2 from heterogeneous in situ observations, how can we incorporate the physical mechanisms known to control the ocean carbon cycle? 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 (SST), chlorophyll-a (Chl-a), Mixed layer depth (MLD) climatology, Sea Surface Salinity (SSS), winds, geographic location, and time of year impact the resulting surface ocean pCO_2. Each of these features impacts pCO_2. Chl-a provides a measure of the biological production that removes dissolved inorganic carbon (DIC) from the surface ocean, thereby reducing surface ocean pCO_2. MLD is a proxy for ocean stratification. During highly stratified times, the phytoplankton are held within the lit surface ocean, setting up biological production. During periods of deeper mixing, DIC from depth is brought to the surface, and an increase in surface ocean pCO_2 occurs. But MLDs strongly co-vary with temperature. In other words, temperature has both direct and indirect effects on surface ocean pCO_2. The direct effect of temperature, due to solubility and chemical equilibrium, is that increasing (decreasing) temperatures directly causes in an increase (decrease) of pCO_2 (Takahashi et al., 2002). Temperature variations are also associated with 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.

Early efforts in pCO_2 reconstruction (Lefèvre & Taylor, 2002) relied only on SST and atmospheric pCO_2. All current machine learning approaches (Denvil-Sommer et al., 2019; Gregor et al., 2019; Landschützer et al., 2014) have SST as an input. To build these reconstruction algorithms through data-driven training, the statistics must identify a single function that disentangles the competing effects SST on pCO_2. In this work, we remove the direct temperature impact on pCO_2, pCO_2-T (Takahashi et al., 2002) from the target variable for the machine learning. This pre-processing step explicitly removes the well-known direct effect of temperature on pCO_2 from

our regression. This means that the machine learning algorithm has to learn only the biogeochemical/physical component pCO_2, supported by the information from SST and the other input variables.

The philosophy of the method described here is to apply pre-processing that removes from the target variable signals that are quantifiable based on prior knowledge. This leaves behind for the statistical fit the components that are not directly quantifiable. By removing known signals that influence the data, we aim to create a cleaner target variable for the machine learning. As described above, it is well-known that the pCO_2 signal is a mix of influences from direct temperature impacts and biogeochemical/physical effects (Takahashi et al., 2002, 2009). We calculate pCO_2-T from observations, remove this from observed pCO_2, and then use an XGBoost (XGB) algorithm to reconstruct the remaining component, "pCO_2-Residual." The final estimate of pCO_2 is the sum of pCO_2-T and pCO_2-Residual. In other studies from our research group, we have applied this philosophy in a substantially different way in developing and then extending the LDEO-Hybrid Data Physics (HPD) approach (Bennington et al., 2022; Gloege et al., 2022). In LDEO-HPD, the pre-existing knowledge is pCO_2 estimated by hindcast ocean biogeochemical models. An XGB algorithm identifies relationships between observed driver data and model errors (model error = $pCO_2^{model} - pCO_2^{SOCAT observed}$). Full-coverage model errors are estimated and then added to the original model fields to estimate full-coverage real-world pCO_2.

While both of these approaches follow the philosophy of applying pre-processing to incorporate prior knowledge, use an XGB algorithm and have similar input data, they are otherwise independent. A critical distinction is that this pCO_2-Residual method depends only on observations to both create the target variable and as algorithm input. In contrast, LDEO-HPD requires both observed pCO_2 and ocean biogeochemical model fields to create the target variable, and then uses observations as input. Both approaches are useful, given that they could have different practical applications. The pCO_2-Residual approach has the benefit that it could be applied before model outputs become available (Ciais et al., 2022). The merged model-data approach of LDEO-HPD can be used to make a data-constrained projection backward in time, before pCO_2 observations are available (Bennington et al., 2022).

In the following sections, we introduce the pCO_2-Residual approach and show that the resulting algorithm captures key physical processes of the surface ocean carbon cycle. Additionally, it performs modestly better than other data-only reconstruction approaches when compared to independent 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

Our approach is to pre-process pCO_2 data to remove the direct influence of temperature. This focuses the statistical algorithm on the spatio-temporal variance in pCO_2 due to biogeochemistry and other physical processes. Specifically, we calculate a residual (pCO_2-Residual), the difference between observed pCO_2 and the purely temperature (solubility and chemical equilibrium) driven component of pCO_2 (pCO_2-T, Takahashi et al. (2002)). 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. These steps are described in detail below.

2.1.1. Pre-Processing SOCAT Observations

We calculate surface ocean pCO_2 from the SOCAT v2021 monthly $1° \times 1°$ gridded fCO_2 product (Bakker et al., 2016). This is a quality-controlled data set 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 SST in Kelvin from the National Oceanic and Atmospheric Administration (NOAA) optimally interpolated SST version 2 (OISSTv2), B and δ are virial coefficients from Weiss (1974), and R is the gas constant (Dickson et al., 2007). All data products are regridded from their native resolutions to the SOCAT resolution of $1° \times 1°$ using bilinear interpolation. SOCAT fCO_2 data

Table 1
Summary of the Products, Variables, and Processing Steps Used for Feature and Target Data Sets

Product	Variable	Abbreviation	Processing
NOAA OISSTv2[a]	Sea surface temperature	SST	–
	SST anomaly	SST'	SST—monthly clim
Met Office: EN4[b]	Salinity	SSS	–
	SSS anomaly	SSS'	SSS—monthly clim
NOAA: GLOBALVIEW[c]	Atmospheric CO_2	xCO_2	–
ESA GlobColour[d]	Chl-a	Chl-a	Log_{10}(Chla)
	Chl-a anomaly	Chl-a'	Chl a—monthly clim
deBoyer Montegut[e]	Mixed layer depth	MLD	Log_{10}(MLD)
SOCATv2020[f]	Partial pressure of CO_2	pCO_2	Equations 1 and 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	$\cos\left(\frac{j*2\pi}{365}\right)$
$\overline{pCO_2}$	Mean pCO_2	pCO_2 clim	Equation 2 and Section 2.1.2

Note. Data products are regridded from their native resolutions to the SOCAT resolution of 1° × 1° using bilinear interpolation. [a]Source: https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html, Reynolds et al. (2002). [b]Source: https://www.metoffice.gov.uk/hadobs/en4/, Good et al. (2013). [c]Source: https://gml.noaa.gov/ccgg/mbl/, Masarie (2012). [d]Source: http://www.globcolour.info/, Maritorena et al. (2010). [e]Source: http://www.ifremer.fr/cerweb/deboyer/mld/home.php, de Boyer Montégut et al. (2004). [f]Source: https://www.socat.info/, Bakker et al. (2016).

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.

2.1.2. Initial pCO_2 Reconstruction

Calculation of the temperature influence on pCO_2 requires a gridded field of the long-term mean pCO_2. To estimate this, we use SOCAT pCO_2 data processed as in Section 2.1.1 and an XGBoost algorithm. Input features are the monthly observations and time/space variables on the first eight lines in Table 1. For the final pCO_2-Residual algorithm, the only use of this direct pCO_2 reconstruction is to determine the spatially resolved 1985–2019 long-term mean pCO_2 $\left(\overline{pCO_2}\right)$, required for calculation of the pCO_2-Residual target variable (Section 2.1.3 and Equation 2).

2.1.3. Calculating pCO_2-Residual

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

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

where $\overline{pCO_2}$ is mean surface ocean pCO_2 from the initial pCO_2 reconstruction (Section 2.1.2), SST is temperature in Celsius from NOAA OISSTv2, and $\overline{SST}$ is the local long term mean in SST in Celsius from NOAA OISSTv2. All data products are regridded from their native resolutions to the SOCAT resolution of 1° × 1° using bilinear interpolation. The residual (pCO_2-Residual) is calculated as the difference between observed pCO_2 and pCO_2T for all observations (Figure 1).

$$pCO_2^{Residual} = pCO_2 - pCO_2T \quad (3)$$

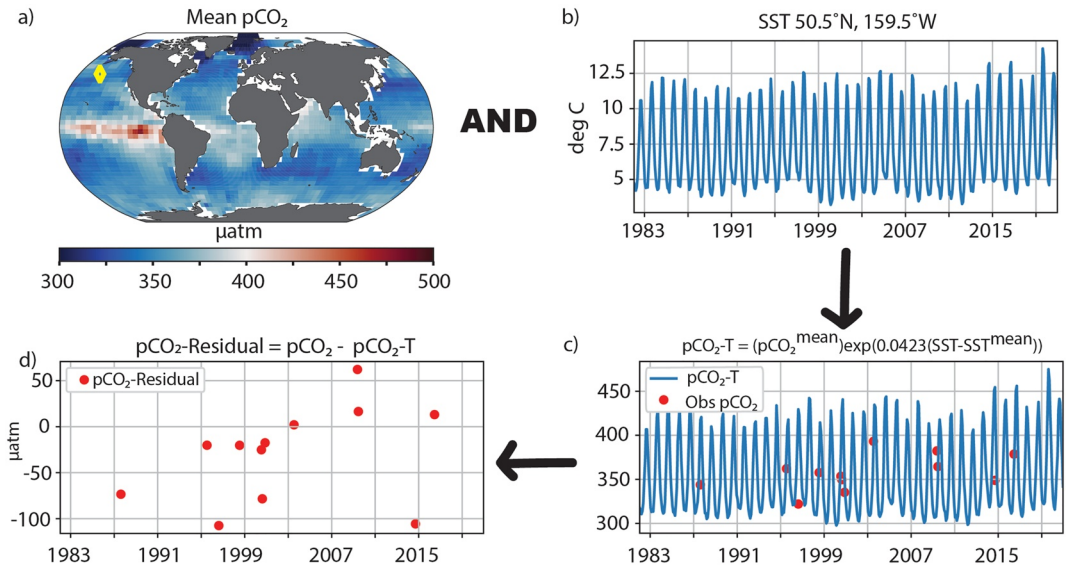

Figure 1. (a) Mean surface ocean pCO_2 from the initial run. (b) Observed satellite sea surface temperature time series from location of yellow diamond in subplot (a). (c) Calculated pCO_2-T (blue) and observed pCO_2 (red dots) at yellow diamond located in panel (a). (d) The calculated pCO_2-Residual,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

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 µatm, while subpolar regions may have residuals on the order of 100 µatm. Looking at the seasonality of the residual in Figures 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 narrower than a Gaussian distribution (Figure 2b), perhaps because of sampling bias that oversamples regions with a small residual. There is 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.

2.2. 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 MLD. 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-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. An example decision tree may increase the predicted pCO_2-Residual because the observed chlorophyll-a observation is below a certain threshold. 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 cost function (loss) is the 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

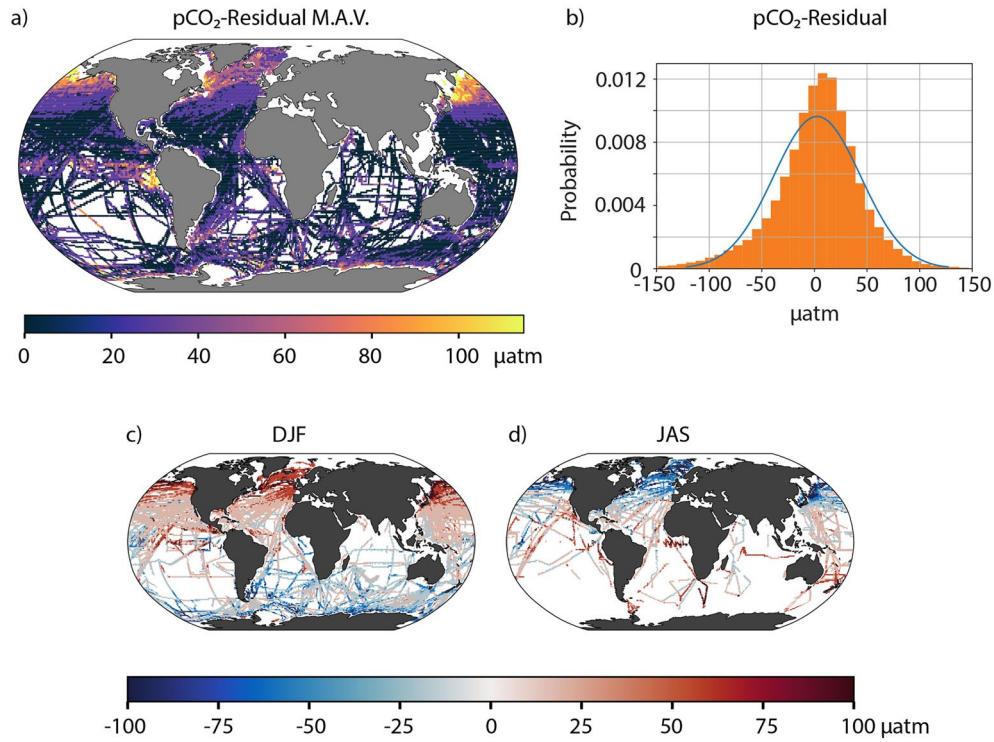

Figure 2. (a) Mean of the absolute value of the pCO_2-Residual calculated from all observations in the SOCAT database. (b) Histogram of the calculated pCO_2-Residual from SOCAT observations. (c) Mean pCO_2-Residual calculated for all observations during the northern hemisphere winter (DJF). Panel (d)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

of each tree. Our final XGBoost algorithm uses 1,000 decision trees with a maximum depth of seven levels. The training set is used to build the function between the features and the residual; that is, 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 data set. 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, selecting every fifth month for testing, and our final estimate of the residual is the ensemble mean of the five predictions, which then utilizes all of the available training data (20% are still held out for testing).

2.3. Features

In order to reconstruct the residual across both space and time, datasets with approximately full global coverage are used (Table 1): SST and Chlorophyll-a (Chl-a) from satellite; SSS from in situ data (Good et al., 2013); 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. For example, is this June's chlorophyll value higher/lower/equal to its climatological June value? Geographic location and time of year are incorporated using an N-vector transformation of latitude and longitude which transforms the latitude and longitude values to continuous values between 0 and 1, and a time transformation of day of year (D.O.Y.). 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 (S4).

We tested the sensitivity of the reconstruction to the source of mean pCO_2 $\left(\overline{pCO_2}\right)$ 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 the Lamont-Doherty Earth Observatory (LDEO) pCO_2 climatology (Takahashi et al., 2009) and the mean pCO_2 of the SeaFlux observation-based products (Fay et al., 2021). The alternative sources of mean pCO_2 did not significantly impact reconstructed pCO_2 or resulting air-sea CO_2 exchange, so we maintain our own method for the initial reconstruction of pCO_2 (Section 2.1.2).

2.3.1. 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 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 to 2019 observations at all locations and months prior to 1998. Within the Large Ensemble Testbed (Gloege et al., 2021), we show that utilizing climatological chlorophyll prior to 1998 introduced a mean uncertainty of 0.1 PgC/yr to the global air-sea CO_2 exchange (Text S2 in Supporting Information S1).

2.4. Feature Importance

One of the benefits of the XGBoost algorithm is that it facilitates the determination of 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 statistical 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.

2.5. Independent Data Sets

In our algorithm training, we use 80% of the observations contained within the SOCAT database, and hold out 20% for testing. We also wish to 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 LDEO data set (Takahashi et al., 2009) (data already in SOCAT are removed) and the GLobal Ocean Data Analysis Project version 2 (GLODAPv2 (Olsen et al., 2016)). For LDEO, pCO_2 is directly measured. For the other datasets, pCO_2 is calculated from observations of Total Alkalinity, 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 µatm in LDEO (directly measured) to >12 µatm in GLODAPv2 (calculated). Given the

Table 2
Observation-Based Products Used for Comparison (Fay et al., 2021; Gregor & Fay, 2021)

Product	Reference
CSIR ML6	Gregor et al. (2019)
CMEMS	Denvil-Sommer et al. (2019)
HPD	Gloege et al. (2022)
MLS	Rödenbeck et al. (2015)
MPI-SOMFFN	Landschützer et al. (2014) and Landschützer et al. (2020)

known larger biases in some of the other observation-based products in the 1980s, we compare to observations within the time frame 1990–2019.

2.6. Regression Skill

To compare predicted pCO_2 (P) to the observations (O), we examine the correlation (r), bias, 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 central value of the sorted absolute errors. 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.

2.7. 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 the all data products shown here.

2.8. CO_2 Flux Calculations

The bulk air-sea CO_2 flux (FCO_2) is calculated as:

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

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 Naegler (2009); K0 is the solubility calculated using EN4 salinity and OISST temperatures (Weiss, 1974); ice fraction 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. All data products are regridded from their native resolutions to $1° \times 1°$ using bilinear interpolation. For a consistent comparison K_w, K0, ice fraction, and pCO_2^{atm} from SeaFlux are used (Fay et al., 2021). The SeaFlux data set (Gregor & Fay, 2021) includes K_w for three wind speed products: CCMPv2, ERA5, and JRA55. Fluxes presented are the mean flux across the three wind products.

2.8.1. 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 (Fay et al., 2021; Gregor & Fay, 2021).

2.8.2. Anthropogenic Carbon Flux

Observation-based products that 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 ($F_{net} = F_{natural} + F_{ant}$) (Crisp et al., 2022). The natural outgassing of riverine carbon is understood to be the dominant component of $F_{natural}$ (Aumont et al., 2001; Crisp et al., 2022; Hauck et al., 2020). Additional non-riverine components such as outgassing due to ocean circulation change have been proposed for 1994–2007 (Gruber et al., 2019), with

Table 3
pCO_2 Test Statistics for Each of the Five Ensemble Members and Their Mean Values

	Run 1	Run 2	Run 3	Run 4	Run 5	Mean
RMSE (µatm)	16.13	16.02	16.76	16.51	16.25	16.33
Bias (µatm)	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 (µatm)	10.88	10.87	11.20	11.13	10.92	11.00
Median AE (µatm)	7.41	7.49	7.57	7.68	7.46	7.52

large uncertainty in magnitude and no evidence of long-term impact (Crisp et al., 2022); these are assumed to be zero here. To quantify the anthropogenic air-sea CO_2 flux from our estimate of F_{net}, $F_{natural}$ due to rivers must be subtracted. Quantifying $F_{natural}$ due to riverine carbon is a complex scientific problem with substantial remaining uncertainties. Here, as in Gloege et al. (2022), we use an average of three estimates: Jacobson 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 PgC/yr). The combined estimated efflux due to riverine carbon is 0.49 ± 0.26 PgC/yr, and we remove the efflux of 0.49 PgC/yr from the estimated annual global air-sea CO_2 fluxes calculated using pCO_2 from the pCO_2-Residual method and from other observation-based products.

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 3. We have a mean test RMSE of 16.33 µatm, lower than the recent data product of Gregor et al. (2019) (17.16 µatm). Each run has a relatively small bias and is highly correlated with the test observations. The Mean Absolute Error (Mean AE) is near 11 µatm, and the Median Absolute Error (Median AE) is less than 8 µatm. For the ensemble, RMSE is lowest (below 10 µatm) in the subtropical regions as we would expect due to the small possible values of the residual there, and higher in the equatorial Pacific, Southern Ocean, and subpolar North Atlantic and subpolar North 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).

We tested this pCO_2-Residual approach with the Large Ensemble Testbed in which a suite of Earth System Models are used to evaluate reconstruction performance given real-world sampling (Gloege et al., 2021) (Text S1 in Supporting Information S1). These tests demonstrate that pCO_2-Residual can reconstruct pCO_2 with higher skill than a similar algorithm that uses pCO_2 as the target variable (Section 2.1.2). With respect to bias, improvements are greatest in the poorly sampled regions in the southern hemisphere where temperature is a primary control on pCO_2 (Figure S2 in Supporting Information S1). The phasing and amplitude of interannual to decadal variability is also improved across much of the open ocean in the pCO_2-Residual approach compared to when pCO_2 is the target variable (Figures S3 and S4 in Supporting Information 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 time series sites Hawaii (HOT) and Bermuda (BATS), reconstructed surface ocean pCO_2 is highly correlated with observations (Figure 3, top row). The seasonal cycle is the dominant signal in these two datasets, and the correlation here primarily indicates how well these cycles are captured. It has previously been shown that subtropical mean and seasonality are very similar across observation-based products (Fay & McKinley, 2021; Gloege et al., 2022; Rödenbeck et al., 2015). These comparisons primarily tell us that the pCO_2-Residual technique gives us the right phasing and amplitude of the seasonality at BATS and HOT. Further analysis (Section 3.4, Text S1, and Figure S2 in Supporting Information S1) indicates low bias in this region. More detailed analysis would be required to provide a detailed quantification of product skill in interannual variability and trend at BATS and HOT (Gloege et al., 2021, 2022).

GLODAP and LDEO are observations taken along ship transects traveled irregularly. These datasets provide groundtruth for spatial patterns of pCO_2, but the variability timescales that they represent are less clear. They should not primarily represent seasonal variability given very limited regular reoccupation of the sampling sites (Olsen et al., 2016; Takahashi et al., 2009), but instead a variety of timescales from interannual to decadal. Correlations to the data are slightly lower than at BATS and HOT for all reconstruction approaches (Figure 3, bottom row). pCO_2-Residual and LDEO-HPD (Gloege et al., 2022) perform marginally better than other observation-based data products. All approaches underestimate the amplitude of observed variability, except JENA-MLS. The unbiased RMSE in pCO_2-Residual is approximately equal to LDEO-HPD. A decadal breakdown of these Taylor

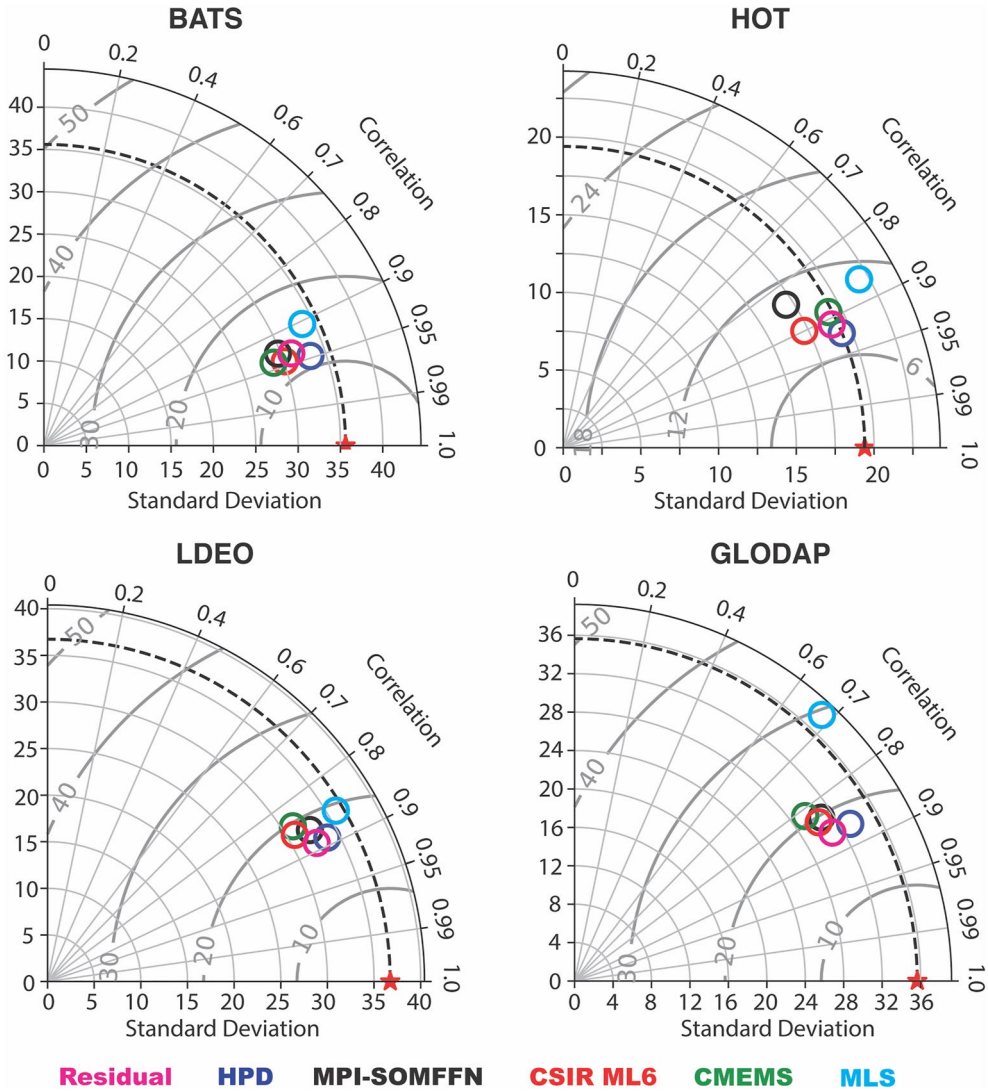

Figure 3. Taylor diagrams (Taylor, 2001) of Correlation (along circumference), Standard Deviation (along radii), and root mean squared error (gray arcs centered at red star) of five previous observation-based approaches (Hybrid Data Physics in blue, MPI-SOMFFN in black, CSIR ML6 in red, CMEMS in green, and MLS in cyan) and the new pCO_2-Residual technique (magenta). Bermuda Atlantic Time-Series Study is shown at the top left, Hawaii Ocean Time-Series at the top right, LDEO at bottom left, and GLobal Ocean Data Analysis Project at bottom right.

diagrams (Figure S5 in Supporting Information S1) demonstrates that both pCO_2-Residual and LDEO-HPD compare similarly to the data as other approaches in the 1990s. However, in the 2000s and 2010s, both are clearly closer to the observations. However, this doesn't necessarily indicate that these approaches will better capture the decadal variability of the globally integrated air-sea CO_2 flux, and indeed the two approaches disagree substantially on the magnitude of decadal variability in the globally integrated air-sea CO_2 flux (Figure 6).

3.3. Physical Mechanisms

With additional analysis, we can reveal the relative contributions of features to the algorithm's prediction. The first column in Figure 4 shows the mean (1982–2019) importance of MLD; geographic location and D.O.Y.; 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 D.O.Y., because there is no seasonal cycle in location, but there is geographic variation of the impact of D.O.Y. on the pCO_2-Residual. The second column of Figure 4 examines the mean seasonal cycles of feature importance for

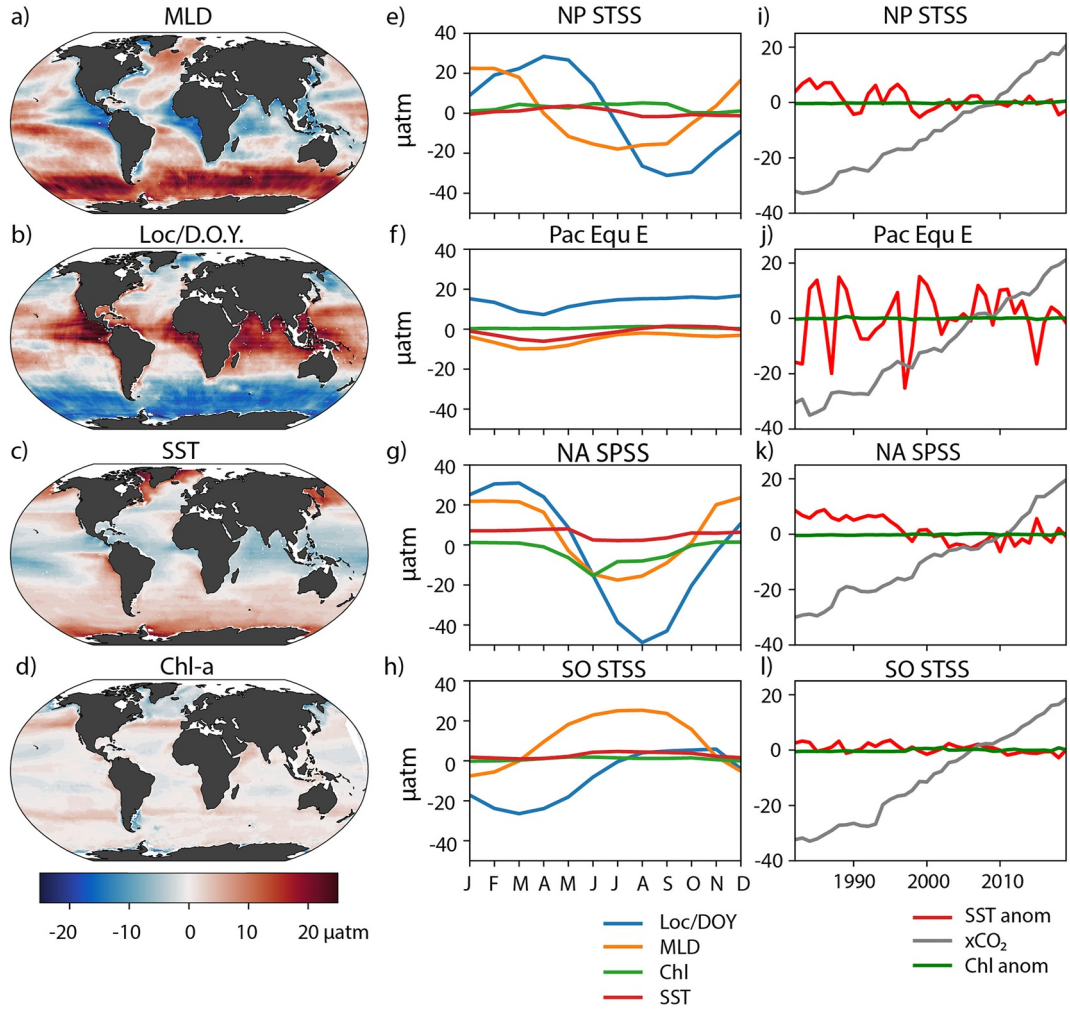

Figure 4. Mean feature importance and seasonal cycles and interannual variations in feature importance in sample biomes (µatm). (a) Mean feature importance of mixed layer depth (MLD). (b) Mean feature importance of Location and Day of Year (D.O.Y.). (c) Mean feature importance of sea surface temperature (SST). (d) Mean feature importance of Chl-a. (e) Mean seasonal cycles of feature importance of Location/D.O.Y., MLD, Chl-a, and SST in the North Pacific Subtropical Seasonally Stratified (NP STSS) biome. (f) Same as in (e) except for the Pacific Equatorial East (Pac Equ E) biome. (g) Same as in (e) except for the North Atlantic Subpolar Seasonally Stratified (NA SPSS) biome. (h) Same as in (e) except for the Southern Ocean STSS (SO STSS) biome. (i) Interannual variations in feature importance for SST, Chl-a, and $x\mathrm{CO}_2$ within the NP STSS biome. (j) Same as in (i) except for within the Pac Equ E 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.

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 MLD, which has large mean feature importance (Figure 4a), but also large seasonal variations away from the equator (Figures 4e–4h). Deep winter mixing brings up DIC and increases pCO_2, whereas shallower mixed layer depths set up biological production which decreases surface DIC. During northern hemisphere winter (DJF), the algorithm's estimate of the pCO_2-Residual is significantly increased (decreased) by MLD 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 D.O.Y. significantly increases the pCO_2-Residual on the mean in equatorial zones and decreases the pCO_2-Residual in the Southern Ocean (Figure 4b). The small mean impact of these combined features in the subpolar northern regions is due to cancellation of significant seasonal variations in its importance to the reconstructed pCO_2-Residual (Figures 4d–4f).

While the direct impacts of SST have been set aside by the pre-processing (Equation 2), Figures 4g–4i show the importance of SST to the reconstructed residual. Summertime stratification sets up biological production, and wintertime deep mixing brings up older, remineralized 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 (Figures 4e, 4g and 4h). On the mean, 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 (Figures 4e, 4f and 4h), 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 (Figures 4i–4l). However, within the algorithm, interannual anomalies in SST do cause significant adjustments to the predicted pCO_2-Residual, particularly in the eastern equatorial Pacific.

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 (Figures 4i–4l). 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 to the long-term growth of xCO_2. Within the algorithm, interannual variability in the reconstructed pCO_2 Residual is largely controlled by interannual anomalies in SST in all regions.

Considering the spatial impact of interannual variability, 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 time series 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 long-term positive trend in atmospheric CO_2, and year-to-year variations within the model 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:

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

Figure 5 displays the mean value (1985–2019) of the second term 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, that is, there are the lowest uncertainties, within the subtropical oceans (±less than 10 µatm at 67% confidence). Uncertainties become larger

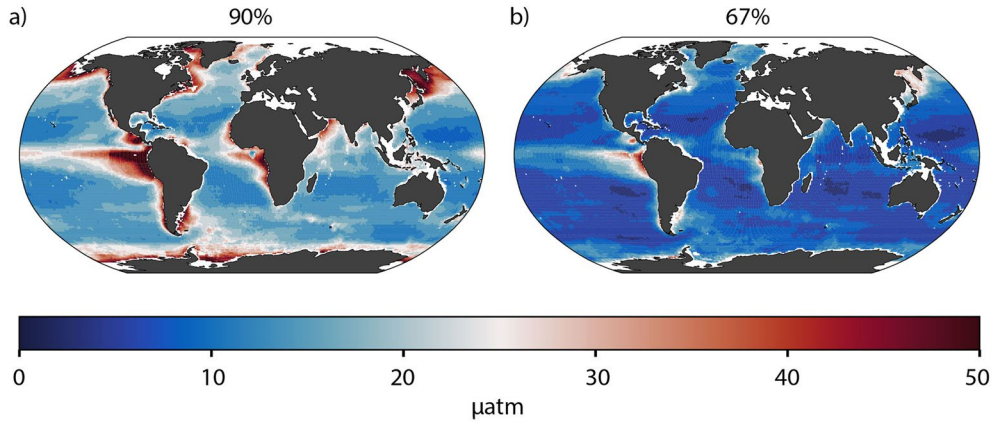

Figure 5. Mean pCO_2 uncertainty within the (a) 90% and (b) 67% confidence bounds. At a given location, the shading represents the mean value that would be added and subtracted to form the confidence interval of reconstructed pCO_2.

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 observations (off the southwestern tip of South America and in 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 features common to other reconstructions (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.

3.6. Uncertainty in CO_2 Fluxes Due To pCO_2 Convolution With Winds

In order to determine the uncertainty in CO_2 flux caused by our uncertainty in surface ocean pCO_2 when combined with uncertainty in the winds, we assume zero global mean bias in the pCO_2 reconstruction. This assumption is supported by the analysis of Gloege et al. (2021) and our own analysis with the Large Ensemble Testbed (Text S1 in Supporting Information 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° by 1° 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 important 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 at high latitudes where there are moderate pCO_2 uncertainties (Figure 5) and significant piston velocities (Figure 7b). In the equatorial Pacific, the flux impact of large pCO_2 uncertainties are damped by the 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

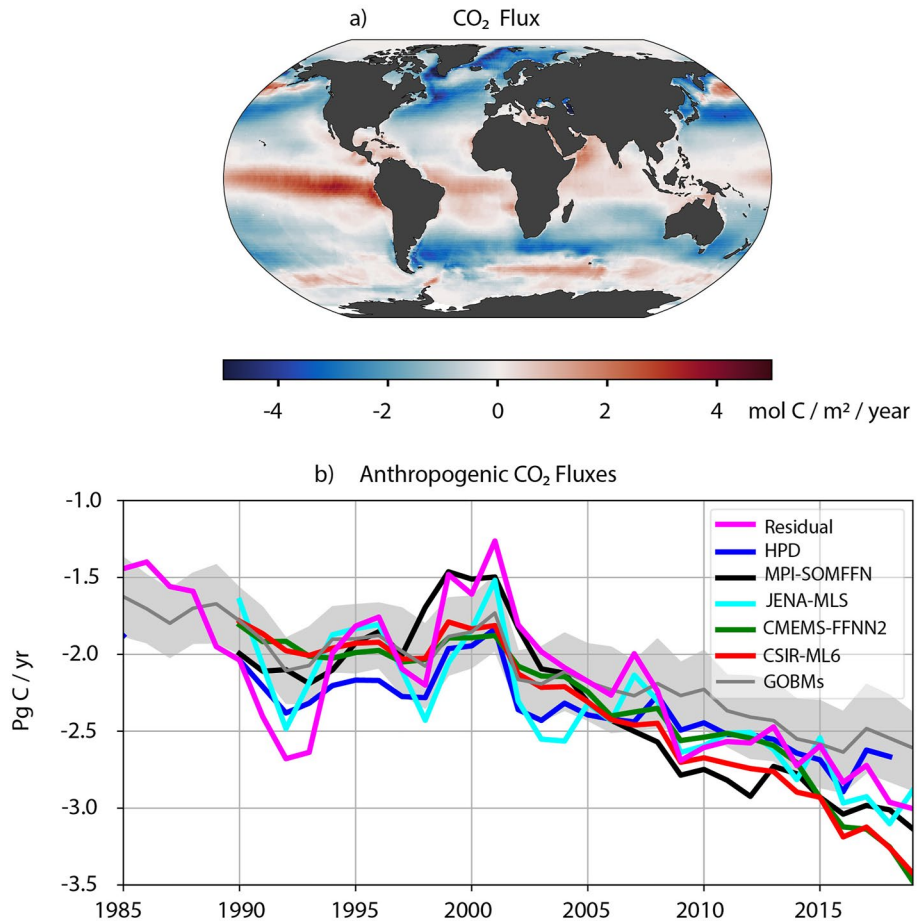

Figure 6. (a) Map of mean (1985–2019) anthropogenic air-sea CO_2 Flux reconstructed by the pCO_2-Residual Technique. (b) Annual mean (1985–2019) air-sea CO_2 fluxes estimated by the pCO_2-Residual (magenta), Hybrid Data Physics (blue), MPI-SOMFFN (black), CSIR ML6 (red), CMEMS (green), and MLS (cyan) data products. Mean of the nine Global Ocean Biogeochemical Models and one standard deviation shading in gray. Harmonized observation-based products begin in 1990 (Gregor & Fay, 2021).

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 convolved with wind speeds (≈0.01 PgC/yr). However, uncertainties in piston velocities estimated by different wind products cause a standard deviation of annual fluxes of 0.04–0.10 PgC/yr (not shown). Combining these with the square root sum of squares, we estimate a total uncertainty from wind speeds, both due to the convolution with pCO_2 uncertainty and the piston velocity, at 0.10 PgC/yr, one standard deviation, for the 67% confidence interval.

Fay et al. (2021) perform a complete uncertainty analysis for globally integrated CO_2 flux estimates from observation-based pCO_2 products. They account for differences across products (0.19 PgC/yr) and uncertainties in gas transfer parameterizations (0.39 PgC/yr, Wanninkhof (2014)). If we assume that this Monte-Carlo analysis of the convolution of winds with pCO_2 for one product is representative, we can use the calculation above to replace the winds uncertainty of Fay et al. (2021) (0.09 PgC/yr) with 0.10 PgC/yr. Combining this again with square root sum of squares, we arrive at a total uncertainty of 0.45 PgC/yr. This is the same as found by Fay et al. (2021), indicating that the convolution of pCO_2 with wind uncertainty is not a significant factor in the total uncertainty of globally integrated CO_2 fluxes estimated from observation-based pCO_2 products. However, for regional and local estimates in regions of high piston velocity, this factor does have the potential to be important (Figure 7a).

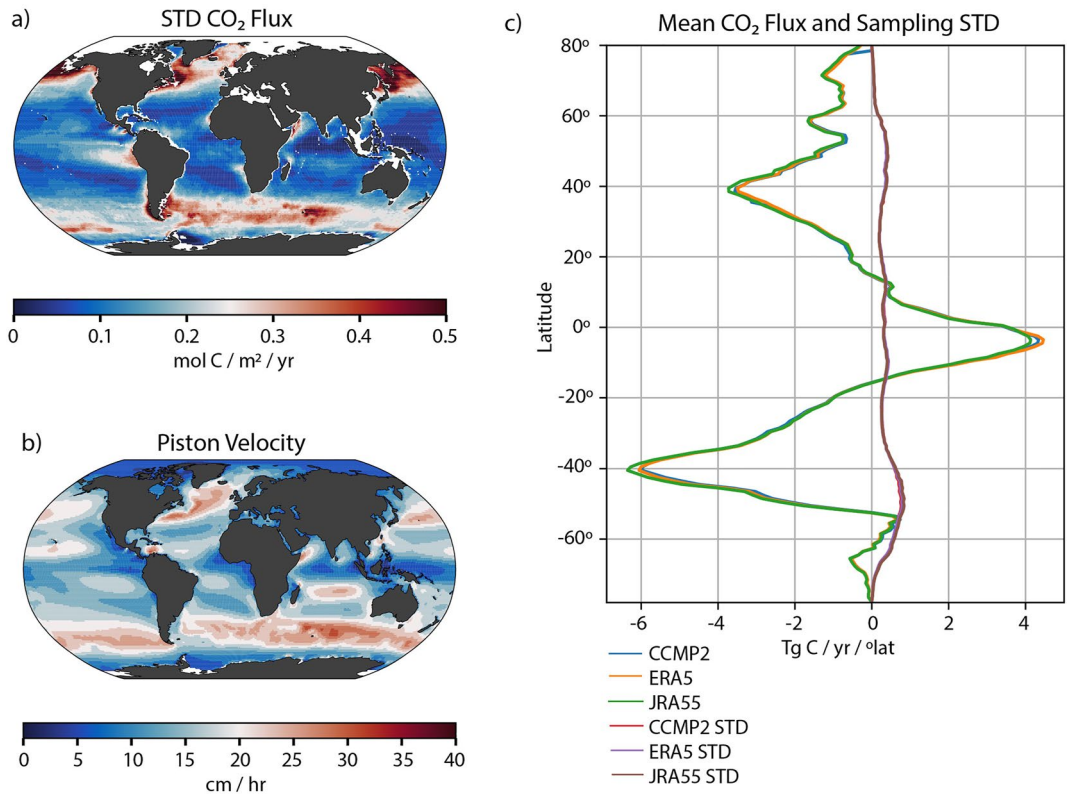

Figure 7. Uncertainty in CO_2 fluxes due to convolution of pCO_2 with winds. (a) Standard deviation of annual CO_2 flux from Monte Carlo simulations (mol C/m^2/yr). (b) Mean piston velocity, average of CCMP2, ERA5, and JRA55 (Kw: cm/hr). (c) Mean CO_2 flux by latitude band (Tg C/yr) and wind product,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

4. Discussion

The new pCO_2 reconstruction algorithm presented here captures expected relationships between input features and pCO_2 output. This result is obtained by pre-processing the data with explicit physical knowledge so that the target variable for the machine learning has reduced complexity. By reconstructing the difference 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 and 2), the pCO_2-Residual approach requires the machine learning algorithm to learn only the biogeochemical/physical component of 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.

In the pCO_2-Residual algorithm, MLD, location, season, SST, Chl-a, and xCO_2 impact the reconstructed pCO_2 in physically expected ways (Figure 4). MLD, location, and time of year strongly control the seasonal cycles of reconstructed pCO_2, while atmospheric CO_2 concentrations and SST control interannual variations. Year-to-year variations in Chl-a do not drive significant variability. 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). It could also be due to correlations between Chl-a anomalies and other features, such as SST, that the algorithm uses to determine pCO_2. These mechanistically interpretable connections between observed inputs and algorithm output are a step forward for pCO_2 reconstruction.

pCO_2-Residual 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 and Figure S5 in Supporting Information S1). Uncertainties in reconstructed pCO_2 are smallest in the subtropical ocean regions and largest in the equatorial Pacific and subpolar regions (Figure 5),

as expected given the small range of the pCO_2-Residual target variable in these regions (Figure 2). The pattern of uncertainty in pCO_2 is very similar to the pattern of test RMSE (not shown), and the magnitude of the global mean test RMSE (16.33 µatm) lies between the global mean uncertainty magnitude at the 67% confidence interval (9.8 µatm) and the 90% confidence interval (19.71 µatm). Thus, our multiple approaches to understanding pCO_2 uncertainty are internally consistent. Air-sea CO_2 fluxes are in agreement with previous observation-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 opposed to a neural network in which non-linearities are controlled by the activation function (Baughman & Liu, 1995).

Uncertainty in the air-sea CO_2 fluxes due to the convolution of pCO_2 and winds, as determined by a Monte Carlo approach, are largest where both piston velocities and pCO_2 uncertainty are large (Figure 7). Although there are regions of significant local flux uncertainty, globally integrated air-sea CO_2 flux uncertainty due to pCO_2 convolved with wind remains small (0.01 PgC/yr). Uncertainty due to the piston velocity itself is larger, with annual flux uncertainty ranging from 0.04 to 0.1 PgC/yr, for a total uncertainty bound due to winds of 0.10 PgC/yr. This is consistent with our Large Ensemble Testbed analysis (Text S1 in Supporting Information S1) that suggests a pCO_2-Residual global mean flux error is ≈0.1 PgC/yr. It is important to note that we have assumed no bias in the observations (Fay et al., 2021; Gloege et al., 2022). If observational bias exists, particularly in regions with moderate to high piston velocities, the flux uncertainty could be larger. This uncertainty estimate is only barely larger than the wind product uncertainty estimated by Fay et al. (2021), and does not change their comprehensive estimate of uncertainty in CO_2 flux from observation-based products (0.45 PgC/yr) that also accounts for uncertainty in gas exchange parameterization and pCO_2 reconstruction approaches.

Our group has also developed another algorithm that uses pre-processing to add physical information to the target variable. In LDEO-HPD (Gloege et al., 2022), this knowledge is the pCO_2 estimated by GOBMs. In that approach, model-observation discrepancy is the target variable for the machine learning algorithm, and this is reconstructed and added to the original model output to estimate full pCO_2 field. That approach requires having a suite of GOBMs for input, while here only observations are needed as input. As discussed in Introduction, while both LDEO-HPD and pCO_2-Residual use pre-processing to reframe the target variable for the machine learning, they are separate approaches. Both approaches use similar observed features and are limited by the same sparsity of pCO_2 observations, but their target variables are unique and the algorithms are separately trained.

The Global Ocean Carbon Budget 2020 (Friedlingstein et al., 2020) estimates an anthropogenic ocean carbon sink of -2.5 ± 0.4 PgC/yr for the period 2000–2019 based on a suite of ocean hindcast models. With pCO_2-Residual, we estimate a similar anthropogenic flux of -2.35 ± 0.5 PgC/yr, after adjusting the net flux of -1.86 PgC/yr for outgassing of riverine carbon. The trend in the ocean carbon sink since 2005 estimated by this method (-0.05 PgC/yr^2) is on the lower end of previous observation-based estimates (Figure 6).

Analysis with the Large Ensemble Testbed (Text S1 in Supporting Information S1), demonstrates that pCO_2-Residual provides an improved reconstruction of the mean and interannual to decadal variability over our initial reconstruction based on pCO_2 alone. Previous work with the Large Ensemble Testbed (Gloege et al., 2021), demonstrated that MPI-SOMFFN reconstruction technique overestimated CO_2 flux decadal amplitude by 21% for the global average, and 31% in the Southern Ocean (<35°S). Considering pCO_2 in the Testbed, MPI-SOMFFN overestimates decadal variability amplitude by 33% globally, 37% in S. Ocean. In contrast, pCO_2-Residual in Testbed has reduced overestimation of pCO_2 decadal variability (24% globally, 28% S. Ocean). These amplitude comparisons are the median of 100 Earth System Model realizations for 1982–2016, with substantial spread across the individual members (Gloege et al., 2021).

In the real-world reconstructions from pCO_2-Residual and MPI-SOMFFN (Figure 6a), the global and Southern Ocean CO_2 flux decadal amplitude is 15% larger in pCO_2-Residual than in MPI-SOMFFN (pCO_2-Residual: 1.1 PgC/yr globally, 0.59 PgC/yr S. Ocean; MPI-SOMFFN: 0.95 PgC/yr globally, 0.51 PgC/yr S. Ocean). The same approach as in Gloege et al. (2021) is used here to extract the decadal signal, but amplitude is defined as maximum-minimum since the true state is unknown. This finding is not inconsistent with the Testbed results discussed in the previous paragraph. Testbed results indicate the median skill of a reconstruction method across a range of climate conditions estimated by four ESMs, 25 ensembles members from each. In the real world, only one evolution of climate occurred. Real-world reconstructions over the last few decades should not be expected to linearly scale to the true decadal variability based on Testbed median skill estimates (Gloege et al., 2021). In

other words, its quite reasonable that, in the Testbed, pCO_2-Residual would have less overestimation of decadal variability compared to another method, but at the same time estimate slightly greater variability from recent observations.

While Gregor et al. (2019) suggest ocean surface pCO_2 reconstructions may have "hit a wall," here we illustrate additional progress. The LDEO-HPD (Bennington et al., 2022; Gloege et al., 2022) and this pCO_2-Residual technique, both applying physical knowledge for pre-processing in a machine learning workflow, offer improvements in skill based on comparisons to independent datasets (Figure 3 and Figure S5 in Supporting Information S1). In addition, pCO_2-Residual offers improved reconstruction skill based on analysis with the Large Ensemble Testbed (Figures S2–S4 in Supporting Information S1), clearly explainable links between inputs and output (Figure 4), and new uncertainty quantification (Figure 7).

5. Conclusions

We develop a new machine learning approach to reconstruct global ocean pCO_2, an approach that uses pre-processing to explicitly incorporate 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 (Takahashi et al., 2002) from the target variable for the machine learning. The resulting algorithm exhibits mechanistically appropriate relationships between inputs and outputs. Spatially resolved estimates of uncertainties illustrate the differential impacts of pCO_2 and wind uncertainties on CO_2 fluxes. We find an ocean carbon sink within the range of previous observation-based approaches, but with a lesser trend since 2005.

Data Availability Statement

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 available at https://github.com/valbennington/JAMES_pub_2022. Reconstruction available for download at: https://zenodo.org/record/6438445.

Acknowledgments
The authors acknowledge support from NOAA (NA20OAR4310340), NSF through the LEAP STC (Award 2019625), and the Data Science Institute of Columbia University. We thank all the data providers and quality controllers who work tirelessly to create the SOCAT pCO_2 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. https://doi.org/10.1029/1999GB001238

Bakker, D. C. E., Pfeil, B., Landa, C. S., Metzl, N., O'Brien, K. M., Olsen, A., et al. (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. https://doi.org/10.5194/essd-8-383-2016

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

Bennington, V., Gloege, L., & McKinley, G. A. (2022). Variability in the Global Ocean Carbon Sink from 1959-2020 by correcting models with observations. *Geophysical Research Letters*, *49*, e2022GL098632. https://doi.org/10.1029/2022GL098632

Bennington, V., McKinley, G. A., Dutkiewicz, S., & Ullman, D. (2009). What does chlorophyll variability tell us about export and air-sea CO_2 flux variability in the North Atlantic? *Global Biogeochemical Cycles*, *23*(3), GB3002. https://doi.org/10.1029/2008gb003241

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

Ciais, P., Davis, S., Saatchi, S., Deng, Z., Poulter, B., Chevallier, F., et al. (2022). Towards near-real-time estimates of greenhouse gas budgets. *Nature*. in review.

Crisp, D., Dolman, H., Bastos, A., Sitch, S., Tanhua, T., McKinley, G., et al. (2022). How well do we understand the land-ocean-atmosphere carbon cycle? *Reviews of Geophysics*, *60*(2), e2021RG000736. https://doi.org/10.1029/2021RG000736

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*, *109*, C12003. 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 pCO_2 over the global ocean. *Geoscientific Model Development*, *12*(5), 2091–2105. https://doi.org/10.5194/gmd-12-2091-2019

Descamps, B. (2020). 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 CO_2 measurements*. North Pacific Marine Science Organization.

Fay, A. R., Gregor, L., Landschützer, P., McKinley, G. A., Gruber, N., Gehlen, M., et al. (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. https://doi.org/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*(2), 273–284. https://doi.org/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*, *48*(20), e2021GL095325. https://doi.org/10.1029/2021gl095325

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C., Hauck, J., et al. (2021). Global carbon budget 2021. *Earth System Science Data Discussions*, *14*(4), 1917–2005. https://doi.org/10.5194/essd-14-1917-2022

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Olsen, A., et al. (2020). Global carbon budget 2020. *Earth System Science Data*, *12*(4), 3269–3340. https://doi.org/10.5194/essd-12-3269-2020

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

Gloege, L., Yan, M., Zheng, T., & McKinley, G. A. (2022). Improved quantification of ocean carbon uptake by using machine learning to merge global models and pCO_2 data. *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002620. https://doi.org/10.1029/2021MS002620

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. https://doi.org/10.1002/2013JC009067

Gregor, L., & Fay, A. (2021). SeaFlux: Harmonised sea-air CO_2 fluxes from surface pCO_2 data products using a standardised approach [Dataset]. Zenodo. https://doi.org/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 CO_2 estimates using a machine-learning ensemble (CSIR-ML6 version 2019a) – Have we hit the wall? *Geoscientific Model Development*, *12*, 5113–5136. https://doi.org/10.5194/gmd-12-5113-2019

Gruber, N., Landschützer, P., & Lovenduski, N. S. (2019). The variable southern ocean carbon sink. *Annual Review of Marine Science*, *11*(1), 159–186. https://doi.org/10.1146/annurev-marine-121916-063407

Hauck, J., Zeising, M., Le Quéré, C., Gruber, N., Bakker, D. C. E., Bopp, L., et al. (2020). Consistency and challenges in the ocean carbon sink estimate for the global carbon budget. *Frontiers in Marine Science*, *7*, 571720. https://doi.org/10.3389/fmars.2020.571720

Humphreys, M. P., Lewis, E. R., Sharp, J. D., & Pierrot, D. (2021). $PyCO_2SYS$ v1.7: Marine carbonate system calculations in Python. *Geoscientific Model Development Discussions*, *159*, 1–45. https://doi.org/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), GB1019. https://doi.org/10.1029/2005GB002556

Khatiwala, S., Tanhua, T., Fletcher, S. M., Gerber, M., Doney, S. C., Graven, H. D., et al. (2013). Global ocean storage of anthropogenic carbon. *Biogeosciences*, *10*(4), 2169–2191. https://doi.org/10.5194/bg-10-2169-2013

Lacroix, F., Ilyina, T., & Hartmann, J. (2020). Oceanic CO_2 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. https://doi.org/10.5194/bg-17-55-2020

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. https://doi.org/10.1002/2014GB004853

Landschützer, P., Laruelle, G. G., Roobaert, A., & Regnier, P. (2020). A uniform pco_2 climatology combining open and coastal oceans. *Earth System Science Data*, *12*(4), 2537–2553. https://doi.org/10.5194/essd-12-2537-2020

Lefèvre, N., & Taylor, A. (2002). Estimating pCO_2 from sea surface temperatures in the Atlantic gyres. *Deep Sea Research Part I: Oceanographic Research Papers*, *49*(3), 539–554. https://doi.org/10.1016/S0967-0637(01)00064-4

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. https://doi.org/10.1016/j.rse.2010.04.002

Masarie, K. A. (2012). *Islscp II globalview: Atmospheric CO_2 concentrations*. ORNL Distributed Active Archive Center. https://doi.org/10.3334/ORNLDAAC/1111

McKinley, G. A., Fay, A. R., Eddebbar, Y. A., Gloege, L., & Lovenduski, N. S. (2020). External forcing explains recent decadal variability of the ocean carbon sink. *AGU Advances*, *1*(2), e2019AV000149. https://doi.org/10.1029/2019AV000149

McKinley, G. A., Fay, A. R., Lovenduski, N. S., & Pilcher, D. J. (2017). Natural variability and anthropogenic trends in the ocean carbon sink. *Annual Review of Marine Science*, *9*(1), 125–150. https://doi.org/10.1146/annurev-marine-010816-060529

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*(7591), 469–472. https://doi.org/10.1038/nature16958

Naegler, T. (2009). Reconciliation of excess ^{14}C-constrained global CO_2 piston velocity estimates. *Tellus B: Chemical and Physical Meteorology*, *61*(2), 372–384. https://doi.org/10.1111/j.1600-0889.2008.00408.x

Olsen, A., Key, R. M., van Heuven, S., Lauvset, S. K., Velo, A., Lin, X., et al. (2016). The global ocean data analysis Project version 2 (GLODAPv2) – An internally consistent data product for the world ocean. *Earth System Science Data*, *8*(2), 297–323. https://doi.org/10.5194/essd-8-297-2016

Peters, G. P., Le Quéré, C., Andrew, R. M., Canadell, J. G., Friedlingstein, P., Ilyina, T., et al. (2017). Towards real-time verification of CO_2 emissions. *Nature Climate Change*, *7*(12), 848–850. https://doi.org/10.1038/s41558-017-0013-9

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

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Resplandy, L., Keeling, R. F., Rödenbeck, C., Stephens, B. B., Khatiwala, S., Rodgers, K. B., et al. (2018). Revision of global carbon fluxes based on a reassessment of oceanic and riverine carbon transport. *Nature Geoscience*, *11*(7), 504–509. https://doi.org/10.1038/s41561-018-0151-3

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. https://doi.org/10.1175/1520-0442(2002)015<1609:aiisas>2.0.co;2

Rödenbeck, C., Bakker, D. C. E., Gruber, N., Iida, Y., Jacobson, A. R., Jones, S., et al. (2015). Data-based estimates of the ocean carbon sink variability – First results of the surface ocean pCO_2 mapping intercomparison (SOCOM). *Biogeosciences*, *12*(23), 7251–7278. https://doi.org/10.5194/bg-12-7251-2015

Rödenbeck, C., Keeling, R. F., Bakker, D. C. E., Metzl, N., Olsen, A., Sabine, C., & Heimann, M. (2013). Global surface-ocean pCO_2 and sea-air CO_2 flux variability from an observation-driven ocean mixed-layer scheme. *Ocean Science*, *9*(2), 193–216. https://doi.org/10.5194/os-9-193-2013

Sabine, C. L., Feely, R. A., Gruber, N., Key, R. M., Lee, K., Bullister, J. L., et al. (2004). The oceanic sink for anthropogenic CO_2. *Science*, *305*(5682), 367–371. https://doi.org/10.1126/science.1097403

Shapley, L. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.). *Ann. Math. Study28, contributions to the theory of games*, (pp. 307–317).

Takahashi, T., Sutherland, S. C., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., et al. (2002). Global sea–air CO_2 flux based on climatological surface ocean pCO_2, and seasonal biological and temperature effects. *Deep Sea Research Part II: Topical Studies in Oceanography*, *49*(9), 1601–1622. 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., et al. (2009). Climatological mean and decadal change in surface ocean pCO_2, and net sea–air cCO_2 flux over the global oceans. *Deep Sea Research Part II: Topical Studies in Oceanography*, *56*(8), 554–577. 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*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

Toms, B. A., Barnes, E. A., & Uphoff, I. E. (2020). Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), 1. https://doi.org/10.1029/2019ms002002

Wanninkhof, R. (2014). Relationship between wind speed and gas exchange over the ocean revisited. *Limnology and Oceanography: Methods*, *12*(6), 351–362. https://doi.org/10.4319/lom.2014.12.351

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

References From the Supporting Information

Moosavi, V., Packmann, S., & Vall´es, I. (2014). Sompy: A python library for self organizing map (SOM). Retrieved from https://github.com/sevamoo/SOMPY