

# Machine learning-based modeling of ecosystem-atmosphere CO<sub>2</sub> exchange in support of carbon accounting for nature-based climate solutions

Jeffrey Uyekawa<sup>a</sup>, John Leland<sup>a</sup>, Darby Bergl<sup>b,d</sup>, Yujie Liu<sup>b</sup>, Andrew D. Richardson<sup>b,c</sup> and Benjamin Lucas<sup>a,\*</sup>

<sup>a</sup>Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, AZ, USA

<sup>b</sup>Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, AZ, USA

<sup>c</sup>School of Informatics, Computing & Cyber Systems, Northern Arizona University, Flagstaff, AZ, USA

<sup>d</sup>Department of Biology, Northern Arizona University, Flagstaff, AZ, USA

## ARTICLE INFO

### Keywords:

Carbon dioxide flux  
nature-based climate solutions  
machine learning  
XGBoost  
NEON  
Ameriflux  
phenocam

## ABSTRACT

In recent years, support for nature-based climate solutions to reduce climate change has grown in response to our inability to curb fossil fuel emissions through behavioral change. Nature-based solutions reduce atmospheric CO<sub>2</sub> by managing, restoring, and/or conserving ecosystems, which then act as carbon reservoirs by storing the CO<sub>2</sub> in soils and woody biomass. An important aspect of the future success of these strategies is the ability to accurately quantify carbon dioxide turbulent flux (FCO<sub>2</sub>) under different ecological conditions.

In this study, we predicted FCO<sub>2</sub> at 44 core terrestrial sites across the National Ecological Observatory Network in the United States using 35 environmental drivers and site-specific variables as predictors. We compared the accuracy of seven different machine learning algorithms and found that Extreme Gradient Boosting (XGBoost) consistently produced the most accurate predictions (Root Mean Squared Error of 1.81 μmolm<sup>-2</sup>s<sup>-1</sup>, R<sup>2</sup> of 0.86). The model showed excellent performance testing on sites that are ecologically similar to other sites (the Mid Atlantic, New England, and the Rocky Mountains), and poorer performance at sites with fewer ecological similarities to other sites in the data (Pacific Northwest, Florida, and Puerto Rico). The results show strong potential for machine learning-based models to make more skillful predictions than state-of-the-art process-based models, being able to estimate the multi-year mean carbon balance to within an error ± 50gCm<sup>-2</sup>y<sup>-1</sup> for 29 of our 44 test sites.


## 1. Introduction

Rising levels of atmospheric CO<sub>2</sub> are the primary cause of climate change (Lee et al., 2023). Human-caused emissions of CO<sub>2</sub> from fossil fuel burning and land use change are too large to be fully offset by the uptake of CO<sub>2</sub> that occurs by terrestrial ecosystems and the oceans (Friedlingstein et al., 2023). The important role of terrestrial ecosystems in the global carbon cycle has been relatively well-understood for decades (Wofsy and Harris, 2002; Schimel et al., 2001). One strategy to reduce future climate change is to manage, restore, or otherwise conserve ecosystems so that they remove even more CO<sub>2</sub> from the atmosphere and store it in slow-turnover carbon reservoirs, e.g. in the soil or woody biomass. Support for these so-called natural climate solutions (also known as nature-based climate solutions) has increased in recent years (Fargione et al., 2018; Griscom et al., 2017; Bossio et al., 2020), in part because efforts to reduce fossil fuel emissions have not yet been successful. However, a prerequisite to strategically implementing natural climate solution strategies is a thorough reliable estimation of the carbon uptake potential of different ecosystem types, and

how this varies in space and time, e.g., with site factors and year-to-year variation in weather.

Hemes et al. (2021) have argued that ecosystem-scale CO<sub>2</sub> flux measurements can play an important role in developing strategies for, and evaluating natural climate solutions. For example, Hollinger et al. (2021) noted the value of CO<sub>2</sub> flux measurements for quantifying not only the magnitude of CO<sub>2</sub> uptake by an evergreen forest in Maine but also the persistence of this strong sink over 25 years of measurements. In this case most of the annual net uptake of CO<sub>2</sub> ended up in woody biomass which can sequester atmospheric carbon for decades to centuries. Both Hemes and Hollinger also highlighted the value of CO<sub>2</sub> flux measurements in the context of natural climate solutions. Here, it was shown that these measurements are useful for estimating the rates of carbon sequestration in hard-to-observe storage pools such as soils.

Tower-based, ecosystem-scale CO<sub>2</sub> flux measurements quantify the exchange of turbulence flux of CO<sub>2</sub> (FCO<sub>2</sub>, measured in μmolm<sup>-2</sup>s<sup>-1</sup>) between the land surface and the atmosphere. In short, FCO<sub>2</sub> measures how much CO<sub>2</sub> is moving into or out of an ecosystem, per unit area and per unit time. During daytime hours, most ecosystems are a strong sink for CO<sub>2</sub> (negative FCO<sub>2</sub>, following the micrometeorological sign convention), as they remove CO<sub>2</sub> from the atmosphere through the process of photosynthesis.

 ben.lucas@nau.edu (B. Lucas)

ORCID(s): 0009-0007-7324-0485 (J. Uyekawa); 0009-0004-8258-6806 (J. Leland); 0009-0007-3584-4974 (D. Bergl); 0000-0003-0335-6400 (Y. Liu); 0000-0002-0148-6714 (A.D. Richardson); 0000-0002-2021-3076 (B. Lucas)

By comparison, during the night, ecosystems are generally a moderate source of CO<sub>2</sub> (positive FCO<sub>2</sub>), as they release CO<sub>2</sub> back into the atmosphere through the process of respiration. FCO<sub>2</sub> is measured using a method known as eddy covariance (EC) (Baldocchi, 2020). Eddy covariance measurements are continuous in time (24 hours a day, 7 days a week, 365 days a year) and are generally reported at an hourly or half-hourly temporal resolution. Global networks of eddy covariance flux towers collect in situ carbon flux measurements, providing information on photosynthesis dynamics across different ecosystems and under various environmental conditions. Currently, FCO<sub>2</sub> is measured at hundreds of research sites across the USA, with 385 of these sites being members of the AmeriFlux Network (Novick et al., 2018; Chu et al., 2023). Indeed, the driving motivation for the establishment of AmeriFlux almost three decades ago was to measure the carbon balance of different ecosystems, and more specifically to better understand the distribution of CO<sub>2</sub> sinks and sources across the continent (United States Department of Energy, 2023).

While these measurements run continuously at high frequency (e.g. at 5 Hz), practical limitations such as technical failures, instrument malfunction, and the necessity for filtering out data with low turbulent conditions can lead to gaps in the collected data. This in turn results in the compromised validity of the measured fluxes. Moreover, there is no attempt to standardize the measurements across sites within the AmeriFlux network, meaning that when measurements are available, they may be more or less reliable than another site.

From the perspective of understanding the distribution of CO<sub>2</sub> sinks across the entire continent, the sampling provided by AmeriFlux is woefully inadequate; even assuming that all 385 AmeriFlux sites are currently active, this equates to approximately 1 flux measurement site every 25,000 km<sup>2</sup>. Therefore, extrapolation and upscaling from individual sites to fine resolutions and regional and continental scales must be done using either process-based or statistically-based models. The former approach is attractive because these simulation models are based on state-of-the-art understanding of how the carbon cycle works. However, parameterization and initial conditions remain outstanding challenges, and past model validation efforts have highlighted serious model errors. By comparison, the latter approach is unattractive because many of these statistical approaches are essentially black boxes from which it is impossible to verify process-level representation. Standardization of inputs for statistical models is also a challenge, and, to the best of our knowledge, validation of model predictions has generally not been conducted against independent datasets.

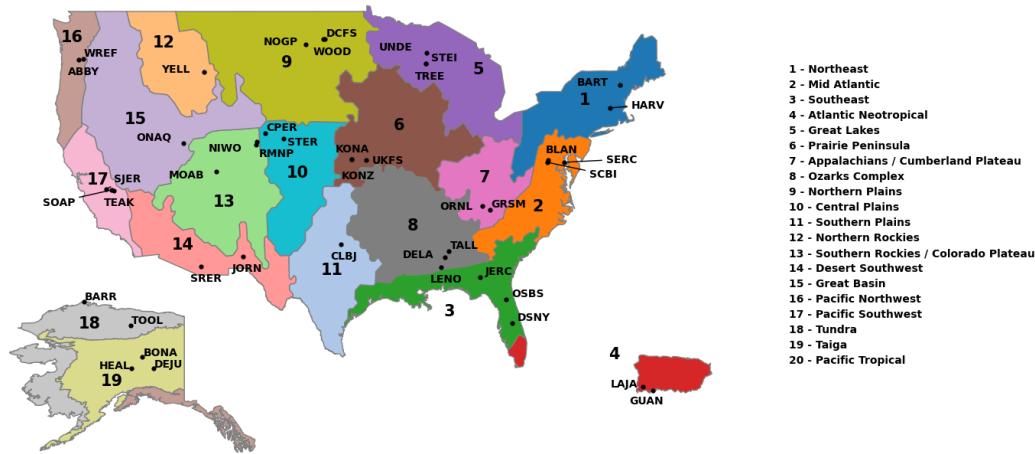
An extensive model-data comparison project of over 20 ecosystem models conducted under the North American Carbon Program found that process-based models generally performed poorly in representing site-level carbon flux dynamics across sites with varying land cover. Specifically, substantial model errors in representing FCO<sub>2</sub> were found

at annual, seasonal, and diurnal time scales (Dietze et al., 2011; Schwalm et al., 2010); models misrepresented the inter-annual variability in observed CO<sub>2</sub> uptake (Keenan et al., 2012); models did not properly represent phenological transitions in Spring or Fall (Richardson et al., 2012a); and models could not predict photosynthetic uptake within the uncertainty of observations (Schaefer et al., 2012). These results lead to valid questions about the viability of using process-based models to evaluate natural climate solution strategies.

Statistically-based upscaling of FCO<sub>2</sub> began about two decades ago with the pioneering work of Papale and Valentini (2003). They used an artificial neural network, trained with CO<sub>2</sub> flux data from 16 measurement sites in Europe to calibrate a simulation model to predict CO<sub>2</sub> fluxes of European forests at 1 km resolution. Several years later, (Xiao et al., 2008) calibrated a modified regression tree model to FCO<sub>2</sub> measurements across the AmeriFlux network, using satellite observed greenness indicators, such as vegetation indices, leaf area index, and fraction of observed photosynthetically active radiation (Kang et al., 2023). The sophistication of these kinds of upscaling efforts has matured over the last 15 years. The current state of the art is probably defined by the FLUXCOM project (Jung et al., 2020), which uses satellite remote sensing and gridded meteorological products to calibrate a model trained on FCO<sub>2</sub> measurements from sites around the world.

However, a challenge with past efforts to upscale site-level measurements is the lack of standardization in measurement protocols across sites. For example, across the AmeriFlux network, the choice of instrument setup and configuration, and even the details of flux data processing and corrections (which are critically important), may be different for each site. Furthermore, key instrumentation principles (e.g., open vs. closed path gas analyzer or sonic anemometer geometry), installation protocols (e.g., depth profiles of soil temperature and moisture measurements), measured and calibrated quantities (gravimetric vs. volumetric soil water content vs. soil water potential), and even units (hPa vs. kPa for vapor pressure deficit – easily converted, but also easily incorrectly reported or interpreted) are not consistent across sites. In particular, this lack of consistency of site variables across sites is a major barrier for any predictive modeling methods which use machine learning techniques.

While AmeriFlux has been characterized as a “coalition of the willing” (Novick et al., 2018), the USA’s National Ecological Observatory Network (NEON) was specifically established to “collect long-term open access ecological data to better understand how U.S. ecosystems are changing” (Battelle, 2024). Implicit in this mission statement is the need for standardization of measurement protocols and techniques across sites. NEON sites are strategically located, following a clustering algorithm to identify and group distinct regions of vegetation, landforms, and ecosystem dynamics into 20 different



**Figure 1:** A map of the NEON core terrestrial sites and their locations within the 19 ecological domains.

domains, as shown in Figure 1. Within each domain, at one or more monitoring sites, standardized measurements of environmental drivers (weather, solar radiation, etc.) are conducted along with ecosystem-level measurements of FCO<sub>2</sub> and other quantities measured by eddy covariance (e.g. sensible and latent heat fluxes). This standardization opens up the possibility to use a machine learning algorithm to predict site-level FCO<sub>2</sub> without relying on gridded or reanalysis products as is necessary when using sites from AmeriFlux as a whole. Thus, the network of NEON sites represents an opportunity to train models on observational data across numerous sites which might be viewed as analogous to a model emulator (Fer et al., 2018). The key difference being that this model is trained on real observations rather than the output of a simulation model.

In this paper we investigate the potential to use cutting-edge machine learning algorithms in conjunction with standardized CO<sub>2</sub> flux measurements and environmental data (“drivers”) from NEON towers to make predictions about the half-hourly, daily, and annual FCO<sub>2</sub> in ecosystems across the continental US. We implement seven machine learning algorithms, of varying complexities, across two experimental scenarios: (1) A randomized 10-fold cross-validation, and (2) A cross-validation stratified by site, which we refer to as ‘Leave-one-site-out’ (LISO). In this scenario, a single site is left out of the training data and the resulting trained model is used to predict the FCO<sub>2</sub> values of this ‘unknown’ site (unknown to the model, not the experimenters). The first scenario assesses how well we can gap-fill missing FCO<sub>2</sub> values with a machine learning model when other values from that site are known, while the second assesses how well we could predict a completely unseen site based only on the environmental drivers. In each scenario, we found that the lowest error was obtained using an optimized Extreme Gradient-Boosted Tree (XGBoost) model. Our analysis

shows that XGBoost can predict FCO<sub>2</sub> values to within a root mean squared error of  $1.81\mu\text{molm}^{-2}\text{s}^{-1}$ , with our predictions having an R-squared of 0.86 with the actual measurements. The LISO results predict FCO<sub>2</sub> to within a root mean squared error of  $2.45\mu\text{molm}^{-2}\text{s}^{-1}$ , however these results varied greatly ( $0.66\text{--}6.22\mu\text{molm}^{-2}\text{s}^{-1}$ ) depending on the domain of the site and its similarity to other sites in the dataset. We use our optimal model to gap-fill a complete FCO<sub>2</sub> dataset and provide it online for use by future researchers.

## 2. Methods

Our experiments compared the performance of seven machine learning algorithms to predict half-hourly FCO<sub>2</sub> measurements collected between January 1st, 2016 and June 30, 2022. The experimental details are all provided in the following section and the code for the experiments is available at: <https://github.com/js1339/AmeriFlux>.

### 2.1. Data

There are 47 NEON core terrestrial sites located across the U.S. and Puerto Rico, which strategically represent a range of vegetation, climate, and ecosystems divided into 20 different ecological domains as shown in Figure 1. Our experiments used data collected at 44 sites, as three sites—Marvin Klemme Range Research Station (OAES), Mountain Lake Biological Station (MLBS), and Puu Makaala Natural Area Reserve (PUUM)—were removed from the analysis due to inconsistencies in predictor variables, missing flux measurements, and errors arising during preprocessing. With these sites removed, our 44 sites represented 19 out of the 20 ecological domains (see National Ecological Observatory Network (NEON) (2024) for general information about the data product).

We preprocessed the data using the R package REdDyproc, as is the standard approach for gap-filling and  $u^*$  filtering of carbon flux values. We used the U50 threshold to filter our  $u^*$  values.

Table 1 shows a general explanation and summary statistics for the environmental drivers that we used as feature variables to learn our models. The data were sourced from 3 locations: AmeriFlux, the Phenocam Network, and MODIS satellite imagery (Richardson et al., 2018; Seyednasrollah et al., 2019; United States Department of Energy, 2023).

Each site was assigned both a primary and secondary vegetation type from the following categories:

1. Agricultural (AG)
2. Deciduous Broadleaf (DB)
3. Evergreen Broadleaf (EB)
4. Evergreen Needleleaf (EN)
5. Grassland (GR)
6. Shrub (SH)
7. Tundra (TN)

After preprocessing, our final dataset consisted of 961,340 observations unevenly divided among the 44 NEON sites.

## 2.2. Experimental design

We compared the predictive performance of 6 machine learning algorithms (explained in section 2.3 below) in two experimental scenarios. In the first experimental scenario, we performed 10-fold cross validation on the data. This means that the data were randomly divided into 10 ‘folds’, with each containing approximately 10% of the data. The models were then trained using 9 folds (90% of the available data) and tested on the remaining fold. This process was repeated so that each fold was used in the training set 9 times and appeared as the test set once (see Figure 2a for an illustrated explanation). The performance of each algorithm was reported as the average across the 10 different runs. We note that the data were divided into the same 10 folds for each predictive algorithm.

$K$ -fold cross-validation is a common technique in the testing and comparison of machine learning algorithms as it removes selection bias (whether deliberate or not), and demonstrates the ability of the models to generalize to unseen data (Rodriguez et al., 2009).

In the second experimental scenario, which we will refer to as Leave-one-site-out cross-validation (LISO CV), we began by partitioning the data by site, resulting in 44 uneven groups of data. We then employed a similar process to scenario one, where the models were trained on all-but-one group and tested on the remaining group (an example is shown in Figure 2b). This was repeated so that each site was used as the test data once, and therefore the stated performance metrics are the average of the 44 models fitted and tested.

The LISO CV experiments present an inherently more difficult problem than the prior scenario as a predictive

model significantly benefits from learning from data belonging to the test site. These experiments were included to replicate a situation where a site has no prior carbon flux recordings, i.e. it could be a new site or the instrumentation might not be functioning correctly. In addition, this experimental setup also tests whether we might be able to make a minimal set of measurements at a site with lower standardization in measurement protocols in order to predict the FCO<sub>2</sub>. This would be helpful for carbon accounting purposes and nature-based carbon solutions, and also to enable a benchmark for land surface model simulations and checking existing datasets.

The performance of each model was assessed using 2 evaluation metrics—Root Mean Squared Error (RMSE) and the Coefficient of Determination ( $R^2$ ). The RMSE is the square root of the average of the squared prediction errors over all of the data in the test set. Specifically,

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

Where  $y$  is the measured (true) value and  $\hat{y}$  is the predicted value for a test set of size  $N$ . Due to the squared component of the metric, the RMSE is sensitive to large errors in any of the individual predictions.

The  $R^2$  evaluation metric is a measure of the goodness-of-fit of the linear model found by regressing the predicted values against the true values. It is calculated as:

$$R^2(\hat{y}, y) = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

In contrast to RMSE,  $R^2$  is not sensitive to large errors in any of the individual predictions as it measures the amount of total variance accounted for by the predictions. When used together, these metrics complement each other and provide a more comprehensive picture of the performance of the algorithms. Each metric can then be analyzed on half-hourly, daily, and annual timescales. Ensuring accurate model predictions on an annual scale is important for reliable carbon accounting. However, it is also critical to evaluate model performance at finer temporal resolutions, such as half-hourly and daily scales, to ensure that our models produce accurate annual predictions for scientifically sound reasons. In order to produce meaningful predictions of annual sums of FCO<sub>2</sub> for each test site, we must first use models optimized in the 10-fold experimental setting to fill in missing FCO<sub>2</sub> values for each site before making predictions per site in the LISO experimental setting.

## 2.3. Machine Learning Models

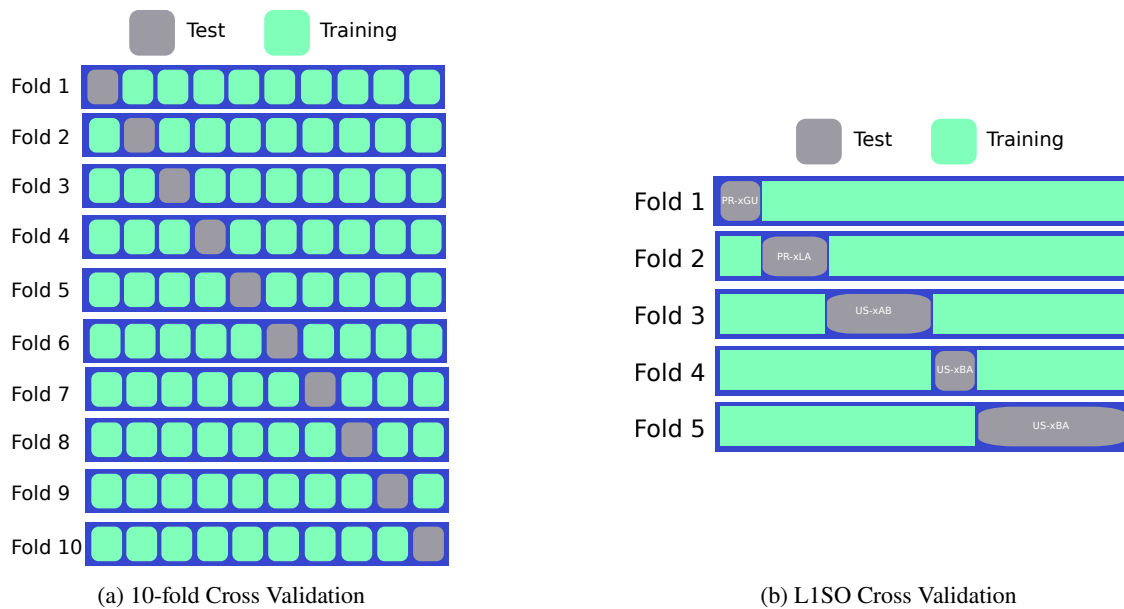
We compared the performance of 6 different machine learning models on predicting carbon dioxide flux. They are:

1. *Linear Regression (all predictors)*: This is a linear model including all of the variables using the

**Table 1**

Environmental drivers (feature variables) used as input to our machine learning models to predict carbon dioxide flux

Variable	Description	Source	Mean	Min	Max
DOY	Day Of Year	Ameriflux/NEON	0.49	0	1
HOUR	Hour Of Day	Ameriflux/NEON	0.49	0	1
TS_1_1_1	Soil Temperature Depth 1	Ameriflux/NEON	12.64	-29.82	56.15
TS_1_2_1	Soil Temperature Depth 2	Ameriflux/NEON	12.17	-29.85	52.52
PPFD	Photosynthetic Photon Flux Density	Ameriflux/NEON	563.72	-2.27	2772.22
TAIR	Air Temperature	Ameriflux/NEON	12.14	-36.39	41.85
VPD	Vapor Pressure Deficit	Ameriflux/NEON	8.48	-0.57	74.49
SWC_1_1_1	Soil Water Content	Ameriflux/NEON	19.74	0.25	40.96
PPFD_OUT	Photosynthetic Photon Flux Density, Outgoing	Ameriflux/NEON	60.92	-2.29	2054.03
PPFD_BC_IN_1_1_1	Photosynthetic Photon Flux Density, Below Canopy Incoming	Ameriflux/NEON	193.89	-9.44	2638.5
RH	Relative Humidity	Ameriflux/NEON	57.03	1.35	101.95
NETRAD	Net Radiation	Ameriflux/NEON	152.55	-308.42	1056.68
USTAR	Friction velocity	Ameriflux/NEON	0.46	0.05	2.78
GCC_50	Green Chromatic Coordinate, 50th Quantile	Phenocam	0.36	0.29	0.46
RCC_50	Red Chromatic Coordinate, 50th Quantile	Phenocam	0.4	0.26	0.58
MAT_DAYMET	Mean Annual Temperature	DAYMET	9.7	-11.6	26.1
MAP_DAYMET	Mean Annual Precipitation	DAYMET	872.85	86	2290
PVEG	Primary Vegetation Type	Phenocam			categorical
SVEG	Secondary Vegetation Type	Phenocam			categorical
LW_OUT	Longwave Radiation, Outgoing	Ameriflux/NEON	378.09	165.3	694.8
DAILY PRECIPITATION	Daily Precipitation	Ameriflux/NEON	2.2	0	225.19
PRCP1WEEK	Cummulative Precipitation 1 Week	Ameriflux/NEON	16.42	0	262.73
PRCP2WEEK	Cummulative Precipitation 2 Week	Ameriflux/NEON	33.59	0	324.87
NDVI	Normalized Difference Vegetation Index	MODIS	0.47	-0.2	0.96
EVI	Enhanced Vegetation Index	MODIS	0.26	-0.13	0.76
LAT	Latitude	Phenocam	41.19	17.97	71.28
LON	Longitude	Phenocam	-101.8	-156.62	-66.87
ELEV	Elevation	Phenocam	813.93	7	3493
DOMAIN	NEON Field Site Domain	Phenocam			categorical
organic_C	Total Organic Carbon Stock in Soil Profile	Ameriflux/NEON	255.87	5	1339
total_N	Total Nitrogen Stock in Soil Profile	Ameriflux/NEON	13.47	0.3	43.6
O_thickness	Total Thickness of Organic Horizon	Ameriflux/NEON	3.49	0	110
A_pH	pH of A Horizon	Ameriflux/NEON	6.03	0	8.5
A_sand	Texture of A Horizon (% Sand)	Ameriflux/NEON	47.78	0	97
A_silt	Texture of A Horizon (% Silt)	Ameriflux/NEON	32.57	0	61.9
A_clay	Texture of A Horizon (% Clay)	Ameriflux/NEON	15.08	0	55.3
A_BD	Bulk Density of A Horizon	Ameriflux/NEON	0.93	0	1.59



**Figure 2:** A visual explanation of the two cross-validation techniques used in our experiments.

maximum likelihood estimates for the coefficients. Linear regression assumes a linear relationship between the predictors and the response variable, which is unlikely in complex modeling problems, but does provide a baseline for the comparison of the performance of other models.

2. *Stepwise Linear Regression*: This model began by testing for the most significant single variable in a linear regression model, and then iteratively added variables and tested for greatest improvement. A threshold number of selection variables was set to 15 for this forward selection technique. In this way, we simplify the basic linear regression model to find feature variables with greater importance for linear prediction.
3. *Decision Tree*: A decision tree is a model based on recursively splitting the data on values of variables to maximize the difference between observations. Decision trees are most effective on problems where there is a non-linear relationship between the predictors and response variable (Nie et al., 2020; Vanli et al., 2019). The optimal tree depth was found to be 10 which was found through cross-validation.
4. *Random Forest*: A random forest model (Breiman, 2001) is a bagged ensemble of decision trees. The algorithm creates an uncorrelated forest of decision trees by using random subsets of features in each tree. When predicting a regression variable with a random forest model, the overall prediction is the average of the results of each of its constituent trees.
5. *Extreme Gradient Boosting (XGBoost)*: The XGBoost model (Chen and Guestrin, 2016) is a boosted ensemble of  $n$  underfit decision tree models. In practice, a decision tree is fit to the data and the errors in prediction are measured. Next, a second decision tree is used to fit the errors of the first tree. Then a third decision tree is fit to the errors of the second tree, and we continue until we have  $n$  trees in our ensemble. The optimal number of trees in our ensemble was found to be 2000. We also set the number of rounds for early stopping to be 50, and we used a learning rate of 0.05, max depth of 10, subsample ratio of 0.5, and subsample ratio of columns for each node of 0.45. Finally we used the histogram-optimized approximate greedy algorithm for tree construction to optimize our XGBoost model. All hyperparameters were optimized through 10-fold cross-validation using an exhaustive grid search.
6. *Neural Network (single-layer)*: A neural network is the sum of weighted non-linear functions of the predictor variables. This model is a single-layer neural network, with 256 neurons in the hidden layer, and uses a feed-forward architecture with ReLU activation. Early stopping was implemented to prevent model overfitting, and training was performed with a data loader with a batch size of 128. The learning rate was set to 0.0003, and the best performance was achieved with

no weight decay using the Adam optimizer. For more information on the mathematics of neural networks, see: James et al. (2021); Mahabbati et al. (2021).

7. *Deep Neural Network*: The model uses the same mathematical structure as the single-layer neural network, but increases the number of hidden layers to 3, each consisting of 256 neurons. Compared to the single-layer neural network, the increased depth of the model increases the number of parameters to learn, meaning the model is capable of modeling more complex relationships, but also takes longer to learn from the data.

### 3. Results

#### 3.1. 10-fold cross-validation results

The results for fitting each model and testing on each fold of the 10-fold cross-validation experiments are shown in Table 2 (RMSE). The XGBoost model and deep Neural Network were the only two models with a RMSE less than  $2\mu\text{molm}^{-2}\text{s}^{-1}$ . The strength of these models suggests that there are non-linearities in the relationships between the environmental drivers and FCO<sub>2</sub>. It is important to note that the XGBoost model outperformed our deep Neural Network at each stage throughout model development, and in addition the XGBoost model requires significantly less training time than either neural network.

After determining the optimal algorithm, we used the trained XGBoost model to gap-fill all of the missing values for each of the 44 sites. The resulting dataset, consisting of 4,068,459 observations, is freely available at <https://zenodo.org/records/10719776> for use by other researchers in the climate science community.

#### 3.2. L1SO cross-validation results

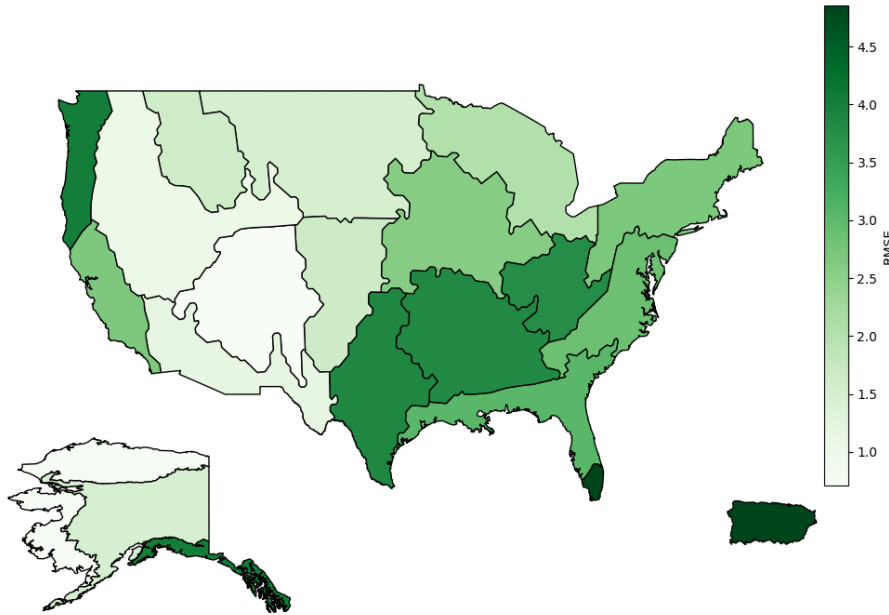
The results for fitting each model and testing on each site of the leave-one-site-out cross-validation experiments are shown in Tables 3 (RMSE) and 4 ( $R^2$ ). Again, the XGBoost model was superior to all others with a mean prediction RMSE of  $2.45\mu\text{molm}^{-2}\text{s}^{-1}$ . This is 35% greater than the RMSE of same model in the 10-fold cross-validation experiments, demonstrating the substantial information the model gains from seeing data from the test site in the training set (as is the case in the 10-fold experiments).

The results also varied greatly between test sites—from a RMSE of  $0.66\mu\text{molm}^{-2}\text{s}^{-1}$  up to  $6.22\mu\text{molm}^{-2}\text{s}^{-1}$ . The model performed best on Toolik (TOOL), as well as other sites with Tundra as the primary vegetation—Barrow Environmental Observatory (BARR), Healy (HEAL), and Niwot Ridge Mountain Research Station (NIWO)—suggesting that the environmental drivers for these sites are highly similar. Another justification for lower model RMSE across sites with Tundra primary vegetation is that these sites in general experience smaller magnitude fluxes. Random errors scale with flux magnitude, so it's almost inevitable that sites with higher magnitude fluxes will have somewhat larger model-data mismatch.

**Table 2**

Comparison of the RMSE and R<sup>2</sup> in predicting FCO<sub>2</sub> using seven machine learning models in a 10-fold cross-validation experimental setting (values shown are the average across the 10 validation folds)

	Linear reg	Stepwise	Decision Tree	Random Forest	XGB	NN 1-layer	NN deeper
RMSE	3.49	3.58	2.39	2.26	1.81	2.06	1.91
R <sup>2</sup>	0.48	0.46	0.76	0.77	0.86	0.82	0.85



**Figure 3:** The average RMSE ( $\mu\text{molm}^{-2}\text{s}^{-1}$ ) per domain for the leave-one-site-out experiments.

The model performed worst on Lajas Experimental Station (LAJA), which is one of two sites in Puerto Rico, and together these two represent the only two sites with an evergreen broadleaf primary vegetation type. While we cannot separate the domain and primary vegetation effects here, we can say that our training data, which is mostly from the United States mainland, does not generalize well when predicting FCO<sub>2</sub> in vastly different climates and ecosystems.

A map of the average RMSE per domain is shown in Figure 3.

### 3.3. XGBoost Feature Importance

The XGBoost algorithm has a built-in method for calculating the importance of each feature variable based on the amount that each feature's split point improves model performance. A plot of the twenty most important features for prediction is shown in Figure 4.

There are two input features that are noticeably more important to the model than others—EVI and net radiation. This is interesting as these are not measurements taken through site-level instrumentation, which suggests that we can learn a lot about the FCO<sub>2</sub> of a site just by knowing the greenness and thermal radiation of the vegetation. Furthermore, six of the ten most important variables are continuous measurement variables, as opposed to the

domain or vegetation categorical variables, meaning the model should generalize easier to any new sites of interest.

## 4. Discussion

In this section, we present a detailed discussion of the results of the XGBoost model at the site- and domain-level. We also analyze our results by vegetation type and how our results look on an annual scale.

### 4.1. Comparison of 10-fold and LISO experimental results

By making predictions on each site in both 10-fold and LISO contexts, we are able to gain greater understanding of model performance across the 44 NEON sites. We partitioned our model's 10-fold RMSE by site, and treated a site's average RMSE value as the irreducible error—that is, error that can be attributed to variability in the dataset, measurement errors, and the error inherent in using a model to predict a physical process. From there, we compare this irreducible error to the average RMSE values for each site obtained through LISO CV experiments and therefore obtain an estimate of the amount of error attributable to testing on an 'unseen' sight, which we call the LISO Remainder. A visualization of the baseline error and its

**Table 3**

A comparison of the RMSE ( $\mu\text{molm}^{-2}\text{s}^{-1}$ ) in predicting FCO<sub>2</sub> using seven machine learning models in a stratified leave-one-site-out cross-validation experimental setting

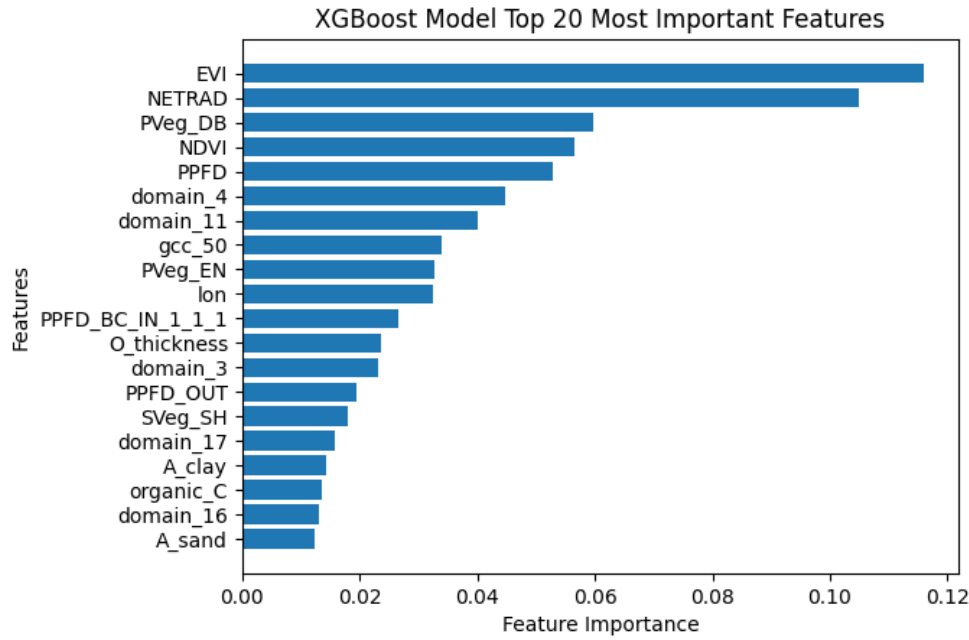
Test Set	Site Code	Site Name	Primary Vegtype	Linear reg	Stepwise	Decision Tree	Random Forest	XGB	NN 1-layer	NN deeper
1	PR-xGU	Guanica Forest (GUAN)	EB	4.83	4.47	5.83	5.32	3.49	5.95	6.48
2	PR-xLA	Lajas Experimental Station (LAJA)	EB	7.52	6.99	7.60	6.68	6.22	6.02	6.60
3	US-xAB	Abby Road (ABBY)	EN	7.25	4.45	4.72	3.86	3.43	3.55	3.66
4	US-xBA	Barrow Environmental Observatory (BARR)	TN	135.35	1.30	1.51	1.49	0.86	2.91	0.89
5	US-xBL	Blandy Experimental Farm (BLAN)	DB	4.10	3.96	2.77	2.69	2.62	2.89	2.98
6	US-xBN	Caribou Creek - Poker Flats Watershed (BONA)	EN	14.61	2.41	2.12	2.01	1.93	2.70	1.92
7	US-xBR	Bartlett Experimental Forest (BART)	DB	5.21	4.41	3.33	3.06	2.77	3.13	3.06
8	US-xCL	LBJ National Grassland (CLBJ)	DB	5.19	4.17	4.38	4.16	3.88	4.11	3.31
9	US-xCP	Central Plains Experimental Range (CPER)	GR	4.24	2.47	1.38	1.29	1.22	1.60	1.48
10	US-xDC	Dakota Coteau Field School (DCFS)	GR	20.35	2.70	1.79	1.70	1.61	1.64	1.74
11	US-xDJ	Delta Junction (DEJU)	EN	5.52	2.28	2.05	1.64	1.44	1.56	1.44
12	US-xDL	Dead Lake (DELA)	DB	9.86	5.29	4.36	4.21	3.84	4.23	4.26
13	US-xDS	Disney Wilderness Preserve (DSNY)	GR	10.21	3.03	3.64	3.25	3.33	2.67	3.35
14	US-xGR	Great Smoky Mountains National Park, Twin Creeks (GRSM)	DB	6.51	6.06	4.21	3.99	3.87	4.12	3.94
15	US-xHA	Harvard Forest (HARV)	DB	5.24	4.50	3.05	2.91	2.60	2.73	2.92
16	US-xHE	Healy (HEAL)	TN	5.03	1.72	2.00	1.65	1.15	1.77	1.17
17	US-xJE	Jones Ecological Research Center (JERC)	DB	6.07	4.37	3.75	3.46	3.19	3.43	3.41
18	US-xJR	Jornada LTER (JORN)	GR	2.56	1.79	1.25	1.23	1.17	1.76	1.26
19	US-xKA	Konza Prairie Biological Station - Relocatable (KONA)	AG	6.57	3.64	3.02	2.95	2.61	3.05	3.56
20	US-xKZ	Konza Prairie Biological Station (KONZ)	GR	6.88	3.57	2.60	2.23	2.21	2.06	2.16
21	US-xLE	Lenoir Landing (LENO)	DB	6.83	5.27	4.92	4.53	4.32	4.25	4.19
22	US-xMB	Moab (MOAB)	GR	8.63	1.86	0.73	0.71	0.68	1.54	0.68
23	US-xNG	Northern Great Plains Research Laboratory (NOGP)	GR	5.07	2.29	1.67	1.59	1.46	1.55	1.96
24	US-xNQ	Onaqui-Ault (ONAQ)	SH	4.01	1.73	1.17	1.11	1.05	1.90	1.21
25	US-xNW	Niwot Ridge Mountain Research Station (NIWO)	TN	9.63	1.46	0.85	0.80	0.74	1.86	1.76
26	US-xRM	Rocky Mountain National Park, CASTNET (RMNP)	EN	8.49	3.18	2.70	2.31	1.92	2.45	1.94
27	US-xRN	Oak Ridge National Lab (ORNL)	DB	5.75	5.11	4.43	4.22	3.68	3.92	3.61
28	US-xSB	Ordway-Swisher Biological Station (OSBS)	EN	7.77	3.40	3.06	2.78	2.63	3.17	3.08
29	US-xSC	Smithsonian Conservation Biology Institute (SCBI)	DB	4.53	4.11	3.36	3.00	2.86	3.12	2.98
30	US-xSE	Smithsonian Environmental Research Center (SERC)	DB	6.79	4.62	3.40	3.21	3.08	3.35	3.32
31	US-xSJ	San Joaquin Experimental Range (SJER)	EN	5.13	4.23	3.23	3.11	3.02	3.23	3.81
32	US-xSL	North Sterling, CO (STER)	AG	6.10	2.40	2.00	1.93	1.83	1.90	2.08
33	US-xSP	Soaproot Saddle (SOAP)	EN	3.57	3.58	4.16	3.86	2.50	2.78	2.67
34	US-xSR	Santa Rita Experimental Range (SRER)	SH	3.22	2.19	4.23	3.63	1.18	2.42	1.12
35	US-xST	Steigerwaldt Land Services (STEL)	DB	3.96	4.06	2.44	2.10	1.91	2.34	1.78
36	US-xTA	Talladega National Forest (TALL)	EN	5.36	5.16	4.53	4.33	3.34	3.77	3.98
37	US-xTE	Lower Teakettle (TEAK)	EN	6.11	3.07	2.99	2.93	2.53	2.48	2.95
38	US-xTL	Toolik (TOOL)	TN	134.54	1.44	1.24	0.79	0.66	2.12	0.96
39	US-xTR	Treehaven (TREE)	DB	5.13	3.89	2.41	2.35	2.12	2.61	2.21
40	US-xUK	The University of Kansas Field Station (UKFS)	DB	5.16	4.12	3.20	3.06	2.92	3.56	2.92
41	US-xUN	University of Notre Dame Environmental Research Center (UNDE)	DB	3.79	3.81	2.51	2.47	2.11	2.53	1.92
42	US-xWD	Woodworth (WOOD)	GR	5.16	2.21	1.77	1.61	1.49	1.52	1.70
43	US-xWR	Wind River Experimental Forest (WREF)	EN	7.53	5.31	5.89	5.82	4.67	4.92	4.68
44	US-xYE	Yellowstone Northern Range (Frog Rock) (YELL)	EN	5.05	2.49	2.10	2.05	1.61	1.71	1.74
AVERAGE				12.28	3.51	3.05	2.82	2.45	2.88	2.70

**Table 4**

A comparison of the R<sup>2</sup> in predicting FCO<sub>2</sub> using seven machine learning models in a stratified leave-one-site-out cross-validation experimental setting

Test Set	Site Code	Site Name	Primary Vegtype	Linear reg	Stepwise	Decision Tree	Random Forest	XGBoost	NN (1-layer)	NN (deep)
1	PR-xGU	Guanica Forest (GUAN)	EB	0.07	0.21	-0.35	-0.12	0.52	-0.40	-0.67
2	PR-xLA	Lajas Experimental Station (LAJA)	EB	0.31	0.40	0.29	0.45	0.53	0.56	0.47
3	US-xAB	Abby Road (ABBY)	EN	-0.37	0.48	0.42	0.61	0.69	0.67	0.65
4	US-xBA	Barrow Environmental Observatory (BARR)	TN	-16320.00	-0.51	-1.03	-0.97	0.34	-6.54	0.29
5	US-xBL	Blandy Experimental Farm (BLAN)	DB	0.54	0.57	0.79	0.80	0.81	0.77	0.76
6	US-xBN	Caribou Creek - Poker Flats Watershed (BONA)	EN	-33.28	0.07	0.28	0.35	0.40	-0.17	0.41
7	US-xBR	Bartlett Experimental Forest (BART)	DB	0.34	0.53	0.73	0.77	0.81	0.76	0.77
8	US-xCL	LBJ National Grassland (CLBJ)	DB	0.35	0.58	0.54	0.58	0.64	0.59	0.74
9	US-xCP	Central Plains Experimental Range (CPER)	GR	-4.44	-0.85	0.42	0.50	0.55	0.22	0.33
10	US-xDC	Dakota Coteau Field School (DCFS)	GR	-28.15	0.49	0.78	0.80	0.82	0.81	0.79
11	US-xDJ	Delta Junction (DEJU)	EN	-3.89	0.17	0.32	0.57	0.67	0.61	0.67
12	US-xDL	Dead Lake (DELA)	DB	-0.89	0.46	0.63	0.66	0.71	0.65	0.65
13	US-xDS	Disney Wilderness Preserve (DSNY)	GR	-3.07	0.64	0.48	0.59	0.57	0.72	0.56
14	US-xGR	Great Smoky Mountains National Park, Twin Creeks (GRSM)	DB	0.39	0.48	0.75	0.77	0.79	0.76	0.78
15	US-xHA	Harvard Forest (HARV)	DB	0.31	0.49	0.77	0.79	0.83	0.81	0.79
16	US-xHE	Healy (HEAL)	TN	-4.45	0.36	0.14	0.41	0.72	0.33	0.71
17	US-xJE	Jones Ecological Research Center (JERC)	DB	0.19	0.58	0.69	0.74	0.78	0.74	0.75
18	US-xJR	Jornada LTER (JORN)	GR	-2.75	-0.85	0.11	0.13	0.21	-0.77	0.09
19	US-xKA	Konza Prairie Biological Station - Relocatable (KONA)	AG	-1.33	0.28	0.51	0.53	0.63	0.50	0.31
20	US-xKZ	Konza Prairie Biological Station (KONZ)	GR	-0.85	0.50	0.74	0.81	0.81	0.83	0.82
21	US-xLE	Lenoir Landing (LENO)	DB	0.19	0.52	0.58	0.64	0.67	0.69	0.69
22	US-xMB	Moab (MOAB)	GR	-145.46	-5.79	-0.05	0.01	0.09	-3.66	0.09
23	US-xNG	Northern Great Plains Research Laboratory (NOGP)	GR	-2.17	0.36	0.66	0.69	0.74	0.71	0.52
24	US-xNQ	Onaqui-Ault (ONAQ)	SH	-7.30	-0.54	0.29	0.37	0.43	-0.87	0.25
25	US-xNW	Niwot Ridge Mountain Research Station (NIWO)	TN	-120.13	-1.77	0.05	0.17	0.28	-3.53	-3.04
26	US-xRM	Rocky Mountain National Park, CASTNET (RMNP)	EN	-5.45	0.09	0.35	0.52	0.67	0.46	0.66
27	US-xRN	Oak Ridge National Lab (ORNL)	DB	0.25	0.41	0.56	0.60	0.69	0.65	0.71
28	US-xSB	Ordway-Swisher Biological Station (OSBS)	EN	-1.39	0.54	0.63	0.69	0.73	0.60	0.62
29	US-xSC	Smithsonian Conservation Biology Institute (SCBI)	DB	0.42	0.52	0.68	0.74	0.77	0.72	0.75
30	US-xSE	Smithsonian Environmental Research Center (SERC)	DB	-0.01	0.53	0.75	0.77	0.79	0.75	0.76
31	US-xSJ	San Joaquin Experimental Range (SJER)	EN	-0.51	-0.03	0.40	0.44	0.47	0.40	0.17
32	US-xSL	North Sterling, CO (STER)	AG	-4.83	0.10	0.38	0.42	0.47	0.44	0.32
33	US-xSP	Soaproot Saddle (SOAP)	EN	-0.98	-0.98	-1.68	-1.31	0.03	-0.19	-0.10
34	US-xSR	Santa Rita Experimental Range (SRER)	SH	-7.73	-3.04	-14.04	-10.11	-0.18	-3.93	-0.06
35	US-xST	Steigerwaldt Land Services (STEL)	DB	0.53	0.50	0.82	0.87	0.89	0.83	0.90
36	US-xTA	Talladega National Forest (TALL)	EN	0.39	0.44	0.57	0.60	0.76	0.70	0.66
37	US-xTE	Lower Teakettle (TEAK)	EN	-2.27	0.17	0.22	0.25	0.44	0.46	0.24
38	US-xTL	Toolik (TOOL)	TN	-12181.30	-0.40	-0.03	0.58	0.71	-2.01	0.38
39	US-xTR	Treehaven (TREE)	DB	0.24	0.57	0.83	0.84	0.87	0.80	0.86
40	US-xUK	The University of Kansas Field Station (UKFS)	DB	0.24	0.52	0.71	0.73	0.76	0.64	0.76
41	US-xUN	University of Notre Dame Environmental Research Center (UNDE)	DB	0.56	0.55	0.81	0.81	0.86	0.80	0.89
42	US-xWD	Woodworth (WOOD)	GR	-2.01	0.45	0.65	0.71	0.75	0.74	0.67
43	US-xWR	Wind River Experimental Forest (WREF)	EN	-0.65	0.18	-0.01	0.02	0.37	0.30	0.36
44	US-xYE	Yellowstone Northern Range (Frog Rock) (YELL)	EN	-2.28	0.20	0.43	0.46	0.67	0.62	0.61
AVERAGE				-656.42	-0.02	0.06	0.23	0.60	-0.01	0.44





**Figure 4:** The twenty most important features of our XGBoost model

corresponding LISO remainder for each site, ordered by ecological domain, is shown in Figure 5.

The LISO remainder gives us a reasonable way to identify sites that are difficult to predict without having that site's data available in the training set. We identified five NEON terrestrial sites that had a LISO remainder greater than 0.85. These sites are Guanica Forest (GUAN), Lajas Experimental Station (LAJA), LBJ National Grassland (CLBJ), Disney Wilderness Preserve (DSNY), and Wind River Experimental Forest (WREF). There are several reasons that can justify why these sites in particular may be difficult for a model in a LISO scenario. Firstly, Guanica Forest and Lajas Experimental Station are the only two sites in Puerto Rico and in ecological domain 4. In addition, these two sites are the only two whose primary vegetation type is evergreen broadleaf (EB).

Wind River Experimental Forest is a site in Washington state, located in an old growth forest with very tall trees with a real summer dry-down that restricts FCO<sub>2</sub>. Overall, Wind River Experimental Forest is a very unusual site in comparison with the other NEON terrestrial sites.

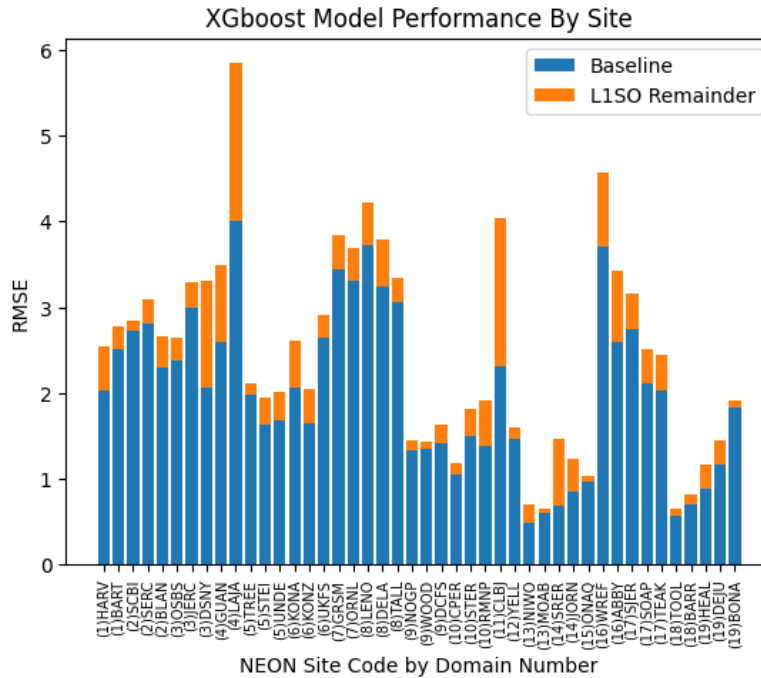
For each of these five sites, we created time series of predicted FCO<sub>2</sub> values and actual FCO<sub>2</sub> values reported both in half hourly increments, and aggregated as an average for each day of the year, as well as a scatter plot of predicted FCO<sub>2</sub> vs. actual FCO<sub>2</sub> for analysis. We then compared these results to sites with the same primary vegetation types for which our model had superior performance. In the case of Guanica Forest and Lajas Experimental Station, since there were no other sites with the same primary vegetation type, both sites are included in Figure 6.

Steigerwaldt Land Services (STEI), Dakota Coteau Field School (DCFS), and Delta Junction (DEJU) were used as

comparison for our other three sites representing primary vegetation types of DB, GR and EN respectively. These comparisons are found in Figures 7, 8, and 9.

Note that in most cases we observed large systematic errors in model performance for our 5 sites with the greatest LISO Remainder values. For example, when considering scatterplots of predicted vs observed FCO<sub>2</sub> for LAJA and WREF, the slope of predicted vs observed FCO<sub>2</sub> is less than 1. At CLBJ, the magnitude of summertime uptake is under-predicted. At DSNY, the seasonality is represented well but there is a consistent offset of several  $\mu\text{molm}^{-2}\text{s}^{-1}$ , with predicted values higher than measured values. By comparison, at STEI, DCFS, and DEJU, the magnitude and timing of predicted FCO<sub>2</sub> is much better.

What is interesting from this analysis is that even on sites with relatively high LISO remainder, our model seems to do a good job on average predicting patterns and dips in daily average FCO<sub>2</sub>. It appears that most of the errors associated with sites with large LISO remainder can be attributed to the model being too conservative in its predictions, that is predicting values closer to zero than the true measured flux values. As seen by the right column of plots in Figures 7-9, sites of the same primary vegetation type where our model had stronger performance seem to generally have less large positive and negative flux values. This makes sense, since our model learns to minimize prediction error, and since each error ends up being squared, predicting very large positive or negative values in general would be more heavily penalized. A good example of this fact can be seen in the half-hourly time series for Lajas Experimental Station in Figure 6. This site has a mix of large positive and negative observed flux values, and our model rarely made large positive or negative predictions. Compare this



**Figure 5:** Visualization of XGBoost L1SO RMSE remainder organized by domain number (shown as a prefix to the site code)

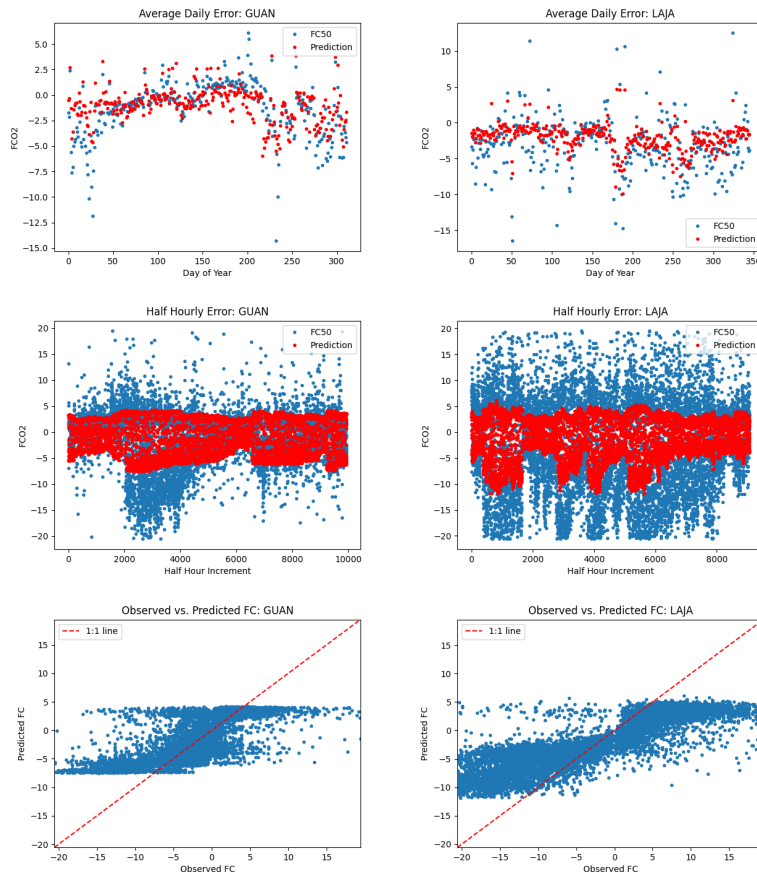
to a site like Delta Junction in Figure 8. Here, there are a number of large negative observed flux values, but not as many large positive values. Spikes in the negative direction are less erratic, and model predictions, as a result, more closely represented measured flux values. When looking at scatter plots of predicted flux vs. observed flux, one can see that the results for sites in the right hand column are more tightly clustered about the 1:1 line, resulting in higher  $R^2$  scores.

#### 4.2. Relevance of L1SO predictions for unseen sites

Historically, process-based models have been considered the “gold standard” for predicting ecosystem CO<sub>2</sub> fluxes. However, past model-data evaluation studies have shown that although process-based models can often predict daily- or sub-daily fluxes that agree reasonably well with measured values, model performance on longer time scales (seasonal, annual and inter-annual) is often quite poor (Dietze et al. (2011); Stoy et al. (2013); Keenan et al. (2012)). Models that cannot accurately predict ecosystem carbon budgets on annual and inter-annual time scales are not likely to be useful for carbon accounting purposes or for developing strategies for nature-based climate solutions. This suggests that alternatives to process-based models are needed. While machine learning-based models have been used for flux upscaling for almost two decades (Papale and Valentini (2003); Xiao et al. (2008); Jung et al. (2020)), these analyses have generally attempted to extrapolate from individual sites to regions and continents using only remotely-sensed variables as drivers. While this strategy is intuitively appealing, it is unable to leverage

the site-level characteristics that are undoubtedly relevant for making fine-scale predictions. Indeed, basic ecosystem theory suggests that without accounting for these site-level characteristics such as disturbance and land use history, it is impossible to predict ecosystem carbon balance. Notably, we found that site characteristics related to vegetation type, as well as to soils, were identified as among the most important features for predicting FCO<sub>2</sub>. However, remotely-sensed variables from MODIS such as EVI and NDVI were found to be more important than site-level vegetation indices (e.g. Gcc, Rcc) derived from PhenoCam imagery. We can hypothesize that while PhenoCam imagery can provide phenological information at a fine spatial and temporal scale, it may be subject to issues related to the mismatch of footprints with eddy covariance flux measurements. In the case of heterogeneous landscapes, MODIS vegetation indices with larger spatial coverage may actually be more representative of seasonal variations in vegetation dynamics within the flux tower footprint.

Finally, we note that although site-level meteorological and environmental drivers (e.g. air temperature, relative humidity or VPD, soil temperature, soil moisture, and precipitation) were not ranked highly in terms of feature importance, this is not to say that these variables do not matter. Rather, it is likely that in the context of variation in FCO<sub>2</sub> from the Arctic to the Tropics, from winter to summer, and from day to night, that the additional information contributed by these variables explains only a small amount of the half-hourly variation in FCO<sub>2</sub>, although it may contribute greatly to improved estimates of annual FCO<sub>2</sub>.



**Figure 6:** Time series and scatter plots of FCO<sub>2</sub> prediction error for sites with EB primary vegetation type

A persistent challenge in estimating site-level carbon balance via FCO<sub>2</sub> measurements has always been that small but selectively systematic measurement errors in 30 minute data can accumulate to large errors in annual integrals Richardson et al. (2012a). In our machine learning approach, selectively systematic prediction errors could occur if important meteorological or environmental variables were not accounted for as covariates. Omission of these variables might do little to impact the  $R^2$  calculated on 30 minute values but could seriously impact annual flux integrals. Adoption of model optimization criteria that place more weight on reducing selectively systematic bias (which might not even show up when bias is calculated over a multi-year dataset) and improving predictive power on annual and multi-year time scales could be important for further improving the application of machine learning methods to carbon accounting and nature-based climate solutions.

### 4.3. Leveraging Site-level Data when Standardized Model Inputs are not Available

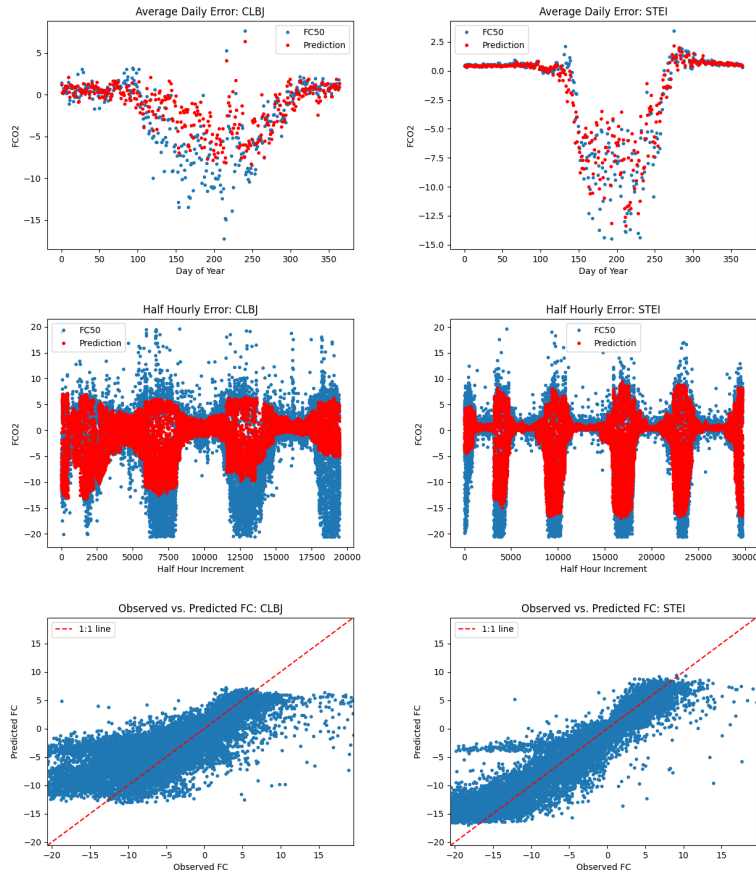
Our feature importance plot (Figure 4) shows that, in spite of our assertion that site-level data are critical for correctly predicting ecosystem carbon balance, much of the information needed to predict half-hourly FCO<sub>2</sub> actually comes from variables that are already available

from gridded land cover maps (i.e. vegetation type classifications), satellite data products characterizing phenology (i.e. EVI, NDVI), and basic energy balance data that are also widely available as satellite data products (e.g., net radiation). This suggests that there is the potential for leveraging the much greater abundance of AmeriFlux towers, (for which site-level measurements are not standardized), together with key remotely sensed data products to generate an initial map of ecosystem carbon balance. This initial map, when fused with elements of the analysis presented here, could lead to a hybrid data product that leverages the sampling intensity of AmeriFlux and the standardized sampling of NEON. Development of a data fusion platform such as we describe here is beyond the scope of the present analysis, but it is potentially an exciting direction to be pursued in future research.

### 4.4. Annual carbon sums

For most sites, we managed to obtain low RMSE and high  $R^2$  for predicting the measured half-hourly FCO<sub>2</sub>, even in the L1SO analysis (Tables 3 and 4). However, in the context of carbon accounting and nature-based climate solutions, it more important to know the overall carbon balance on an annual time scale. That is, we want to answer the question of how much carbon (if any) the ecosystem is removing from the atmosphere and putting into biomass

## ML-based modeling of CO<sub>2</sub> exchange



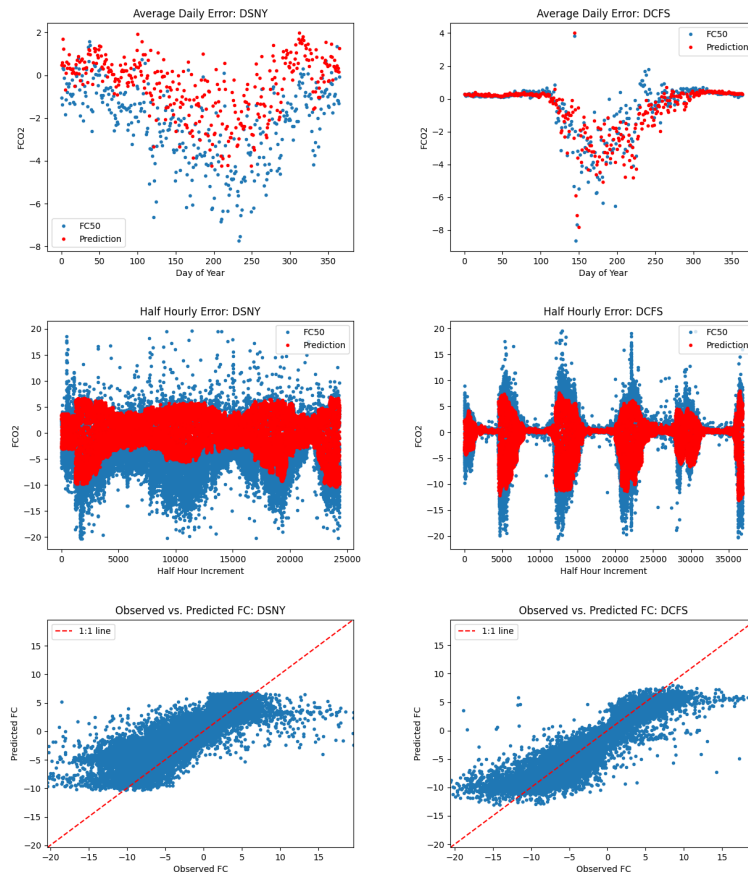
**Figure 7:** Comparison of FCO<sub>2</sub> prediction error for sites with DB primary vegetation type. Site with poor model performance (CLBJ) is on the left and site with better model performance (STEI) on the right.

and soil carbon on an annual basis. This carbon balance reflects the balance between plant photosynthesis (carbon uptake, or negative flux) and ecosystem respiration (carbon release, or positive flux). It is a challenge for models, either process-based or data-driven, to get the overall carbon balance correct because of the opposing nature of these processes on different timescales. For example, in most ecosystems there is a strong seasonal pattern of carbon uptake during the growing season and release during the dormant season. During the growing season there is also a diurnal pattern of carbon uptake during the day and release during the night. Annually, the difference between photosynthesis and respiration is much smaller (0-30%) than the flux associated with either of these two key processes.

A model that predicts the annual carbon balance for an unknown site would be extremely valuable if it successfully estimated the multi-year mean carbon balance. The model would be even more useful if it successfully represented the inter-annual variability in carbon balance. State-of-the-art process-based models have generally failed to meet either of these targets (Keenan et al., 2012). Our results show that across all vegetation types, annual sums predicted in the L1SO analysis did a surprisingly good job at hitting the first target (see Table 5). For 29 out of 44

sites (66%), the L1SO-predicted multi-year mean carbon balance was within  $\pm 50\text{gCm}^{-2}\text{y}^{-1}$  of the “true” value estimated by gap-filling missing values in the CV analysis. This is quite remarkable given that the total uncertainty on the annual carbon balance, derived from gap-filled FCO<sub>2</sub> measurements, is typically estimated to be about  $\pm 50\text{gCm}^{-2}\text{y}^{-1}$  (Richardson et al., 2012b). However, for 7 of 44 sites (16%), the deviation between the L1SO-predicted multi-year mean and the “true” value was greater than  $150\text{gCm}^{-2}\text{y}^{-1}$ . Three of these were deciduous broadleaf forest sites, one was an evergreen needleleaf forest site, and one was a grassland site. We expect that there may be land use history, disturbance, or similar factors that might explain these deviations, but were not included in our model.

Annual sums predicted in the L1SO analysis also did a reasonably good job of representing the “true” inter-annual variability estimated from gap-filled time series. At more than a quarter of sites (12 of 44, 27%), the correlation of L1SO-predicted annual sums and the gap-filled annual sums was greater than 0.75, while for almost half of sites (21 of 44, 47%) the correlation was greater than 0.50. While these results are based on at most 5 years of data per site, they point to the enormous potential of machine learning to predict not only the long-term carbon balance



**Figure 8:** Comparison of FCO<sub>2</sub> prediction error for sites with GR primary vegetation type. Site with poor model performance (DSNY) is on the left and site with better model performance (DCFS) on the right.

of an unknown site, but even the inter-annual variation in that carbon balance. By comparison, it has been known for more than a decade that even the most sophisticated process-based models are unable to capture this inter-annual variability (Braswell et al., 2005; Siqueira et al., 2006; Ricciuto et al., 2008), despite accurately capturing the dynamics of “fast” processes operating on timescales of hours to days.

## 5. Conclusions

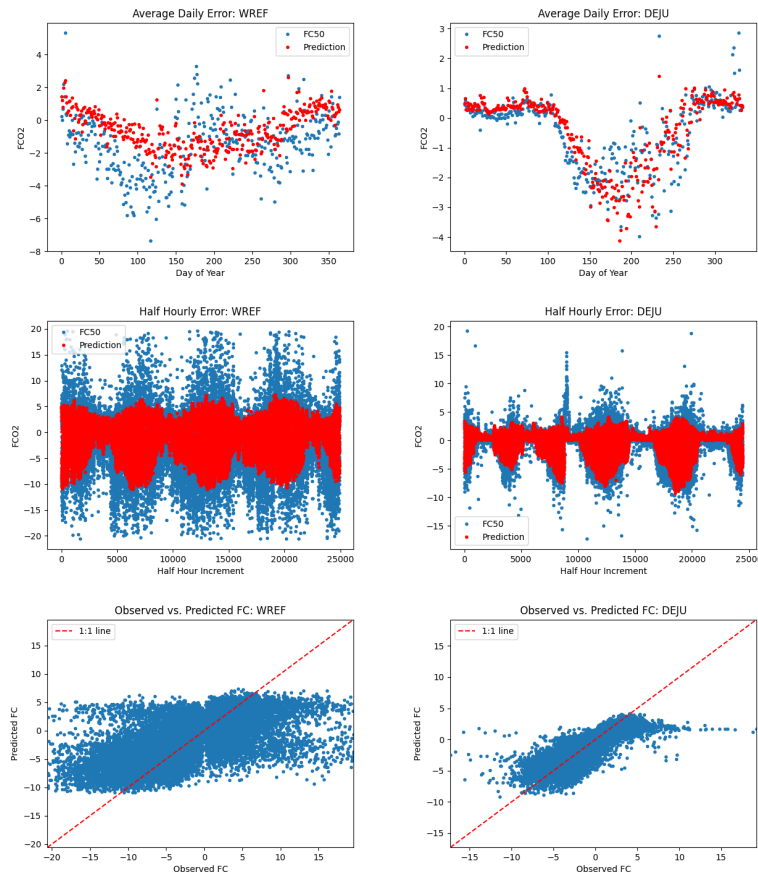
In this paper, we showed the potential for machine learning-based models to make more skillful predictions of FCO<sub>2</sub> than state-of-the-art process-based models. Specifically, we found that an XGBoost model trained on environmental drivers recorded at 43 locations from varying ecological domains can predict FCO<sub>2</sub> at an unseen site to within an average error of  $2.45\mu\text{molm}^{-2}\text{s}^{-1}$ . Furthermore, this error reduces significantly—down to as little as  $0.66\mu\text{molm}^{-2}\text{s}^{-1}$ —when a site in the training data has similar ecological characteristics to the unseen sites. This suggests that, with strategic placement of instrumentation to record future training data, there is potential to predict most locations of interest with high accuracy. Our research underscores the importance

of integrating advanced modeling techniques into carbon accounting frameworks, enabling more accurate quantification of carbon sequestration potential and guiding the implementation of effective nature-based climate mitigation strategies.

## CRedit authorship contribution statement

**Jeffrey Uyekawa:** Data Curation, Formal Analysis, Validation, Writing - Original Draft, Writing - Review & Editing. **John Leland:** Data Curation, Formal Analysis, Validation, Writing - Review & Editing. **Darby Bergl:** Data Curation, Resources, Writing - Review & Editing. **Yujie Liu:** Resources, Data Curation, Writing - Review & Editing. **Andrew D. Richardson:** Project administration, Funding Acquisition, Conceptualization, Writing - Original Draft, Resources, Writing - Review & Editing. **Benjamin Lucas:** Project administration, Supervision, Methodology, Validation, Writing - Original Draft, Writing - Review & Editing.

## ML-based modeling of CO<sub>2</sub> exchange



**Figure 9:** Comparison of sites with EN primary vegetation type. Site with poor model performance (WREF) is on the left and site with better model performance (DEJU) on the right.

### Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Data Availability

All processed data and code is at <https://github.com/js1339/AmeriFlux>. The gap-filled dataset has been made available for download at <https://zenodo.org/records/10719776>.

### Acknowledgements

This project was funded by NSF awards 1702697 and 2105828.

### References

Baldocchi DD (2020) How eddy covariance flux measurements have contributed to our understanding of global change biology. *Global Change Biology* 26(1):242–260

Battelle (2024) National Science Foundation's National Ecological Observatory Network (NEON). <https://www.neonscience.org/>

Bossio D, Cook-Patton S, Ellis P, Fargione J, Sanderman J, Smith P, Wood S, Zomer R, Von Unger M, Emmer I, et al. (2020) The role of soil carbon in natural climate solutions. *Nature Sustainability* 3(5):391–398

Braswell BH, Sacks WJ, Linder E, Schimel DS (2005) Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. *Global Change Biology* 11(2):335–355

Breiman L (2001) Random forests. *Machine learning* 45:5–32

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

Chu H, Christianson DS, Cheah YW, Pastorello G, O'Brien F, Geden J, Ngo ST, Hollowgrass R, Leibowitz K, Beekwilder NF, et al. (2023) Ameriflux base data pipeline to support network growth and data sharing. *Scientific Data* 10(1):614

Dietze MC, Vargas R, Richardson AD, Stoy PC, Barr AG, Anderson RS, Arain MA, Baker IT, Black TA, Chen JM, et al. (2011) Characterizing the performance of ecosystem models across time scales: A spectral analysis of the north american carbon program site-level synthesis. *Journal of Geophysical Research: Biogeosciences* 116(G4)

Fargione JE, Bassett S, Boucher T, Bridgman SD, Conant RT, Cook-Patton SC, Ellis PW, Falcucci A, Fourqurean JW, Gopalakrishna T, et al. (2018) Natural climate solutions for the united states. *Science Advances* 4(11):eaat1869

Fer I, Kelly R, Moorcroft PR, Richardson AD, Cowdery EM, Dietze MC (2018) Linking big models to big data: efficient ecosystem model calibration through bayesian model emulation. *Biogeosciences* 15(19):5801–5830

Friedlingstein P, O'Sullivan M, Jones MW, Andrew RM, Bakker DCE, Hauck J, Landschützer P, Le Quéré C, Lujckx IT, Peters GP, Peters W, Pongratz J, Schwingshackl C, Sitch S, Canadell JG, Ciais P, Jackson RB, Alin SR, Anthoni P, Barbero L, Bates NR, Becker M, Bellouin N, Decharme B, Bopp L, Brasika IBM, Cadule P, Chamberlain MA,

**Table 5**

Table of Mean Bias and Correlation Coefficient (r) using L1SO predicted annual carbon sums and 10-fold projections of annual carbon sums.

Primary Vegetation	Site	Mean Bias	R	
AG	US-xSL	-15.7958525	0.583905043	
	US-xKA	4.735793977	0.219740126	
	AVERAGE	-5.53 ± 10.266	0.402 ± 0.182	
DB	US-xSC	-60.81559653		
	US-xLE	134.1225374		
	US-xJE	76.42205637	0.680557913	
	US-xHA	-46.70929896	0.32489171	
	US-xGR	20.46464793	0.05640022	
	US-xRN	-67.14514317	0.82442625	
	US-xDL	55.56749305	-0.560025347	
	US-xST	21.37893181	0.745566851	
	US-xSE	17.42608627	-0.363818862	
	US-xCL	170.9445378	-0.781250941	
	US-xBR	114.3802585	0.854848083	
	US-xTR	-3.153405104	0.957922058	
	US-xBL	135.4473751	0.018731614	
	US-xUK	1.47867905	0.982685652	
US-xUN	4.947445999	-0.44166956		
	AVERAGE	38.317 ± 71.717	0.254 ± 0.635	
EB	PR-xLA	140.69792		
	PR-xGU	31.93251784		
	AVERAGE	86.315 ± 54.383		
EN	US-xSB	121.0153978	-0.203040655	
	US-xSP	-44.92273294	0.480016019	
	US-xTA	-68.90194021	0.32647278	
	US-xTE	47.39909206	-0.350431189	
	US-xSJ	-15.43605605	-0.650856478	
	US-xRM	-48.54995031	-0.667693298	
	US-xYE	4.313300343	0.573007411	
	US-xDJ	20.68358591	0.242707529	
	US-xWR	-10.30877542	-0.923173358	
	US-xAB	52.29245394	0.098859792	
	US-xBN	-18.25977483	0.558453915	
		AVERAGE	3.575 ± 51.99	-0.047 ± 0.514
	GR	US-xWD	-5.97743641	0.622368211
US-xCP		-9.58174608	0.517780215	
US-xDC		21.98706387	0.882080178	
US-xMB		17.88061649	0.982363967	
US-xDS		230.2646877	-0.901199761	
US-xJR		28.81973346	0.629948639	
US-xKZ		19.33977652	-0.624307388	
US-xNG		34.73729593	0.854563415	
		AVERAGE	42.184 ± 72.567	0.37 ± 0.674
SH	US-xSR	-61.36605941	0.995706643	
	US-xNQ	63.51041138	0.988233651	
	AVERAGE	1.072 ± 62.438	0.992 ± 0.004	
TN	US-xNW	-28.12476053	0.950683757	
	US-xHE	-12.38685133	0.765762711	
	US-xTL	-22.11903194	0.536714819	
	US-xBA	-29.72485677	0.805185014	
	AVERAGE	-23.089 ± 6.798	0.765 ± 0.148	

Chandra N, Chau TTT, Chevallier F, Chini LP, Cronin M, Dou X, Enyo K, Evans W, Falk S, Feely RA, Feng L, Ford DJ, Gasser T, Ghattas J, Gkritzalis T, Grassi G, Gregor L, Gruber N, Gürses O, Harris I, Hefner M, Heinke J, Houghton RA, Hurtt GC, Iida Y, Ilyina T, Jacobson AR, Jain A, Jarníková T, Jersild A, Jiang F, Jin Z, Joos F, Kato E, Keeling RF, Kennedy D, Klein Goldewijk K, Knauer J, Korsbakken JI, Körtzinger A, Lan X, Lefèvre N, Li H, Liu J, Liu Z, Ma L, Marland G, Mayot N, McGuire PC, McKinley GA, Meyer G, Morgan EJ, Munro DR, Nakaoka SI, Niwa Y, O'Brien KM, Olsen A, Omar AM, Ono T, Paulsen M, Pierrot D, Pocock K, Poulter B, Powis CM, Rehder G, Resplandy L, Robertson E, Rödenbeck C, Rosan TM, Schwinger J, Séférian R, Smallman TL, Smith SM, Sospedra-Alfonso R, Sun Q, Sutton AJ, Sweeney C, Takao S, Tans PP, Tian H, Tilbrook B, Tsujino H, Tubiello F, van der Werf GR, van Ooijen E, Wanninkhof R, Watanabe M, Wimart-Rousseau C, Yang D, Yang X, Yuan W, Yue X, Zaehle S, Zeng J, Zheng B (2023) Global carbon budget 2023. *Earth System Science Data* 15(12):5301–5369,

DOI 10.5194/essd-15-5301-2023, URL <https://essd.copernicus.org/articles/15/5301/2023/>

Griscom BW, Adams J, Ellis PW, Houghton RA, Lomax G, Miteva DA, Schlesinger WH, Shoch D, Siikamäki JV, Smith P, et al. (2017) Natural climate solutions. *Proceedings of the National Academy of Sciences* 114(44):11645–11650

Hemes KS, Runkle BR, Novick KA, Baldocchi DD, Field CB (2021) An ecosystem-scale flux measurement strategy to assess natural climate solutions. *Environmental science & technology* 55(6):3494–3504

Hollinger D, Davidson E, Fraver S, Hughes H, Lee J, Richardson A, Savage K, Sihi D, Teets A (2021) Multi-decadal carbon cycle measurements indicate resistance to external drivers of change at the howland forest ameriflux site. *Journal of Geophysical Research: Biogeosciences* 126(8):e2021JG006276

James G, Witten D, Hastie T, Tibshirani R (2021) *An Introduction to Statistical Learning: with Applications in R*: 2nd Edition. Springer, URL

- <https://faculty.marshall.usc.edu/gareth-james/ISL/>
- Jung M, Schwalm C, Migliavacca M, Walther S, Camps-Valls G, Koirala S, Anthoni P, Besnard S, Bodesheim P, Carvalhais N, Chevallier F, Gans F, Goll DS, Haverd V, Köhler P, Ichii K, Jain AK, Liu J, Lombardozzi D, Nabel JEMS, Nelson JA, O'Sullivan M, Pallandt M, Papale D, Peters W, Pongratz J, Rödenbeck C, Sitch S, Tramontana G, Walker A, Weber U, Reichstein M (2020) Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the fluxcom approach. *Biogeosciences* 17(5):1343–1365, DOI 10.5194/bg-17-1343-2020, URL <https://bg.copernicus.org/articles/17/1343/2020/>
- Kang Y, Gaber M, Bassiouni M, Lu X, Keenan T (2023) Cedar-gpp: spatiotemporally upscaled estimates of gross primary productivity incorporating co<sub>2</sub> fertilization. *Earth System Science Data Discussions* 2023:1–51
- Keenan T, Baker I, Barr A, Ciais P, Davis K, Dietze M, Dragoni D, Gough CM, Grant R, Hollinger D, et al. (2012) Terrestrial biosphere model performance for inter-annual variability of land-atmosphere CO<sub>2</sub> exchange. *Global Change Biology* 18(6):1971–1987
- Lee H, Calvin K, Dasgupta D, Krinner G, Mukherji A, Thorne P, Trisos C, Romero J, Aldunce P, Barret K, Blanco G, Cheung WW, Connors SL, Denton F, Diongue-Niang A, Dodman D, Garschagen M, Geden O, Hayward B, Jones C, Jotzo F, Krug T, Lasco R, Lee YY, Masson-Delmotte V, Meinshausen M, Mintenbeck K, Mokssit A, Otto FE, Pathak M, Pirani A, Poloczanska E, Pörtner HO, Revi A, Roberts DC, Roy J, Ruane AC, Skea J, Shukla PR, Slade R, Slangen A, Sokona Y, Sörensson AA, Tignor M, van Vuuren D, Wei YM, Winkler H, Zhai P, Zommers Z, Hourcade JC, Johnson FX, Pachauri S, Simpson NP, Singh C, Thomas A, Totin E, Arias P, Bustamante M, Elgizouli I, Flato G, Howden M, Méndez-Vallejo C, Pereira JJ, Pichs-Madruga R, Rose SK, Saheb Y, Rodríguez RS, Ürge-Vorsatz D, Xiao C, Yassaa N, Alegría A, Armour K, Bednar-Friedl B, Blok K, Cissé G, Dentener F, Eriksen S, Fischer E, Garner G, Guivarch C, Haasnoot M, Hansen G, Hauser M, Hawkins E, Hermans T, Kopp R, Leprince-Ringuet N, Lewis J, Ley D, Ludden C, Niamir L, Nicholls Z, Some S, Szopa S, Trewin B, van der Wijst KI, Winter G, Witting M, Birt A, Ha M, Romero J, Kim J, Haites EF, Jung Y, Stavins R, Birt A, Ha M, Orendain DJA, Ignon L, Park S, Park Y (2023) *ipcc, 2023: Climate change 2023: Synthesis report, summary for policymakers. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change [core writing team, h. lee and j. romero (eds.)]. ipcc, geneva, switzerland. Technical report, Intergovernmental Panel on Climate Change (IPCC), Geneva, Switzerland*
- Mahabbi A, Beringer J, Leopold M, McHugh I, Cleverly J, Isaac P, Izady A (2021) A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers. *Geoscientific Instrumentation, Methods and Data Systems* 10(1):123–140, DOI 10.5194/gi-10-123-2021, URL <https://gi.copernicus.org/articles/10/123/2021/>
- National Ecological Observatory Network (NEON) (2024) Bundled data products - eddy covariance (dp4.00200.001). DOI 10.48443/J9PT-M241, URL <https://data.neonscience.org/data-products/DP4.00200.001/RELEASE-2024>
- Nie F, Zhu W, Li X (2020) Decision tree svm: An extension of linear svm for non-linear classification. *Neurocomputing* 401:153–159
- Novick KA, Biederman J, Desai A, Litvak M, Moore DJ, Scott R, Torn M (2018) The AmeriFlux network: A coalition of the willing. *Agricultural and Forest Meteorology* 249:444–456
- Papale D, Valentini R (2003) A new assessment of european forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biology* 9(4):525–535
- Ricciuto DM, Davis KJ, Keller K (2008) A bayesian calibration of a simple carbon cycle model: The role of observations in estimating and reducing uncertainty. *Global biogeochemical cycles* 22(2)
- Richardson AD, Anderson RS, Arain MA, Barr AG, Bohrer G, Chen G, Chen JM, Ciais P, Davis KJ, Desai AR, et al. (2012a) Terrestrial biosphere models need better representation of vegetation phenology: results from the north american carbon program site synthesis. *Global Change Biology* 18(2):566–584
- Richardson AD, Aubinet M, Barr AG, Hollinger DY, Ibrom A, Lasslop G, Reichstein M (2012b) Uncertainty quantification. In: Aubinet M, Vesala T, Papale D (eds) *Eddy Covariance: A Practical Guide to Measurement and Data Analysis*, Springer Netherlands, Dordrecht, pp 173–209, DOI 10.1007/978-94-007-2351-1\_7, URL [https://doi.org/10.1007/978-94-007-2351-1\\_7](https://doi.org/10.1007/978-94-007-2351-1_7)
- Richardson AD, Hufkens K, Milliman T, Aubrecht DM, Chen M, Gray JM, Johnston MR, Keenan TF, Klosterman ST, Kosmala M, et al. (2018) Tracking vegetation phenology across diverse north american biomes using phenocam imagery. *Scientific data* 5(1):1–24
- Rodriguez JD, Perez A, Lozano JA (2009) Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence* 32(3):569–575
- Schaefer K, Schwalm CR, Williams C, Arain MA, Barr A, Chen JM, Davis KJ, Dimitrov D, Hilton TW, Hollinger DY, et al. (2012) A model-data comparison of gross primary productivity: Results from the north american carbon program site synthesis. *Journal of Geophysical Research: Biogeosciences* 117(G3)
- Schimel DS, House JI, Hibbard KA, Bousquet P, Ciais P, Peylin P, Braswell BH, Apps MJ, Baker D, Bondeau A, et al. (2001) Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature* 414(6860):169–172
- Schwalm CR, Williams CA, Schaefer K, Anderson R, Arain MA, Baker I, Barr A, Black TA, Chen G, Chen JM, et al. (2010) A model-data intercomparison of CO<sub>2</sub> exchange across north america: Results from the north american carbon program site synthesis. *Journal of Geophysical Research: Biogeosciences* 115(G3)
- Seyednasrollah B, Young AM, Hufkens K, Milliman T, Friedl MA, Frohling S, Richardson AD (2019) Tracking vegetation phenology across diverse biomes using version 2.0 of the phenocam dataset. *Scientific data* 6(1):222
- Siqueira M, Katul GG, Sampson D, Stoy PC, Juang JY, McCarthy HR, Oren R (2006) Multiscale model intercomparisons of CO<sub>2</sub> and H<sub>2</sub>O exchange rates in a maturing southeastern us pine forest. *Global Change Biology* 12(7):1189–1207
- Stoy PC, Dietze MC, Richardson AD, Vargas R, Barr AG, Anderson RS, Arain MA, Baker IT, Black TA, Chen JM, Cook RB, Gough CM, Grant RF, Hollinger DY, Izaurreal RC, Kucharik CJ, Lafleur P, Law BE, Liu S, Lokupitiya E, Luo Y, Munger JW, Peng C, Poulter B, Price DT, Ricciuto DM, Riley WJ, Sahoo AK, Schaefer K, Schwalm CR, Tian H, Verbeeck H, Weng E (2013) Evaluating the agreement between measurements and models of net ecosystem exchange at different times and timescales using wavelet coherence: an example using data from the north american carbon program site-level interim synthesis. *Biogeosciences* 10(11):6893–6909, DOI 10.5194/bg-10-6893-2013, URL <http://dx.doi.org/10.5194/bg-10-6893-2013>
- United States Department of Energy (2023) AmeriFlux Management Project. <https://ameriflux.lbl.gov/>
- Vanli ND, Sayin MO, Mohaghegh M, Ozkan H, Kozat SS (2019) Nonlinear regression via incremental decision trees. *Pattern Recognition* 86:1–13
- Wofsy SC, Harris RC (2002) The north american carbon program 2002. Tech. rep., The Global Carbon Project, URL <https://www.globalcarbonproject.org/global/pdf/thenorthamericancprogram2002.pdf>
- Xiao J, Zhuang Q, Baldocchi DD, Law BE, Richardson AD, Chen J, Oren R, Starr G, Noormets A, Ma S, et al. (2008) Estimation of net ecosystem carbon exchange for the conterminous united states by combining modis and ameriflux data. *Agricultural and Forest Meteorology* 148(11):1827–1847