

Article

Globally Scalable Approach to Estimate Net Ecosystem Exchange Based on Remote Sensing, Meteorological Data, and Direct Measurements of Eddy Covariance Sites

Ruslan Zhuravlev ¹^{*}, Andrey Dara ^{1,2}, André Luís Diniz dos Santos ^{1,3}, Oleg Demidov ¹ and George Burba ⁴

¹ Carbonspace Ltd, Dublin, Ireland
² German Climate Computing Center (DKRZ), Hamburg, 20146, Germany
³ Federal University of Paraná, Curitiba, Paraná, Brazil
⁴ LI-COR Biosciences, Daugherty Water for Food Global Institute, University of Nebraska-Lincoln, Lincoln, United States
^{*} Correspondence: ruslan@carbonspace.tech

Abstract: Despite a rapid rise in NBS development in recent years, the methods for evaluating NBS still have certain gaps. We propose an approach based on a combination of remote sensing data and meteorological variables to reconstruct the spatio-temporal variation of net ecosystem exchange from eddy-covariance stations. Lagrangian particle dispersion model was used for upscaling of satellite images and flux towers. We trained data-driven models based on kernel methods separately for each selected land cover class. The results suggest that the proposed approach to quantifying carbon exchange on a medium-to-large scale by blending eddy covariance flux data with moderate resolution satellite and weather data provides a set of key advantages over previously deployed methods: (1) scalability, achieved via the validation design based on a separate set of eddy covariance stations; (2) high spatial and temporal resolution due to use of the Landsat imagery; (3) robust and accurate predictions due to improved data quality control, advanced machine learning techniques, and rigorous validation. The machine learning models yielded high cross-validation results. Overall we present here globally scaled technology for the land sector based on high resolution remote sensing imagery, meteorological variables, and direct carbon flux measurements of eddy covariance flux stations.

Keywords: net ecosystem exchange; eddy-covariance; regression; upscaling; data augmentation; feature selection

1. Introduction

Nature-based solutions (NBS) are increasingly recognized as an essential tool in combating global warming as they remove Greenhouse Gases (GHG) from the atmosphere [1], reduce GHG emissions through sustainable land use practices, and store carbon in the soil and biomass [2]. While NBS are becoming quite advanced and continuously going through significant developments, the methods for accurately quantifying their resulting benefits are still lagging [3]. A vast majority of the existing projects use biometric soil or tree survey data to estimate carbon stock changes [4]. These methods provide solid estimates of the sequestered carbon but come with multiple important limitations. They typically describe either soil layer or partial canopy layer (e.g., above ground), with time resolution limited to multiple months and years and space resolution limited to a particular plot where such measurements have been performed. As a result of such limitations, these methods may not be best suited for large-scale ongoing NBS quantification [2]. Moreover, the measurements required by these methods for a meaningful time-resolved, large-scale assessment are expensive and unaffordable for most users.

A pure modeling approach employing either Earth System Models or Terrestrial Biosphere Models has been used as an alternative to direct measurements at a large scale

[5,6], but the spatial resolution of these models is too coarse, while the uncertainty remains high due to their inability to respond timely to actual changes in the ecosystem. Likewise, satellites that measure CO₂ concentrations and are used in inverse modeling to map GHG emission and sequestration [7,8] have coarse spatial resolution and hence can hardly be used for the NBS assessment at the farm, or forest stand level. Optical, radar and lidar satellites provide a scalable solution for mapping soil carbon [9] and carbon fluxes at a much finer spatial resolution [10,11], and do have the potential to help large-scale space- and time-resolved NBS assessments if anchored in the accurate ground measurements. The current remote sensing-based studies typically use allometric/biometric data for machine learning (ML) model training [2]. The amount of ground data required for capturing the landscape heterogeneity and temporal variation is challenging to sustain and scale. The models thus inherit the disadvantages of the biometric measurements.

There is however another potential solution that combines many advantages of the above methods while overcoming some of their limitations, and can result in a high spatial resolution, scalability to the regional level, and practical affordability. This can be achieved by blending the direct highly time-resolved small-scale observations from the numerous eddy covariance flux towers [2] and the less direct but large-scale remote sensing methods. Considerable efforts have been made in this area lately to improve the quantification of the spatio-temporal carbon dioxide (CO₂) patterns, employing multiple different approaches vs. existing estimates to help resolve the majority of issues with current carbon accounting methods in the land sector.

For example, [12] evaluate how seven light-use-efficiency (LUE) models captured the spatio-temporal gross primary productivity (GPP) variations from the LaThuile FLUXNET dataset. This study used data from 157 eddy covariance (EC) towers located in the six major terrestrial biomes (evergreen broadleaf forest (EBF), deciduous broadleaf forest (DBF), mixed forest (MF), evergreen needle leaf forest (ENF), shrubland (SHR) and grassland (GRA). None of the seven models matched well with estimated GPP based on overall EC measurements. The models explained 41–57% GPP variations in the study sites.

[13] evaluated four LUE models for GPP estimation and compared them with data obtained by EC in 51 sites with diverse vegetation types. The average daily R^2 for all models and all sites was about 66%. [14] used three different models to estimate GPP in a local deciduous forest in the US in combination with flux measurements and remote sensing data. Comparisons were made yearly, but on average, the correlation coefficient was much better, at 0.91. [15] used two enhanced vegetation index (EVI) based models to estimate the GPP in 15 sites across the USA. The model's results vary according to the vegetation; in some cases, coefficients of determination can vary between 0.27 up to 0.94 for specific sites. Using remote sensing data, disturbance information, and a support vector regression model, [16] upscaled the CO₂ fluxes from a network of 21 EC towers to estimate the regional-scale Alaskan CO₂ budget from 2000 to 2011 and compared the results with an inverse model. The predicted values were consistent with the field observations (R^2 = 0.79 for NEE) and the magnitude of the inverse model but were inconsistent with the interannual variations between the two approaches needing more validation. [17] used three remote sensing data-driven models to evaluate and understand GPP estimations in 119 EC sites in the northern hemisphere. Considering all the study sites, the mean R^2 was 0.63. However, the values can vary significantly depending on the model used and the vegetation, and in some cases, they can reach 0.78. Usually, the highest values were found for deciduous broadleaf and mixed forest. [18] mapped the organic carbon stored in Italy using ground observations and ancillary remote sensing data. Compared with all observations, the geographically weighted regression using MODIS Normalized Difference Vegetation Index (NDVI) obtained an R^2 of 0.45. [19] have done comprehensive work generalizing and mapping NEE, GPP, and respiration data for all types of classification as a collaborative FLUXCOM initiative. The overall coefficients of determination for GPP and respiration were more than 0.71 and 0.64, respectively, while about 0.48 for NEE.

As seen in the examples above, despite the significant advances in the use of remote sensing for CO₂ flux estimations, the results point to mixed success and lack of consistency, so there is room for major improvements. The key areas of such improvements include increasing spatial resolution, capturing temporal variability, improving the data quality (primarily by reducing uncertainties and increasing consistency), and developing better algorithms to enable the simultaneous use of multiple relevant parameters and different spatial and temporal scales. In order to achieve such improvements, we propose to rely on the significant advances which have been made in recent years in the fields of machine learning, artificial intelligence, and neural networks to extract patterns and insights from remote sensing data, processing large volumes of remotely sensed data with minor human interference [20]. Here we present the successful use of one of such approaches for globally scaled remote sensing-based carbon accounting models anchored by the direct carbon flux measurements using eddy covariance flux stations. We use input data on the net ecosystem exchange (NEE, [21–24]) as targets for the training process, and then combine the remote sensing and meteorological variables [19] as predictors to reproduce variations of NEE on a daily and monthly scales, and to demonstrate the increased predictive power of the proposed approach over more traditional methods.

2. Data and Methods

2.1. Data

2.1.1. Eddy covariance data

The eddy covariance method is a standard method to directly measure fluxes between ecosystems and the atmosphere by computing the covariance between the vertical wind velocity and the entity of interest [21,25,26]. Such measurements are made routinely on all continents [27–29], with many sites linked across a global network called FLUXNET, dispersed across the world’s climate zones and representative biomes [29–33]. FLUXNET teams standardize the data format, perform uniform quality checks, and use highly vetted gap-filling and flux partitioning, with the newest product version being the FLUXNET2015 dataset, which includes over 212 sites [32] (see Figure 1).

We used data from 171 Tier 1 sites with a data policy consistent with the Attribution 4.0 International (CC-BY-4.0) data usage license. Daily net ecosystem exchange (NEE) was filtered using a variable friction velocity threshold for each year and quality flags indicating the measured and good quality percentage (data higher than 0.8 were used, indicating percentage of measured and good quality gap-fill data). The following land cover classes based on the International Geosphere-Biosphere Programme (IGBP) classification were used: cropland (CRO); grassland (GRA), which includes grasslands, closed shrublands (CSH), and open shrublands (OSH) [34]; evergreen needleleaf forest (ENF); deciduous broadleaf forest (DBF); evergreen broadleaf forest (EBF); mixed forest (MF); wetland (WET); savannas and woody savannas (SAV).



Figure 1. Locations of all Fluxnet sites.

2.1.2. Remote sensing data

We used Landsat surface reflectance (SR) and land surface temperature (LST) as well as a number of indices derived from the former as the main predicting features for our machine learning model. Landsat 5, 7, and 8 (TM, ETM+, and OLI respectively) cover our study period and have a 30-m spatial resolution. USGS provides open access to their Landsat Collection 2 Level-2 Science Products (L2SP) through Amazon s3 cloud service. L2SP data are geometrically and atmospherically corrected using the LEDAPS algorithm for TM and ETM+ and LaSRC for OLI imagery, and thus prepared for scientific analyses. USGS Landsat was accessed on February 1 2022 from <https://registry.opendata.aws/usgs-landsat>.

We filtered out imagery with cloud cover above 75 percent of a scene. To ensure consistently high data quality, we masked out all possible artifacts as marked in the quality assessment bands provided with the product. We also left out high to low-confidence clouds and cloud shadows detected by the CFMask algorithm [35].

We utilized SR from three visible, near-infrared (NIR), and short-wave infrared (SWIR 1 and SWIR 2) spectral bands (bands 1-5 and 7 for TM and ETM+; and 2-7 for OLI). These bands are well-suited for assessing vegetation and soil state and, consequently, together with LST, can serve as accurate predictors of NEE.

In addition, We selected two spectral indices and two tasseled-cap components [36] to map proxy land surface parameters to predict NEE better. Soil-Adjusted Vegetation Index (SAVI) and Tasseled-Cap Greenness (TCG) reflect vegetation intensity and chlorophyll content. In contrast, Normalized-Difference Moisture Index (NDMI), Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and Tasseled-Cap Wetness (TCW) reflect moisture content in vegetation in our areas of interest [37,38].

$$TCG = -0.1603 * b1 + 0.2819 * b2 - 0.4934 * RED + 0.7940 * NIR - 0.0002 * SWIR - 0.1446 * b7 \quad (1)$$

$$TCW = 0.0315 * b1 + 0.2021 * b2 + 0.3102 * RED + 0.1594 * NIR - 0.6806 * SWIR - 0.6109 * b7 \quad (2)$$

$$NDMI = \frac{NIR - SWIR}{NIR + SWIR} \quad (3)$$

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad (4)$$

$$SAVI = \frac{NIR - RED}{NIR + RED + 0.5} * 1.5 \quad (5)$$

$$EVI = 2.5 * \frac{NIR - RED}{NIR + 6 * RED - 7.5 * b2 + 1} \quad (6)$$

where b1, b2, b7 – corresponding original bands from Landsat; RED – band 3; NIR – band 4; SWIR – band 5.

2.1.3. Meteorological data

In addition to the remotely sensed data characterizing the vegetation cover, the performance of NEE is influenced by weather parameters such as temperature, dew-point temperature, solar radiation, and precipitation. The global dataset AgERA5 was chosen as the source of this data [39].

This service is based initially on hourly ECMWF ERA5 data at the surface level, which is available at 30-km spatial resolution (about 0.28125°). Data were processed to daily temporal resolution by summing or averaging depending on a particular parameter and corrected to a more refined surface description at a 0.1° spatial resolution. Aggregated data follow a local time zone definition and include several major agronomic parameters. The correction to the 0.1° grid was performed by applying grid and variable-specific regression equations to an ERA5 dataset interpolated at a 0.1° grid. The equations were trained on operational ECMWF's high-resolution atmospheric model (HRES) data at a 0.1° resolution.

The following set of parameters was chosen for our purposes: 24 hours mean of 2-meter air temperature ('Temperature-Air'), 24 hours mean dew-point temperature ('Dew-

Point'), precipitation flux (sum of total precipitation per day, 'Precipitation-Flux'), surface solar radiation downwards (sum of solar radiation per day, 'Solar'). A description of the additional meteorological features generated and used in this work is provided in Section 2.2.2.

2.2. Simulation experiment

The eddy covariance method measures fluxes integrated over an upwind area of a specific size, known as footprint [40,41]. The size of this area depends significantly on wind speed and direction. The original NEE measurements from stations have a half-hour temporal resolution, which means that the station is sensitive to the area around at a distance that air mass has passed upstream during a specific half-hour. The linear scale of such an area may range from hundreds of meters to several kilometers.

Thus, in the case of 30-m resolution remote sensing data, we cannot directly associate the surface reflectance value in a band's pixel (or calculated index) with the station inside and NEE estimation on this station for the specific date. In order to do so, we need to calculate the relevant weighted footprint [42,43] of the impact area around the station for the corresponding time scale (daily in our case, see Figure 2). To calculate this footprint, we used a meteorological model which can reproduce transport processes on a local scale depending on weather conditions on a given day. For this purpose, we used the Lagrangian particle dispersion model (LPDM) FLEXPART. We weighted six Landsat-derived spectral bands and indices and integrated them by a footprint probability density function estimated using the FLEXPART model in backward mode. There is no need to do such a calculation for meteorological features, as their spatial resolution is coarser (about 9-10 km).

2.2.1. Footprint calculation with LPDM

We used the Lagrangian particle dispersion models (LPDM) FLEXPART [44,45] to estimate the footprint of Fluxnet stations. In Lagrangian models, individual particle trajectories (representing air parcel transport) are calculated following a predetermined atmospheric wind field and other meteorological variables. LPDMs can run forward (source to a receptor) or backward (receptor to source) modes.

We used the FLEXPART version, which utilizes GFS meteorological data on a regular latitude-longitude horizontal grid with 0.25-1.0 degree spatial resolution on hybrid sigma-pressure vertical levels and temporal resolution of 3-6 hours. The model output grid has a spatial resolution on a latitude-longitude grid with 0.002 degrees (about 150-200 meters), and the smoother result, the model synchronization time step, was 15 seconds.

We need to run the model in backward mode to get the weighted footprint for a specific station, setting station parameters such as location coordinates and above-ground height as receptor position. The duration of trajectories was selected corresponding to the original EC system time scale, which is 30 minutes. This model setup was run six times a day (i.e., every 4 hours) for every station for every date where satellite images were available to obtain a daily footprint.

One of the regular output parameters of the FLEXPART model in backward mode is residence time, which is measured in seconds. That gives an understanding of how long one particle, or group of particles, has "lived" in a particular grid cell, which allows it to evaluate the receptor's sensitivity to the surface grid cells. Thus, the resulting residence time map can be interpreted as a weighting function (see Figure 3 right panels), which we need to rely on when processing the remote sensing data..

2.2.2. Feature generation and data augmentation

We selected the SR and the LST values from the Landsat bands, indices calculated from these bands, and the set of meteorological parameters taken from the reanalysis AgERA5 as predictors of a regression model to predict daily NEE, which is a product from Fluxnet stations (see Section 2.1.1).

In this work, a set of additional features has been created to increase the flexibility of the regression model and expand the interactions between the base predictive features and the target variable. First, all possible ratios were created between 6 Landsat bands and LST (e.g., b1/b2, b1/RED, b1/NIR, and vice versa b2/b1, RED/b1, NIR/b1). As a result, 42 additional features were added to the data set at this stage. Second, from meteorological variables such as solar radiation, 2-meter air temperature, and precipitation flux, additional features were also created. For solar radiation, cumulative and mean radiation over five different periods were calculated: four days, one week, two weeks, three weeks, and four weeks (10 features as a result), which may show delayed phenological effects on sub-seasonal and seasonal scales. The same set of parameters was prepared for 2-m air temperature. For precipitation flux, cumulative precipitation for the last one, two, three, four, and five days was calculated (5 features). Thus, 25 more features were added to the base predictive features. In total, 87 features were prepared for further training.

Another issue in the dataset collection process is the availability of clear-sky observations from the Landsat constellation. After merging the dates with available satellite images and good quality NEE observations, fewer training points remain, mainly due to the images' cloud cover and other artifacts. We decided to apply training sample augmentation techniques to solve this problem.

The procedure for increasing the training sample is based on the assumption that the surface reflectance on the day of satellite measurements, and the neighboring days, will be similar given similar weather conditions. Weather variables, i.e., precipitation, air temperature, short-wave, and incoming long-wave radiation, were determined using data from local meteorological measurements at each eddy-covariance station. From these four parameters, a weather profile was generated for the whole time series of daily NEE estimations, which was then compared with a similar profile on the day of satellite observations. If the R^2 value was greater than 0.9, it was assumed that the day had similar weather conditions, and images of the original day could be applied to such a day. As a result, we may use values for all bands and indices at augmented days from neighboring days but apply the weighted footprint function for the specific day. The time considered in this study was a week before the date of Landsat observation. As a result, the size of the training sample was increased by about a factor of 3.

2.2.3. Building regression model

Figure 2 demonstrates the data flow for the dataset creation process. The formulas for calculating weighted parameter values and a regression model are as follows:

$$F_{weighted}^j = \sum_{i=1}^N w_i * F_i^j \quad (7)$$

$$NEE = H(F_{weighted}^1, F_{weighted}^2, \dots, F_{weighted}^M) \quad (8)$$

where w_i – weight value for the i -th grid cell (see Figure 3); F_i^j – value of j -th predictive feature for i -th grid cell; $F_{weighted}^j$ – weighted sum for j -th feature; NEE – eddy covariance data; N – number of grid cells; M – number of features; H – non-linear operator.

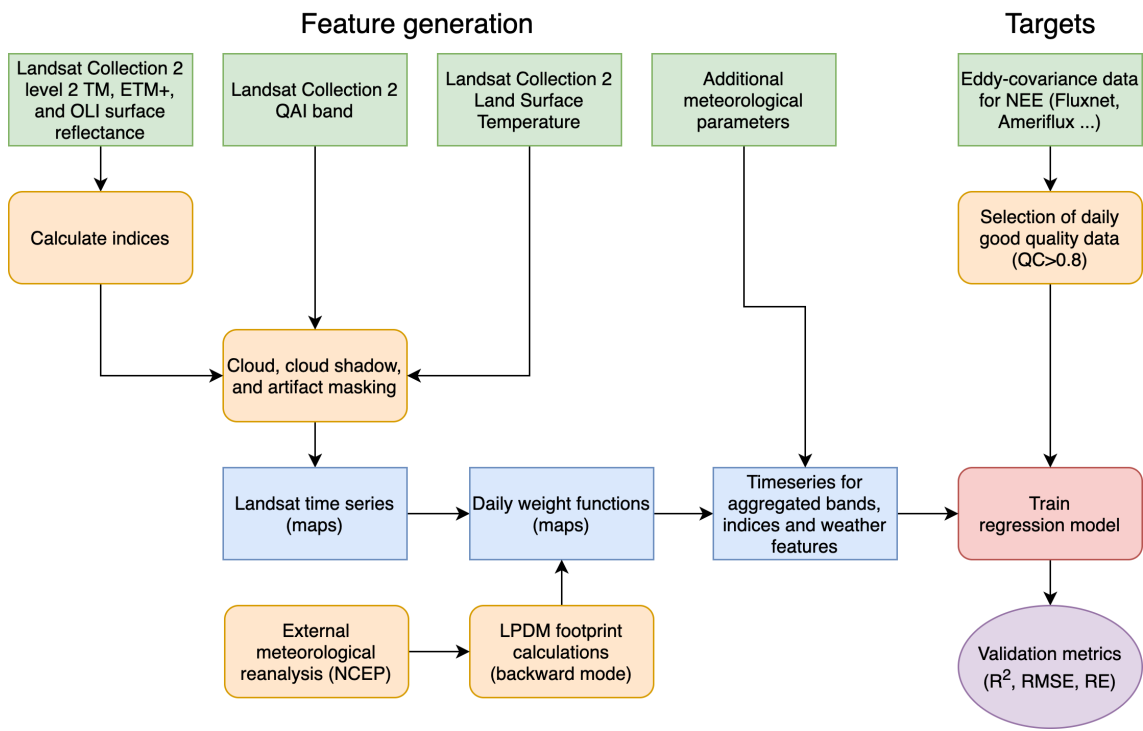


Figure 2. General flowchart of regression model training process.

Figure 3 shows an example of original features calculated from Landsat data and results of LPDM calculations of footprint for the up-scaling process. The weighted sum for each parameter may be interpreted as the dot product of the feature and footprint map. To train the model, we match the resulting input dataset (Table 1) with eddy covariance data (Table 2). An example of targets (daily NEE) is provided in Figure 4.

253
254
255
256
257

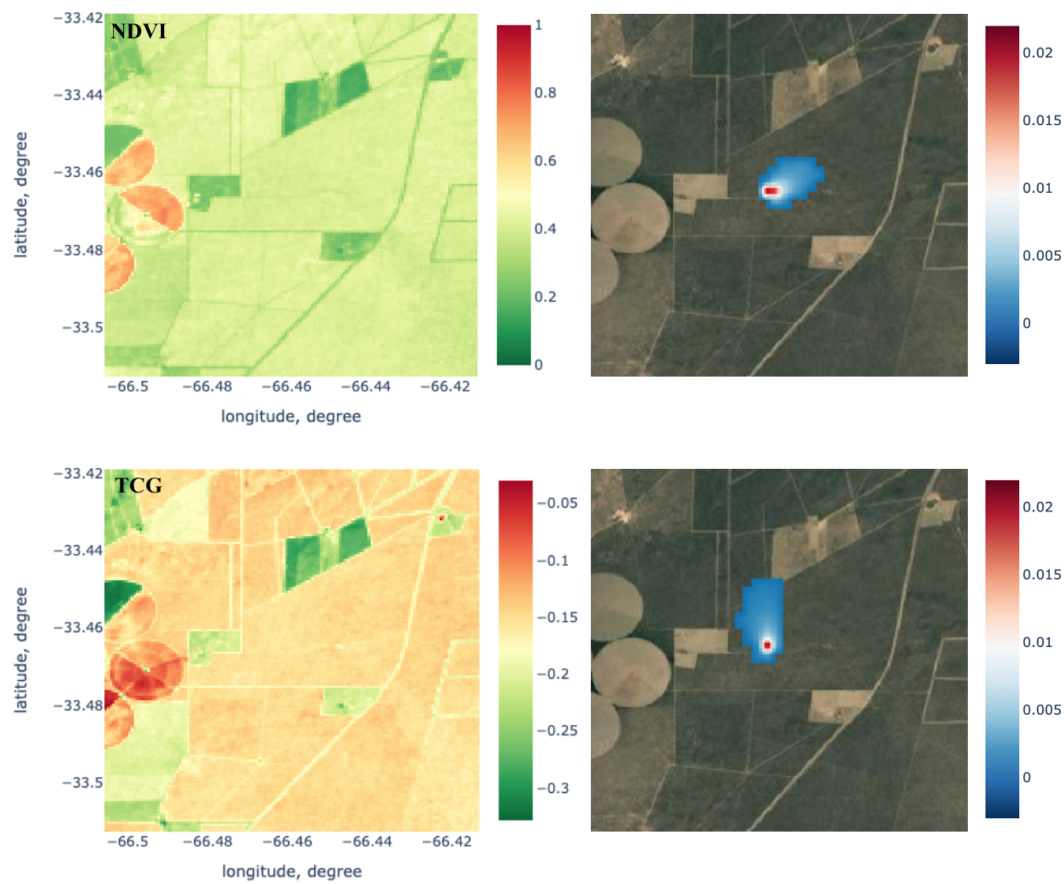


Figure 3. Left panels are maps of calculated indices for the same date; right panels are maps of daily weight functions for different dates.

Table 1. Example of the aggregated dataset for each feature for each date. Note that the values are scaled by 10000.

Date	b1	NDMI	SAVI	...
2009-12-27	344.9	327.6	2314.2	...
2009-12-28	346.8	345.2	2335.3	...
2009-12-29	338.3	315.5	2276.5	...
2009-12-30	342.1	317.5	2289.9	...
2009-12-31	355.2	355.5	2392.7	...
2010-01-01	351.3	356.4	2354.5	...
2010-01-02	349.3	325.1	2346.3	...
2010-01-03	351.4	361.2	2373.4	...
2010-01-04	355.4	363.2	2382.4	...
2010-01-05	352.2	348.9	2367.3	...
...	...	...	...	...

Table 2. Example of “target” NEE from eddy covariance stations for each date.

Date	NEE, gC/m2/day
2009-12-27	-5.77
2009-12-28	-7.43
2009-12-29	-10.24
2009-12-30	-5.49
2009-12-31	-7.83
2010-01-01	-5.47
2010-01-02	-5.24
2010-01-03	-3.87
2010-01-04	-5.52
2010-01-05	-3.45
...	...

We selected an algorithm based on the kernel method Kernel Ridge Regression (KRR) [46,47]. This method is similar to the more widely known class of support vector machine (SVM) methods [48]. However, it uses a different loss function, which allows for creating non-linear dependencies between the features according to the selected kernel, and then searching for the optimal coefficients of these dependencies with control of the regularization process. The result of solving such a problem is a hyperplane in the multidimensional feature space that describes the dependencies of the target values on the selected features. This method works well with small data sets, which is the case in this study and is also faster than the original SVR.

KRR has three main hyperparameters: α – the regularization parameter of the model (the higher the value, the stronger the regularization); kernel – kernel interaction between the features included in the model (see eq. 9-10); γ – internal kernel parameter (see eq. 9-10).

$$k(x,y) = e^{-\gamma\|x-y\|_1} \tag{9}$$

$$k(x,y) = e^{-\gamma\|x-y\|_2} \tag{10}$$

where e – exponent; x, y – input vectors; $\|x - y\|_1$ - Manhattan distance between the input vectors; $\|x - y\|_2$ – Euclidean distance between the input vectors.

Another critical aspect of preparation is the feature selection process. As mentioned in 2.2.2, in addition to the primary remote sensing and weather parameters, 87 additional features in total have been prepared. Also, we use products between all meteorological and remote sensing features that lead to a few thousand of features.

For each IGBP class, a small set of features has to be selected to allow optimal prediction of NEE at the stations. The exact solution for finding the optimal combination of features can be obtained by a complete enumeration of all possible variants. The complexity of this approach is 2^n where n is a number of features, so the approach becomes inapplicable for a large number of features. The utilities in the package machine learning extensions (Mlxtend) [49] were used to select the features.

The selection process was divided into two stages. The first stage estimates mutual information for each feature in the list. A standard method from the sci-kit-learn library was used [50], which allows us to reduce the number of features to select. The second step is a sequential selection of features in the forward mode. That means the model starts to learn from one feature and sequentially adds features to improve the quality of the model. As a result of this selection, the feature sets for each IGBP class have been generated, as shown in Table 3.

Model training and hyperparameters tuning processes were based on N -fold cross-validation. In our case, N equals 3-5-7 depending on the size of the available dataset per each IGBP class. The original dataset was divided into N nearly same-sized folds with a unique set of sites inside. Then, the model was trained on $N - 1$ folds and validated on

the remaining fold after the procedure was repeated N times with different combinations of folds. Each fold contains a unique set of stations without overlapping with other folds [51], which ensures spatial scalability. Thus, N times independent validation is performed on the sets of stations that were not involved in the training of the model, which ensures spatial scalability of the model predictions. Similarly to [19], hyperparameters for each N fold were tuned separately for a fixed set of features.

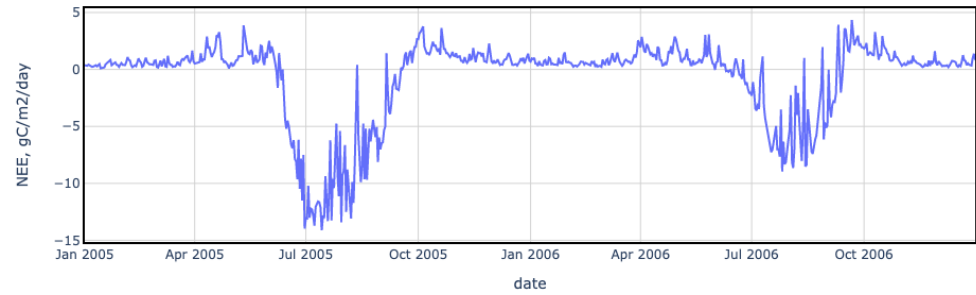


Figure 4. Daily NEE values in gC/m²/day for one station (US-Ne3) for 2005-2006.

Table 3. Feature sets for predicting NEE for all IGBP classes.

List of selected variables
$LST/b7 * Mean_Solar_1week, LST/b7 * Solar, LST, Cum_Solar_2week, Cum_Solar_4days,$ $Cum_Temperature_1week, Cum_Temperature_4days, Cum_Temperature_over_2_1week,$ $Cum_Temperature_over_2_3week, Cum_Temperature_over_3_3week,$ $Cum_Temperature_over_4_3week, Cum_Temperature_over_5_2week,$ $Cum_Temperature_over_5_3week, Cum_precipitation_2days, Cum_precipitation_3days,$ $Cum_precipitation_5days, Delta_T, DewPoint, EVI * DewPoint, Mean_Solar_2week,$ $Mean_Solar_4days, Mean_Solar_4week, Mean_Temperature_1week, NDVI,$ $NDVI * Cum_Solar_3week, NDVI * Mean_Solar_3week, NDVI * Mean_Solar_4week,$ $NDVI * Solar, Solar, TCG, TCG * Cum_Solar_3week, TCG * Cum_Solar_4days,$ $TCG * Dew - Point, TCG * Mean_Temperature_2week, TCG * Solar, TemperatureAir, b1 * SWIR,$ $b1/b2, b1/RED, b1/NIR, b1/b7, b2 * RED, b2/RED, b2/NIR, RED/b1, RED/b2, RED/NIR,$ $RED/SWIR, RED/b7 * Cum_Solar_2week, NIR * SWIR, NIR/LST, NIR/LST * Solar, NIR/b2,$ $NIR/SWIR, NIR/b7, NIR/b7 * Mean_Solar_2week, NIR/b7 * Solar, NIR * Cum_Solar_3week,$ $NIR * Solar, SWIR/b2, SWIR/RED, SWIR/NIR,$ $SWIR/NIR * Cum_Temperature_over_5_4week, b7/b1, b7/b2, b7/NIR, b7/SWIR$

2.2.4. Metrics for accuracy assessment

Several standard metrics were used to assess the level of accuracy for the trained NEE estimation model: mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), and relative error (RE).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i^t - y_i)^2} \quad (11)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i^t - y_i)^2}{\sum_{i=1}^N (y_i^t - \bar{y})^2} \quad (12)$$

$$RE = \frac{\frac{1}{N} \sum_{i=1}^N |y_i^t - y_i|}{\frac{1}{N} \sum_{i=1}^N |y_i^t - \bar{y}|}$$

(13)

where N is the number of observation points to estimate; y_i^t and y_i are the i -th true and modeled NEE value respectively; $\bar{y}$ is the mean NEE value.

3. Results

As a result of N -fold cross-validation for each land cover (Section 2.2.3), we received an ensemble of N different models trained on different independent data sets with the same inputs. Thus, we can use the mean value over the ensemble as a prediction.

First, selected metrics were applied to estimate the model’s accuracy by reproducing the original daily NEE values and then used to estimate the accuracy of the predicted NEE behavior. The results of the training and cross-validation experiments based on independent stations are presented in Table 4. One can see that the trained models can adequately reproduce the behavior of NEE at stations with different land cover (overall mean across land cover NEE , $R^2 > 0.6$). The lowest predicted values are for EBFs because such forests are mostly located in the tropics, where there are few Fluxnet stations, and there could be problems with cloud conditions during some seasons. The largest variation in error (standard deviation) is observed for CRO, possibly due to the difference in crop types and significant local anthropogenic influence on specific fields. An increase or decrease in NEE can be explained by the state of the biomass and the influence of agricultural operations. Comparisons of observed values and predicted values are shown in Figures 5, 6.

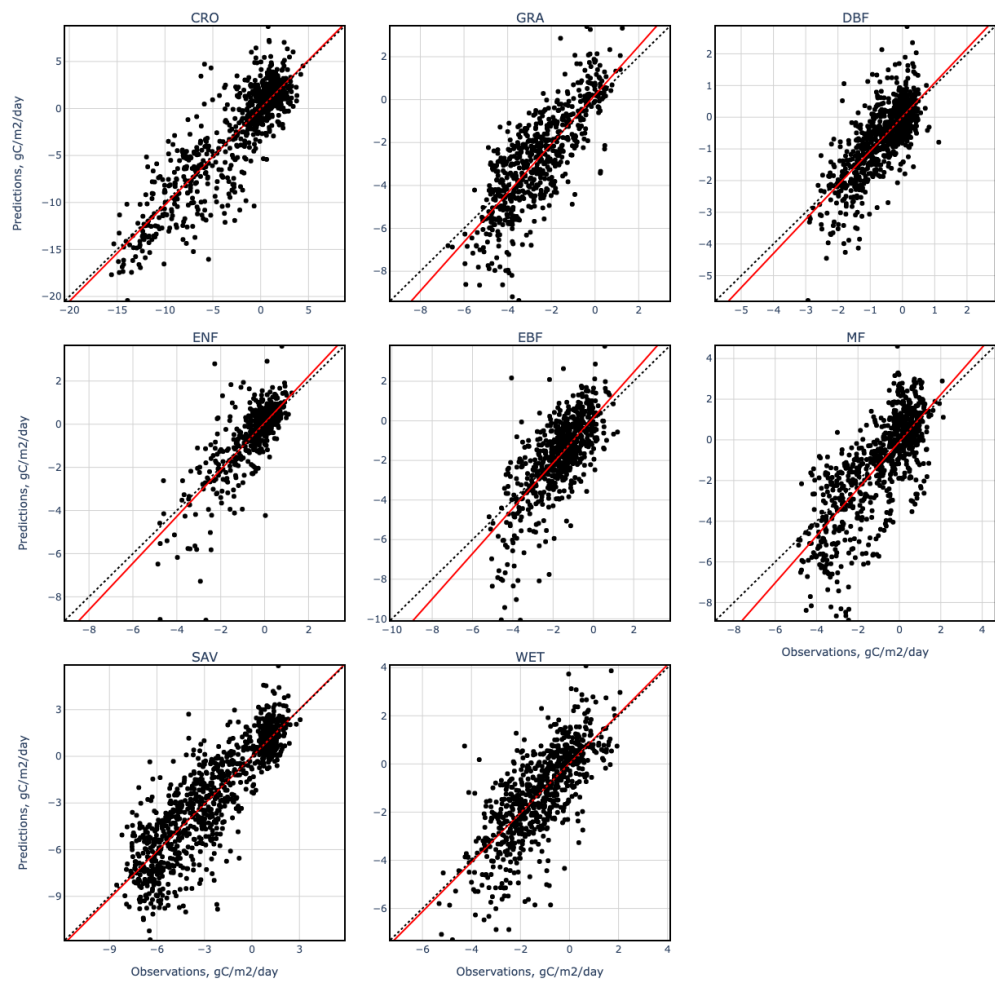


Figure 5. Scatterplots of observed NEE vs. predicted NEE based on remote sensing and meteorological features models per each land cover: cropland (CRO); grassland (GRA); evergreen needleleaf forest (ENF); deciduous broadleaf forest (DBF); evergreen broadleaf forest (EBF); mixed forest (MF); wetland (WET); savanna (SAV).

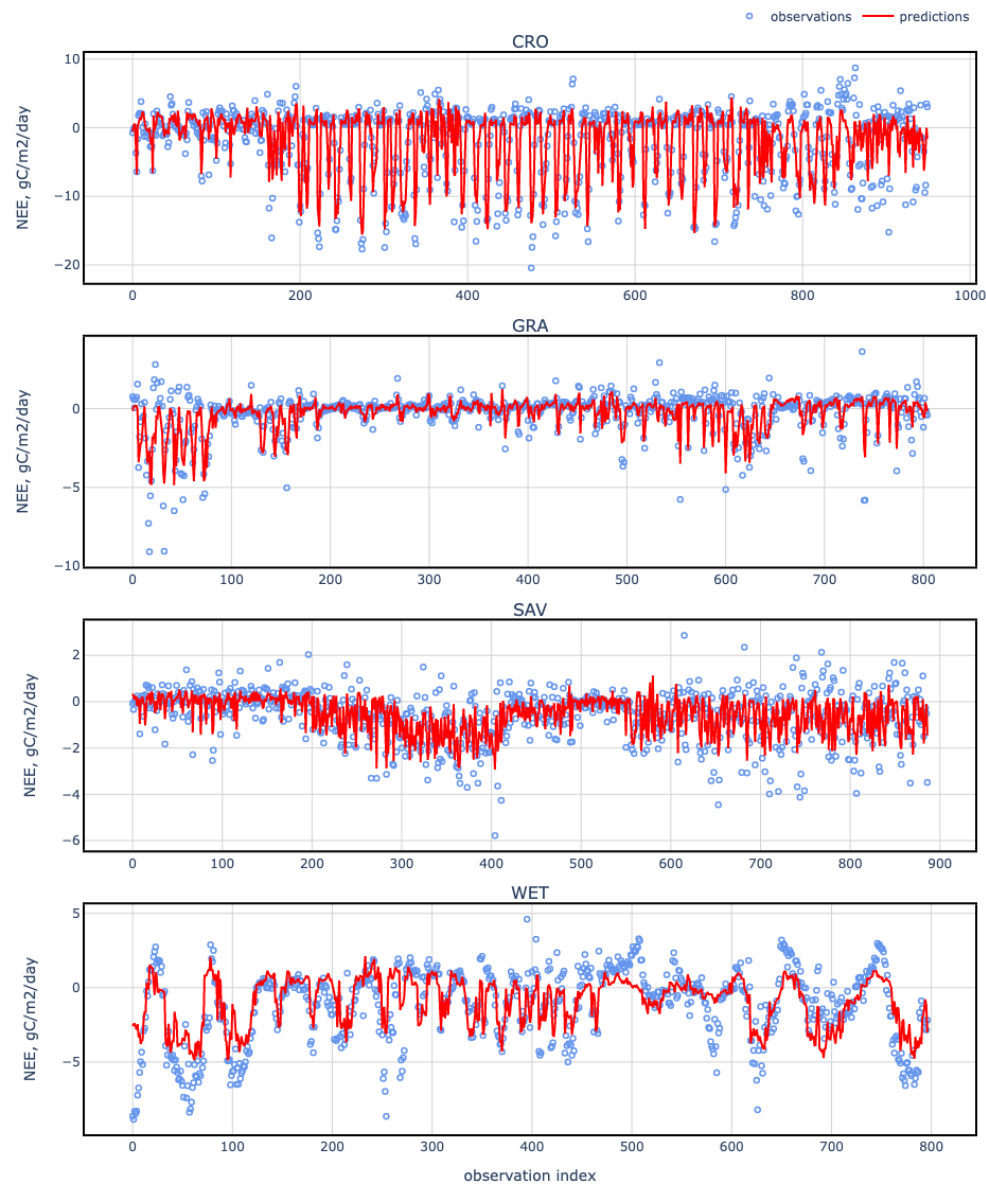


Figure 6. Time series of observed NEE and the predicted NEE based on remote sensing and meteorological features models per each land cover: cropland (CRO); grassland (GRA); evergreen needleleaf forest (ENF); deciduous broadleaf forest (DBF); evergreen broadleaf forest (EBF); mixed forest (MF); wetland (WET); savanna (SAV).

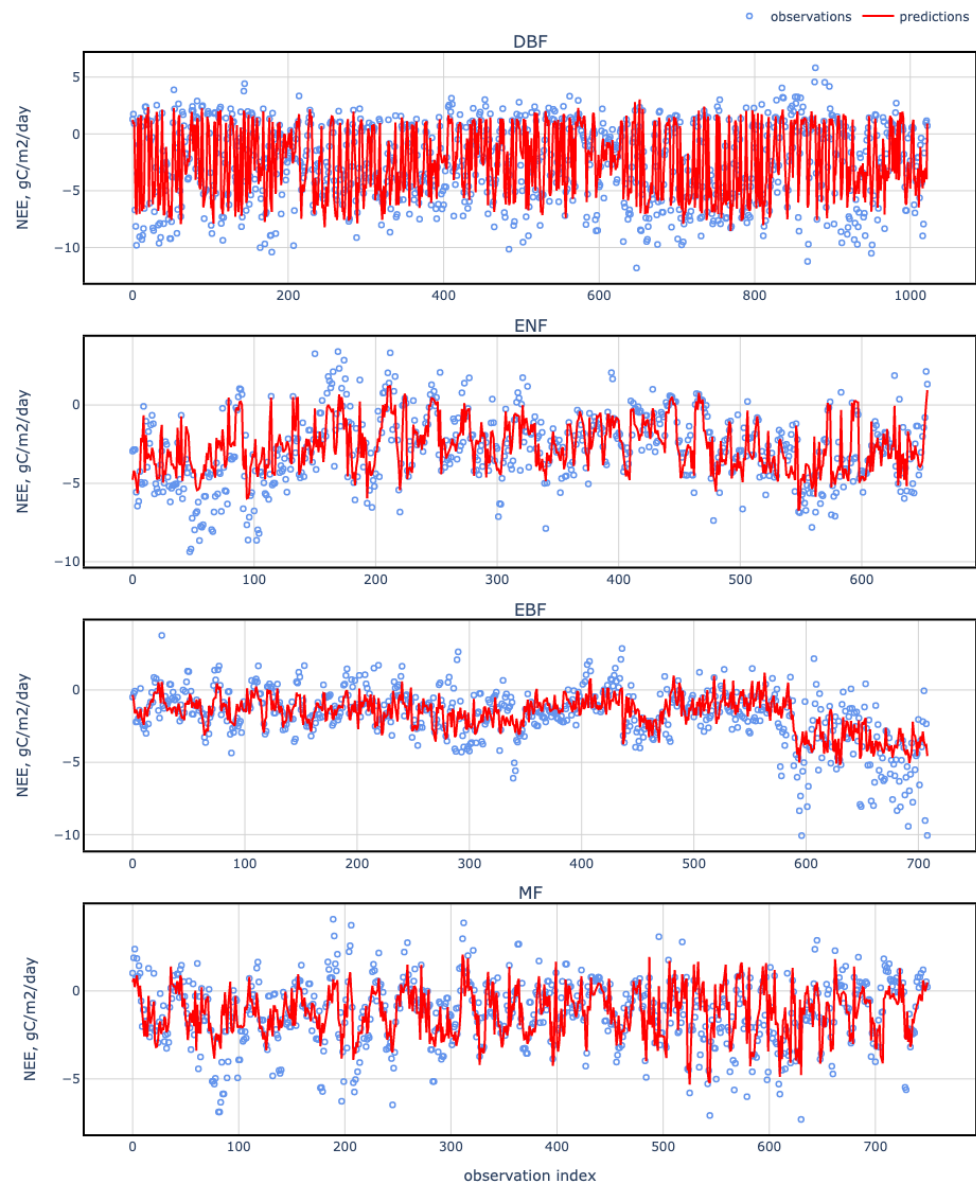


Figure 6. (continued).

Table 4. Cross-validation metrics (and standard deviations for each metric) for daily NEE models per land cover class.

IGBP	R ²	RMSE, gC/m2/day	RE
CRO	0.73 (0.17)	2.32 (0.67)	0.44 (0.19)
GRA	0.61 (0.04)	0.70 (0.35)	0.60 (0.05)
DBF	0.76 (0.05)	1.64 (0.16)	0.42 (0.04)
ENF	0.62 (0.05)	1.30 (0.22)	0.60 (0.06)
MF	0.55 (0.02)	1.25 (0.18)	0.64 (0.01)
EBF	0.42 (0.06)	1.28 (0.32)	0.73 (0.05)
WET	0.53 (0.06)	1.55 (0.19)	0.64 (0.08)
SAV	0.56 (0.03)	0.68 (0.14)	0.63 (0.02)

To compare monthly *NEE*, additional daily footprint calculations similar to Section 2.2.1 were carried out. In this case, the calculation period of one week before and one week after the available satellite images were selected. Usually, periods of 8 days are used [19,34], but in our case, we decided to use a similar averaging backward and forwards. Further, remote sensing data from the original image was used for the additional dates, but with meteorological variables and weighting functions corresponding to the actual date. Monthly model *NEE* data were then calculated for those months where it was possible to predict more or equal to 2/3 days of the month. The resulting monthly *NEE* was compared with monthly *NEE* values from stations for which the quality flag was greater than or equal to 0.8. The results of the comparison are presented in Table 5.

Table 5. Cross-validation metrics for different land covers monthly *NEE* values.

IGBP	R ²	RMSE, gC/m2/day
CRO	0.83	1.94
GRA	0.60	0.54
DBF	0.84	1.23
ENF	0.78	0.89
MF	0.70	0.75
EBF	0.70	0.80
WET	0.66	1.25
SAV	0.69	0.49

4. Discussion

There has been a drastic rise in NBS development in recent years [3]. Although the methods for evaluating NBS are advancing, there are still certain gaps that need to be addressed [2]. Previous studies found that most of the current methods for quantifying NBS benefits use biometric soil or tree survey data to estimate carbon stock changes. These methods do not enable seamless large-scale assessments, while associated regular measurements make the methods financially unsustainable for most players in the land sector. Modeling approaches based on Earth System Models or Terrestrial Biosphere Models, as well as inverse modeling based on CO₂ concentrations data, have spatial resolution too coarse to be used locally. The current remote sensing-based approaches mainly rely on allometric data as ground truth and therefore involve challenges with capturing landscape heterogeneity and temporal variation that negatively impact scalability. As the result of the analysis from the recent studies, it is fair to say that the approaches based on observations from eddy covariance towers and remote sensing are the best fit to enable affordable, high spatial resolution, globally scalable CO₂ estimates. A dozen papers representing a vast amount of scientific efforts targeted on these approaches were analyzed, and the significant challenges were identified. Here we presented an improved approach to quantifying carbon fluxes by blending eddy covariance data with moderate resolution satellite data and weather data. Our method (1) is scalable, due to the validation design based on a separate set of eddy covariance stations, (2) features a high spatial and temporal resolution of the Landsat data and set of meteorological variables, (3) and delivers robust and accurate predictions due to improved data quality control, advanced machine learning techniques, and rigorous validation.

Further in the Discussion, we suggest the key reasons for these improvements and also list our opinion on the current limitations of the proposed approach. Our methodology employed eddy covariance data from the Fluxnet network. Since the network has a standardized methodology for data acquisition and processing, we collected a robust and consistent training dataset, typically unavailable for more traditional carbon accounting methods based on biomass and soil sampling. The availability of a large amount of consistent and uniform data allowed for setting a fairly conservative threshold of 80% for the *NEE* data coverage and resulted in a substantially reduced noise. On the remote sensing side of the approach, the Landsat archive provided a consistent and robust set of data with

great temporal depth and spatial resolution. The FLEXPART model was instrumental in mapping the area that contributed to the NEE values captured by flux stations, helping to match these with remote sensing products better to account for landscape heterogeneity. Utilization of the KRR and its parameterization tailored to the goals of this study, and strict feature selection, allowed achieving high accuracy of the predictions.

Still, the metrics of our regression estimators were not uniform and varied across the land cover classes. The R^2 for the DBF was the highest likely due to clear seasonality patterns and relatively low heterogeneity within the class. The same metric for the EBF class was the lowest, likely due to other carbon exchange process drivers and lower data availability due to high cloud cover in tropical regions.

Current results are presented for each land cover class globally, providing visibility on the unified, scalable calculation approaches and the global accuracy baselines for various land covers. Overall, the achieved accuracy was in line with similar studies but outperformed most of those targeting global geographical coverage (overall land cover cross-validation $R^2 = 0.73$, $RMSE = 1.53$ gC/m²/day). Daily data accuracy has the R^2 range between 0.42 for EBF and 0.76 for DBF (monthly data accuracy is between 0.6 for GRA and 0.84 for DBF). The models have better accuracy in those parts of the world where the number of stations is higher and, conversely, slightly worse accuracy in areas of lower station density.

To resolve or minimize these deficiencies and improve accuracy for specific cases or regions, the current system allows additional training (not a local re-training) in the presence of the additional local flux observations, as suggested by [2]. Another way to improve the confidence of model data is to add other multispectral, hyperspectral, and radar satellite products, e.g., Sentinel-1, 2, or EnMap. It is expected that these new data streams can improve the accuracy of the daily data slightly but can contribute significantly to the monthly and annual data improvement.

Finally, since in this study we have trained N independent models for each land cover (where $N = 3, 5, 7$), in the future, it may be better to use ensemble predictions to allow additional characteristics such as mean, median, and standard deviation. These and other activities are currently in development but not advanced enough to be included in this manuscript.

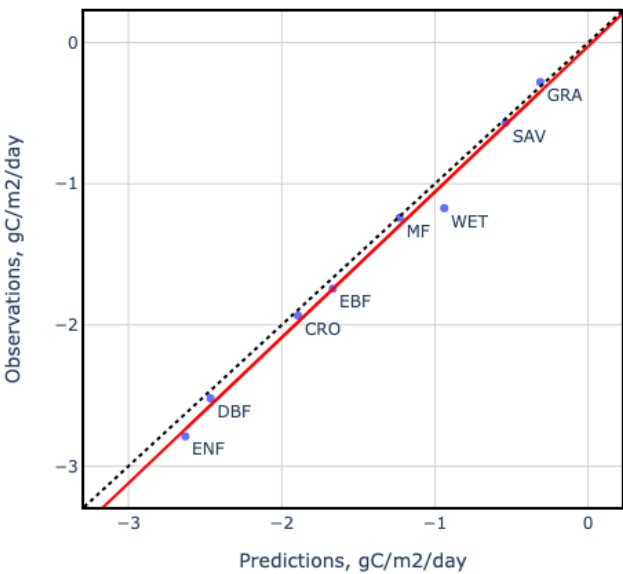


Figure 7. Comparison of observed and estimated mean NEE for different land covers.

Given the urgent need to decarbonize the atmosphere and the critical role of NBS in doing so, a fusion approach of eddy covariance and remote sensing data proved to be an efficient and cost-effective solution for quantifying these efforts. We hope our methods could contribute to mitigating global climate change in a cost-effective and transparent way.

Author Contributions: Conceptualization, R.Z. and A.D.; methodology, R.Z. and A.D.; software, R.Z. and A.D.; validation R.Z., writing - original draft preparation, R.Z., A.D., A.L.D.S., O.D.; writing - review and editing, A.D., O.D., G.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: Not applicable

Acknowledgments: We thank Igor Susman (Chief Technical Officer at CarbonSpace Ltd) and the IT team of CarbonSpace for developing infrastructure and supporting our calculations. We thank Ray Mercedes (Head of Innovation at CarbonSpace Ltd) for extremely valuable suggestions, contributions to the paper, and overall support of the work. We thank Andrey Samodin (Chief Marketing Officer at CarbonSpace Ltd) for significant contribution to the draft preparation.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Not applicable.

References

1. Fargione, J. E.; Bassett, S., Boucher, T., Bridgham, S. D., Conant, R. T., Cook-Patton, S. C., Ellis, P. W., Falcucci, A., Fourqurean, J. W., Gopalakrishna, T., Gu, H., Henderson, B., Hurteau, M. D., Kroeger, K. D., Kroeger, T., Lark, T. J., Leavitt, S. M., Lomax, G., McDonald, R. I., and Griscom, B. W Natural climate solutions for the United States. *Science Advances* **2018**, 4(11).

2. Novick, K. A.; Metzger, S., Anderegg, W. R. L., Barnes, M., Cala, D. S., Guan, K., Hemes, K. S., Hollinger, D. Y., Kumar, J., Litvak, M., Lombardozzi, D., Normile, C. P., Oikawa, P., Runkle, B. R. K., Torn, M., and Wiesner, S. Informing Nature-based Climate Solutions for the United States with the best-available science. *Global Change Biology* **2022**, 28, 3778–3794.

3. Seddon, N.; Chausson, A., Berry, P., Girardin, C. A., Smith, A., and Turner, B. Understanding the value and limits of nature-based solutions to climate change and other global challenges. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2020**, 375(1794).

4. Cook-Patton, S. C.; Leavitt, S. M., Gibbs, D., Harris, N. L., Lister, K., Anderson-Teixeira, K. J., Briggs, R. D., Chazdon, R. L., Crowther, T. W., Ellis, P. W., Griscom, H. P., Herrmann, V., Holl, K. D., Houghton, R. A., Larrosa, C., Lomax, G., Lucas, R., Madsen, P., Malhi, Y., and Griscom, B. W. Mapping carbon accumulation potential from global natural forest regrowth. *Nature* **2020**, 585(7826), 545–550.

5. Heavens, N. G.; Ward, D. S., and Natalie, M. M. Studying and Projecting Climate Change with Earth System Models. *Nature Education Knowledge* **2013**, 4(5), 4.

6. Fisher, R. A.; Koven, C. D., Anderegg, W. R. L., Christoffersen, B. O., Dietze, M. C., Farrior, C. E., Holm, J. A., Hurtt, G. C., Knox, R. G., Lawrence, P. J., Lichstein, J. W., Longo, M., Matheny, A. M., Medvigy, D., Muller-Landau, H. C., Powell, T. L., Serbin, S. P., Sato, H., Shuman, J. K., and Moorcroft, P. R. Vegetation demographics in Earth System Models: A review of progress and priorities. *Global Change Biology* **2018**, 24(1), 35–54.

7. Wang, H.; Jiang, F., Wang, J., Ju, W., and Chen, J. M. Terrestrial ecosystem carbon flux estimated using GOSAT and OCO-2 XCO2 retrievals. *Atmospheric Chemistry and Physics* **2019**, 19, 12067–12082.

8. Kondo, M.; Patra, P. K., Sitch, S., Friedlingstein, P., Poulter, B., Chevallier, F., Ciais, P., Canadell, J. G., Bastos, A., Lauerwald, R., Calle, L., Ichii, K., Anthoni, P., Arneth, A., Haverd, V., Jain, A. K., Kato, E., Kautz, M., Law, R. M., Lienert, S., Lombardozzi, D., Maki, T., Nakamura, T., Peylin, P., Rödenbeck, C., Zhuravlev, R., Saeki, T., Tian, H., Zhu, D., and Ziehn, T. State of the science in reconciling top-down and bottom-up approaches for terrestrial CO2 budget. *Global change biology* **2020**, 26(3):1068-84.

9. Dara, A.; Baumann, M., Kuemmerle, T., Pflugmacher, D., Rabe, A., Griffiths, P., Hölzel, N., Kamp, J., Freitag, M., and Hostert, P. Mapping the timing of cropland abandonment and recultivation in northern Kazakhstan using annual Landsat time series. *Remote Sensing of Environment* **2018**, 213, 49–60.

10. Quegan, S.; Le Toan, T., Chave, J., Dall, J., Exbrayat, J.-F., Minh, D. H. T., Lomas, M., D’Alessandro, M. M., Paillou, P., Papathanassiou, K., Rocca, F., Saatchi, S., Scipal, K., Shugart, H., Smallman, T. L., Soja, M. J., Tebaldini, S., Ulander, L., Villard, L., and Williams, M. The European Space Agency BIOMASS mission: Measuring forest above-ground biomass from space. *Remote Sensing of Environment* **2019**, 227, 44–60.

11. Rodríguez-Veiga, P.; Wheeler, J., Louis, V., Tansey, K., and Balzter, H. Quantifying forest biomass carbon stocks from space. *Current Forestry Reports* **2017**, 3(1), 1–18.

12. Yuan, W.; Cai, W., Xia, J., Chen, J., Liu, S., Dong, W., Merbold, L., Law, B., Arain, A., Beringer, J., Bernhofer, C., Black, A., Blanken, P. D., Cescatti, A., Chen, Y., Francois, L., Gianelle, D., Janssens, I. A., Jung, M., Kato, T., Kiely, G., Liu, D., Marcolla, B., Montagnani, L., Raschi, A., Roupsard, O., Varlagin, A., and Wohlfahrt, G. Global comparison of light use efficiency models for simulating terrestrial vegetation gross primary production based on the LaThuile database. *Agricultural and Forest Meteorology* **2014**, 192-193:108–120.

13. Zhang, L. X.; Zhou, D. C., Fan, J. W., and Hu, Z. M. Comparison of four light use efficiency models for estimating terrestrial gross primary production. *Ecological Modelling* **2015**, 300:30–39.

14. Wu, C.; Munger, J. W., Niu, Z., and Kuang, D. Comparison of multiple models for estimating gross primary production using MODIS and eddy covariance data in Harvard Forest. *Remote Sensing of Environment* **2010**, 114(12):2925–2939.

15. Wu, C.; Chen, J. M., and Huang, N. Predicting gross primary production from the enhanced vegetation index and photosynthetically active radiation: Evaluation and calibration. *Remote Sensing of Environment* **2011**, 115(12):3424–3435.

16. Ueyama, M.; Ichii, K., Iwata, H., Euskirchen, E. S., Zona, D., Rocha, A. V., Harazono, Y., Iwama, C., Nakai, T., and Oechel, W. C. Upscaling terrestrial carbon dioxide fluxes in Alaska with satellite remote sensing and support vector regression. *Journal of Geophysical Research: Biogeosciences* **2013**, 118(3):1266–1281.

17. Xie, X.; Li, A., Tan, J., Jin, H., Nan, X., Zhang, Z., Bian, J., and Lei, G. Assessments of gross primary productivity estimations with satellite data-driven models using eddy covariance observation sites over the northern hemisphere. *Agricultural and Forest Meteorology* **2020**, 280.

18. Gardin, L.; Chiesi, M., Fibbi, L., and Maselli, F. Mapping soil organic carbon in Tuscany through the statistical combination of ground observations with ancillary and remote sensing data. *Geoderma* **2021**, 404:115386.

19. Tramontana, G.; Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences* **2016**, 13, 4291–4313.

20. Dubovik, O.; Schuster, G. L., Xu, F., Hu, Y., Bösch, H., Landgraf, J., and Li, Z. Grand Challenges in Satellite Remote Sensing. *Frontiers in Remote Sensing* **2021**, 2.

21. Burba, G.; *Eddy Covariance Method for Scientific, Regulatory, and Commercial Applications*. LI-COR Biosciences; Publisher: Lincoln, USA, 2022; pp. 702.

22. Campioli, M.; Malhi, Y., Vicca, S., Luyssaert, S., Papale, D., Penuelas, J., Reichstein, M., Migliavacca, M., Arain, A., and Janssens, A. Evaluating the convergence between eddy-covariance and biometric methods for assessing carbon budgets of forests. *Nat Commun* **2016**, 7, 13717.

23. Kirschbaum, M.U.F.; Eamus, D., Gifford, R.M., Roxburgh, S.H., and Sands, P.J. Definitions of some ecological terms commonly used in carbon accounting. In *Proceedings Net Ecosystem Exchange CRC Workshop*, (18-20 April, 2001); p 2-5.

24. Luyssaert, S.; Reichstein, M., Schulze, E.-D., Janssens, I. A., Law, B. E., Papale, D., Dragoni, D., Goulden, M. L., Granier, A., Kutsch, W. L., Linder, S., Matteucci, G., Moors, E., Munger, J. W., Pilegaard, K., Saunders, M., and Falge, E. M. Toward a consistency cross-check of eddy covariance flux-based and biometric estimates of ecosystem carbon balance. *Global Biogeochemical Cycles* **2009**, 23(3).

25. Aubinet, M.; Vesala, T., and Papale, D.; *Eddy Covariance: A Practical Guide to Measurement and Data Analysis*, 1st ed.; Publisher: Springer Dordrecht, 2012; pp. 438.

26. Foken, T.; *Micrometeorology*, 2nd ed.; Publisher: Springer Berlin, Heidelberg, 2017; pp. 362.

27. Pastorello, G. Z.; D. Papale, H. Chu, C. Trotta, D. A. Agarwal, E. Canfora, D. D. Baldocchi, and M. S. A new data set to keep a sharper eye on land-air exchanges. *Eos* **2017**, 98.

28. Burba, G.; Illustrative Maps of Past and Present Eddy Covariance Measurement Locations: II. High-Resolution Images. <https://www.researchgate.net> **2019**, 9.

29. Baldocchi, D.; How eddy covariance flux measurements have contributed to our understanding of global change biology. *Global Change Biology* **2020**, 26(1), 242-260.

30. Baldocchi, D.; Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, Ch., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw U, K. T., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K., and Wofsy, S. FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities. *Bulletin of the American Meteorological Society* **2001**, 82(11), 2415-2434.

31. Friend, A.D.; A. Arneth, N.Y. Kiang, M. Lomas, J. Ogée, C. Rödenbeck, S.W. Running, J.-D. Santaren, S. Sitch, N. Viovy, F.I. Woodward, and S. Zaele FLUXNET and modelling the global carbon cycle. *Glob. Change Biol.* **2007**, 13, 610-633.

32. Pastorello, G.; Trotta, C., Canfora, E. et al. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci Data* **2020**, 7, 225.

33. Williams, M.; Richardson, A. D., Reichstein, M., Stoy, P. C., Peylin, P., Verbeeck, H., Carvalhais, N., Jung, M., Hollinger, D. Y., Kattge, J., Leuning, R., Luo, Y., Tomelleri, E., Trudinger, C. M., and Wang, Y. P. Improving land surface models with FLUXNET data. *Biogeosciences* **2009**, 6, 1341–1359.

34. Fu, D.; Chen, B., Zhang, H., Wang, J., Black, T. A., Amiro, B. D., Bohrer, G., Bolstad, P., Coulter, R., Rahman, A. F., Dunn, A., McCaughey, J. H., Meyers, T., and Verma, S. Estimating Landscape Net Ecosystem Exchange At High Spatial–Temporal Resolution

Based On Landsat Data, An Improved Upscaling Model Framework, And Eddy Covariance Flux Measurements. *Remote Sensing Of Environment* **2014**, *141*, 90-104.

35. Congalton, R.G.; and Green, K. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, 2nd ed.; Publisher: CRC Press, Boca Raton, 2008; pp. 200.

36. Crist, E.P.; A TM tasseled cap equivalent transformation for reflectance factor data. *Remote Sens. Environ* **1985**, *17*, 301–306.

37. Ren, H.; Zhou, G., and Zhang, F. Using negative soil adjustment factor in soil-adjusted vegetation index (SAVI) for aboveground living biomass estimation in arid grasslands. *Remote Sens. Environ* **2018**, *209*, 439–445.

38. Dara, A.; Baumann, M., Freitag, M., Hölzel, N., Hostert, P., Kamp, J., Müller, D., Prishchepov, A. V., and Kuemmerle, T. Annual Landsat time series reveal post-Soviet changes in grazing pressure. *Remote Sensing of Environment* **2020**, *239*, 111667.

39. Boogaard, H.; Schubert, J., De Wit, A., Lazebnik, J., Hutjes, R., and Van der Grijn, G. Agrometeorological indicators from 1979 to present derived from reanalysis, version 1.0. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 30-Jun-2022) **2020**, .

40. Leclerc, M.Y.; and Foken, T. *Footprints in micrometeorology and ecology*, 1st ed.; Publisher: Springer Berlin, Heidelberg, 2014; pp. 239.

41. Kljun, N.; Calanca, P., Rotach, M. W., and Schmid, H. P. A simple two-dimensional parameterisation for flux footprint prediction (FFP). *Geosci. Model Dev.* **2015**, *8*(11), 3695-3713.

42. Griebel, A; Bennett, L T., Metzen, D, Cleverly, J, Burba, G, Arndt, S K. Effects of inhomogeneities within the flux footprint on the interpretation of seasonal, annual, and interannual ecosystem carbon exchange. *Agricultural and Forest Meteorology* **2016**, *221*, 50-60.

43. Griebel, A.; Metzen, D., Pendall, E., Burba, G., Metzger, S. Generating spatially robust carbon budgets from flux tower observations. *Geophysical Research Letters* **2020**, *47*(3).

44. Stohl, A.; Forster, C., Frank, A., Seibert, P., and Wotawa, G. Technical note: The Lagrangian particle dispersion model FLEXPART version 6.2. *Atmos. Chem. Phys.* **2005**, *5*, 2461–2474.

45. Stohl, A.; Seibert, P., Arduini, J., Eckhardt, S., Fraser, P., Grealley, B. R., Lunder, C., Maione, M., Mhle, J., O'Doherty, S., Prinn, R. G., Reimann, S., Saito, T., Schmidbauer, N., Simmonds, P. G., Vollmer, M. K., Weiss, R. F., and Yokouchi, Y. An analytical inversion method for determining regional and global emissions of greenhouse gases: Sensitivity studies and application to halocarbons. *Atmos. Chem. Phys.* **2009**, *9*, 1597–1620.

46. Murphy, K. P.; 14.4.3. Kernelized ridge regression. In *Machine Learning: A Probabilistic Perspective*; Publishing House: The MIT Press, 2012; pp. 492–493.

47. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Publisher: Cambridge University Press, 2004; pp. 462.

48. Vapnik, V.; Golowich, S. E., Smola A. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. In *Advances in Neural Information Processing Systems 9*; Publishing House: MIT Press, 1996; pp. 281–287.

49. Raschka, S.; MLxtend: Providing machine learning and data science utilities and extensions to Python’s scientific computing stack. *J Open Source Softw* **2018**, *3*(24).

50. Kozachenko, L. F.; Leonenko, N. N. Sample Estimate of the Entropy of a Random Vector. *Problems Inform. Transmission* **1987**, *95–101*.

51. Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *J. Geophys. Res.-Biogeo.* **2011**, *116*.