# Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms

Gianluca Tramontana[1], Martin Jung[2], Christopher R. Schwalm[3], Kazuhito Ichii[4,5], Gustau Camps-Valls[6], Botond Ráduly[1,7], Markus Reichstein[2], M. Altaf Arain[8], Alessandro Cescatti[9], Gerard Kiely[10], Lutz Merbold[11,12], Penelope Serrano-Ortiz[13], Sven Sickert[14], Sebastian Wolf[11] and Dario Papale[1].

[1]Department for Innovation in Biological, Agro-food and Forest systems (DIBAF), Univeristy of Tuscia, Viterbo, 01100, Italy

[2]Max Planck Institute for Biogeochemistry, Jena, 07745, Germany

[3]Woods Hole Research Center, Falmouth MA, 02540, USA

[4]Department of Environmental Geochemical Cycle Research, Japan Agency for Marine-Earth Science and Technology, Yokohama, 236-0001, Japan

[5]Center for Global Environmental Research, National Institute for Environmental Studies, Tsukuba, 305-8506, Japan

[6]Image Processing Laboratory (IPL), Universitat de València, Paterna (València), 46980, Spain

[7]Department of Bioengineering, Sapientia Hungarian University of Transylvania, MiercureaCiuc, 530104, Romania

[8]School of Geography and Earth Sciences, McMaster University, Hamilton (Ontario), L8S4L8, Canada

[9]European Commission, Joint Research Centre, Directorate for Sustainable Resources, Ispra,Italy

[10]Civil & Environmental Engineering and Environmental Research Institute, University College, Cork, T12 YN60, Ireland

[11]Department of Environmental Systems Science, Institute of Agricultural Sciences, ETH Zurich, Zurich, 8092, Switzerland

[12] Mazingira Centre, Livestock Systems and Environment, International Livestock Research Institute (ILRI), 00100, Nairobi, Kenya

[13]Department of Ecology, University of Granada, Granada, 18071, Spain

[14]Computer Vision Group, Friedrich Schiller University Jena, 07743 Jena, Germany

*Correspondence to:* G. Tramontana (g.tramontana@unitus.it)

**Abstract.** Spatio-temporal fields of land-atmosphere fluxes derived from data-driven models can complement simulations by process-based Land Surface Models. While a number of strategies for empirical models with eddy covariance flux data have been applied, a systematic intercomparison of these methods is missing so far. In this study, we performed a cross-validation experiment for predicting carbon dioxide, latent heat, sensible heat and net radiation fluxes, across different ecosystem types with eleven machine learning (ML) methods from four different classes (kernel methods, neural networks, tree methods, and regression splines). We applied two complementary setups: (1) 8-day average fluxes based on remotely sensed data, and (2) daily mean fluxes based on meteorological data and mean seasonal cycle of remotely sensed variables. The pattern of predictions from different ML and experimental setups were highly consistent. There were systematic differences in performance among the fluxes, with the following ascending order: net ecosystem exchange ($R^2<0.5$), ecosystem respiration ($R^2>0.6$), gross primary production ($R^2>0.7$), latent heat ($R^2>0.7$), sensible heat ($R^2>0.7$), net radiation ($R^2>0.8$). The ML methods predicted the across site variability and the mean seasonal cycle of the observed fluxes very well ($R^2> 0.7$), while the 8-day deviations from the mean seasonal cycle were not well predicted ($R^2< 0.5$). Fluxes were better predicted at forested and temperate climate sites than at sites in extreme climates or less represented by training data (e.g. the tropics). The evaluated large ensemble of ML based models will be the basis of new global flux products.

*Keywords*: Machine learning, carbon fluxes, energy fluxes, FLUXNET, remote sensing, FLUXCOM

## 1. Introduction

Improving our knowledge of the carbon, water, and energy exchanges between terrestrial ecosystems and the atmosphere is essential to better understand and model the Earth's climate system (IPCC, 2007; Reich, 2010). In situ continuous observations can be obtained with the eddy covariance technique, which estimates the net exchanges of carbon dioxide ($CO_2$), water vapor and energy between land ecosystems and the atmosphere (Aubinet at al., 2012; Baldocchi et al.,2014). The large-scale measurement network, FLUXNET integrates site observations of these fluxes globally and provides detailed time series of carbon and energy fluxes across biomes and climates (Baldocchi et al., 2008). However, eddy covariance measurements are site-level observations (at < 1 $km^2$ scale), and spatial upscaling is required to estimate these fluxes at regional to global scales.

The increasing number of eddy covariance sites across the globe has encouraged the application of data-driven models by machine learning (ML) methods such as Artificial Neural Networks (ANNs, Papale et al., 2003), Random Forest (RF, Tramontana et al., 2015), Model Trees (MTE, Jung et al., 2009; Xiao et al., 2008, 2010) or Support Vector Regression (SVR, Yang et al., 2006, 2007) to estimate land surface-atmosphere fluxes from site level to regional or global scales (e.g. Beer et al., 2010, Jung et al., 2010, 2011; Kondo et al., 2015; Schwalm et al., 2010, 2012; Yang et al., 2007; Xiao et al., 2008, 2010). The ML upscaled outputs are also increasingly used to evaluate process based land surface models (e.g., Anav et al., 2013; Bonan et al., 2010; Ichii et al., 2009; Piao et al., 2013).

The key characteristic of data-driven models compared to process-based ones are the former's intrinsic observational nature, and the fact that functional relationships are not prescribed but rather emerge from patterns found in the measurements. In this context, data-driven models extract multivariate functional relationships between the in situ measured fluxes of the network and explanatory variables. These variables are derived from satellite remote sensing, providing useful (although partial) information on vegetation state (e.g., vegetation indices) and other land surface properties (e.g., surface temperature), along with continuous measurements of meteorological variables at flux towers.

While ML-based upscaling provides a systematic approach to move from point-based flux estimates to spatially explicit gridded fields, various sources of uncertainty exist. For example, individual ML methods can have different responses, especially when these models are applied beyond the conditions represented in the training dataset (Jung et al., 2009; Papale et al., 2015). The information content of the driving input variables may not be sufficient to capture the variability of the fluxes in all conditions (Tramontana et al., 2015). Moreover, remotely sensed and meteorological gridded datasets are affected by uncertainties themselves. Remote sensing data contain noise, biases and gaps, and can be perturbed by atmospheric effects or by the presence of snow. Meteorological gridded datasets are known to contain product specific biases as well (Garnaud et al., 2014; Tramontana et al., 2015; Zhao et al., 2012).

Thorough experiments using multiple data-driven models and explanatory variables are an essential step to identify and assess limitations and sources of uncertainty in the empirical upscaling approach. For this reason several experts in the field gathered together and formed the collaborative FLUXCOM initiative. FLUXCOM aims to better understand the multiple sources and facets of uncertainties in empirical upscaling and, ultimately, to provide an ensemble of machine learning based global flux products to the scientific community. In FLUXCOM we selected machine learning based regression tools that span the full range of commonly applied algorithms: from model tree ensembles, multiple adaptive regression splines, artificial neural networks, to kernel methods, with several representatives of each family. We defined common protocols for two complementary upscaling strategies (setups) based on: (1) 8-day averaged fluxes based on exclusively remotely sensed data, and (2) daily mean fluxes based on remotely sensed and meteorological data. Different ML approaches were then applied to both setups using the same sets of predictor variables, and a thorough 'leave-towers-out' cross-validation was conducted. This study presents the FLUXCOM results obtained from the cross-validation. Our overarching aim was to understand how well fluxes of $CO_2$ (gross primary production (GPP), terrestrial ecosystem respiration (TER) and net ecosystem exchange (NEE)), and energy (latent heat (LE), sensible heat (H) and net radiation (Rn)), as estimated by the eddy covariance technique, are predicted by an ensemble of ML methods. We focused in particular on the ensemble median prediction because the ensemble median global product will likely be used extensively. At first we looked at the consistency of the patterns between the two experimental setups to understand whether satellite remote sensing is sufficient for mapping carbon and energy fluxes or whether instantaneous meteorological conditions need to be considered explicitly. Second, we investigated which characteristics of the predicted fluxes were robust, analyzing how well the median estimates were able to predict the across site variability, the mean seasonal cycle by site and interannual variation, i.e., time-dependent deviations from the mean seasonal cycle. Thirdly, we investigated how the ML performance varies among climate zones or ecosystem types.

## 2 Material and methods

### 2.1 Data

### 2.1.1 Eddy covariance study sites

We used eddy covariance data from 224 flux-tower sites (supplementary material, Sect. S1), which originate from the FLUXNET La Thuile synthesis dataset and CarboAfrica network (Valentini et al., 2014). The study sites were distributed globally and cover most plant functional types (PFT) and biomes over the globe (Table 1).

### 2.1.2 Observation-based $CO_2$ and energy fluxes

All flux measurements were post-processed using standardized procedures of quality control (Papale et al. 2006) and gap-filled following Reichstein et al. (2005). Estimates of GPP and TER were derived from half-hourly NEE measurements using two independent flux partitioning methods: (1) According to Reichstein et al. (2005), where the temperature sensitivity of ecosystem respiration was initially estimated from night-time NEE data and then extrapolated to daytime to estimate TER and GPP. This was done by subtracting NEE (negatively signed for the $CO_2$ uptake) from TER. (2) According to Lasslop et al. (2010), where daytime NEE data were used to constrain an hyperbolic light response curve to directly estimate GPP and TER. In the following we refer to GPP and TER as derived by Reichstein et al. (2005) as $GPP_R$ and $TER_R$; whereas estimates based on the Lasslop et al. (2010) method are referred to as $GPP_L$ and $TER_L$.

Half-hourly data were aggregated to daily values and screened according to multiple quality criteria, as follows:

1) We excluded data when more than 20% of the data were based on gap-filling with low confidence (Reichstein et al., 2005).

2) We identified and removed obviously erroneous periods due to non-flagged instrument or flux partitioning failures based on visual interpretation.

3) We excluded data-points where the two flux-partitioning methods provided extremely different patterns. Specifically, we computed for each site a robust linear regression between (a) $TER_R - GPP_L$ and NEE, and (b) $GPP_R$ and $GPP_L$. Data points with a residual outside the range of ± 3 times of the inter-quartile range were removed. This criterion removed only the extreme residuals, systematic differences between methods were not removed.

4) We removed the 5% of data-points with the largest friction velocity (u*) uncertainty, defined as data points above the 95th percentile of daily u* uncertainty, measured as the inter-quartile range of 100 bootstrap samples (Papale et al., 2006).

We applied the same criteria 1) and 2) above for the energy fluxes as we did for the $CO_2$ fluxes,. Additionally, we removed data with inconsistent energy fluxes, i.e. when the residual of a robust linear regression between LE + H and Rn for each site was outside three-times the inter-quartile range of the residuals.

### 2.1.2 Remote sensing data

We collected data from the Moderate Resolution Imaging Spectroradiometer (MODIS) which provided data at a spatial resolution of 1km or better (Justice et al., 2002). We used MODIS cutouts of 3×3 km pixels centered on each tower to reduce the effect of geolocation error and to better representing the eddy covariance footprint area (Xiao et al., 2008). We used the following products: MOD11A2 Land Surface Temperature (LST) (Wan et al., 2002); MOD13A2 Vegetation Index (Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI), (Huete et al., 2002); MOD15A2 Leaf Area Index (LAI) and fraction of Absorbed Photosynthetic Active Radiation (fPAR) (Myneni et al., 2002); MCD43A2 and MCD43A4 Bidirectional Reflectance Distribution Function (BRDF) corrected surface reflectances (Schaaf et al., 2002). The BRDF-corrected surface reflectance data were further processed to calculate the Normalized Difference Water Index (NDWI) (Gao, 1996) and the Land Surface Water Index (LSWI) (Xiao et al., 2002). These data were obtained from http://daac.ornl.gov/MODIS/.

The remote sensing data were further processed to improve data quality and data gaps were filled to create continuous time-series data, and to minimize non-land surface signals In particular, we identified good quality pixels by the using the quality assurance/quality control (QA/QC) included in the MODIS product. If more than 25% of the pixels had good quality at the time of snapshot, the average of good quality pixels were assigned as the actual value. Otherwise, the data at the time snapshot were marked as blank (no data). Then, we created the mean seasonal variations from 2000-2012 using only good pixels data and the data gaps in the processed data were filled using the mean seasonal variation. Only MOD13 was provided with 16-day composites, and 8-day data were created by assigning the 16-day composite value to the corresponding two 8-day periods.

### 2.1.3 Meteorological data

The air temperature (Tair), global radiation (Rg), vapor pressure deficit (VPD), and precipitation (in situ measured at the flux towers location) were used after data screening according to the criteria 1) and 2) as applied for the measured fluxes (see Sect. 2.1.2). We also used long-term time series of these variables from the dataset ERA-Interim (Dee et al., 2011) for the period 1989-2010, which were bias-corrected for each site based on the period of overlap with the in situ measurements (see http://www.bgc-jena.mpg.de/~MDIwork/meteo/). These long-term meteorological data were primarily used to calculate consistent metrics of climatological variables (e.g. mean annual temperature) for all sites given the temporal coverage of data of the different sites. In addition, we used a composite of these ERA-Interim data and in situ measured data to obtain a gap-free time series for calculating a soil Water Availability Index (WAI, see Sect. 2.3.2 and supplementary material, Sect. S3).

## 2.2 Applied ML methods

For our purpose, eleven ML algorithms for regression from four broad families were chosen: tree-based methods, regression splines, neural networks and kernel methods. Moreover a comprehensive review of ML algorithms in biophysical parameter estimation can be found in Verrelst et al. (2015). At follow a brief description of the characteristics of each family.

Tree based methods

These methods construct hierarchical binary decision trees. The inner nodes of the tree hold decision rules according to explanatory variables (e.g. less/greater than X1), recursively splitting the data into subspaces. The leaf nodes at the end of the decision tree contain models for the response variable. Because a single tree is generally not effective enough to cope with strong non-linear multivariate relationships, ensembles of trees are often used. We applied two different tree ensemble methods: (1) Random Forests (RF) which combines regression trees grown from different bootstrap samples and randomly selected features at each split node (Breiman, 2001; Ho, 1998); and (2) Model Tree Ensembles (MTE) which combine model trees (Jung et al., 2009). The main difference between regression and model trees is the prediction model in the leaf node: a simple mean of the target values from the training in regression trees and a parametric function (here a multiple linear regression) in model trees. In this study, we used three different variants of MTE, which differ mainly with respect to different cost functions for determining the splits, and the technique to create the ensemble of model trees. Further details are described in the supplementary material (Sect. S2).

Regression splines

Multivariate regression splines (MARS) are an extension of simple linear regression adapted to non-linear response surfaces using piecewise (local) functions. The target variable is predicted as the sum of regression splines and a constant value (Alonso Fernández, 2013; Friedman et al., 1991).

Neural networks

Neural networks are based on nonlinear and nonparametric regressions. Their base unit is the neuron, where nonlinear regression functions are applied. The neurons are interconnected and organized in layers. The output of *m* neurons in the current layer are the inputs for *n* neurons of the next layer. We used two types of neural networks: the artificial neural network (ANN) and the group method of data handling (GMDH). In an ANN, each neuron performs a linear regression followed by a non-linear function. Neurons of different layers are interconnected by weights that are adjusted during the training (Haykin et al., 1999; Papale et al., 2003). The GMDH is a self-organizing inductive method (Ungaro et al., 2005) building polynomials of polynomials; the neurons are pairwise connected through a quadratic polynomial to produce new neurons in the next layer (Shirmohammadi et al., 2015).

Kernel methods

Kernel methods (Shawe-Taylor and Cristianini, 2004; Camps-Valls and Bruzzone, 2009) owe their name to the use of kernel functions, which measure similarities between input data examples. Among the available kernel methods we used: (1) support vector regression (SVR) (Vapniket al., 1997), (2) kernel ridge regression (KRR) (Shawe-Taylor and Cristianini, 2004), and (3) Gaussian process regression (GPR) (Rasmussen, 2006). The SVR defines a linear prediction model over mapped samples to a much higher dimensional space, which is non-linearly related to the original input (Yang et al., 2007). The KRR is considered as the kernel version of the regularized least squares linear regression (Shawe-Taylor and Cristianini, 2004). The GPR is a probabilistic approximation to nonparametric kernel-based regression, and both a predictive mean (point-wise estimates) and predictive variance (error bars for the predictions) can be derived. We also used a hybrid approach combining RF with simple decision stumps in the inner nodes and GPR for prediction in the leaf nodes (Fröhlich et al., 2012).

## 2.3 Experimental design

### 2.3.1. Experiment setups

We defined two complementary experimental setups, which differ in the choice of explanatory variables, and the temporal resolution of the target fluxes: 1) at 8-day temporal resolution using exclusively remote sensing data (hereafter RS); and 2) at daily temporal resolution using meteorological data together with the mean seasonal cycle (MSC) of the remote sensing data (hereafter RS+METEO). In the latter case, the MSC of remote sensing data were smoothed and interpolated to a daily time step. Each setup represents a trade-off between spatial and temporal resolution. While RS provides products with high spatial resolution for global upscaling (e.g. 1km), the temporal resolution is coarse (8-day vs. daily) and temporal coverage is limited to the period when satellite observation is available (e.g. 2000-present in the case of MODIS). The uncertainties of remote sensing data at tower locations, due to finer scale spatial heterogeneity, also degraded the performance of the ML methods. In contrast, RS+METEO takes advantage of information from meteorological variables and was resistant to the noise of remote sensing time series because only the mean seasonal cycle of data from satellite RS were used. RS+METEO also allowed for upscaled products over a longer time period (because not constrained by the availability of MODIS data) and finer time scale (daily). Furthermore, the use of meteorological gridded datasets introduced uncertainty due to dataset specific biases and the coarser spatial resolution ( $\geq$ 0.5 degrees or coarser).

### 2.3.2. Variable selection

Combining remote sensing and meteorological data (see Sect. 2.1.2 and 2.1.3) we created additional explanatory variables. In the case of RS+METEO setup we derived the Water Availability Index (WAI) based on a soil water balance model (for more details see supplementary material, Sect. S3) to represent water stress conditions appropriately. For both setups we derived proxies for absorbed radiation as the product between vegetation greenness (e.g. EVI, NDVI, fPAR) and drivers related to the useful energy for photosynthesis(e.g. daytime LST, Rg, and potential radiation). Other derived variables included the MSC of dynamic variables (e.g., LST, fPAR, Rg, air temperature) and associated metrics (minimum, maximum, amplitude, and mean). For remote sensing predictors, the MSC and associated metrics were based on the period 2001-2012, while for climate variables were based on the bias corrected daily long-term ERA-Interim data reference period (1989-2010). In total, 216 potential explanatory variables were created for RS and 231 for RS+METEO (see supplementary material S4 for details).

For each of the two experimental setups we selected a small subset of variables optimally suitable to predict target fluxes using a variable selection search algorithm. Variable selection was an important component in the spatial upscaling since it improved the accuracy of predictions, while the computational costs of the global predictions were minimized. We used the Guided Hybrid Genetic Algorithm (GHGA; Jung and Zscheischler, 2013), which was designed for variable selection problems with many candidate predictor variables and computationally expensive cost functions. The GHGA required the training of a regression algorithm (here RF) to estimate the cost associated with selected variable subsets. We executed GHGA selection runs for the RS and RS+METEO setups and separately for $CO_2$ and energy fluxes (see S5 for details). All ML used exactly the same selected drivers (listed in Table 2) to made predictions. This procedure had the advantage that the resulting global products will be originated from a consistent set of predictor variables.

### 2.3.3. Algorithm training

The capability of ML methods to spatially extrapolate $CO_2$ and energy fluxes was evaluated using a 10-fold cross-validation strategy. The training datasets were stratified into 10-folds, each containing ca. 10% of the data. Entire sites were assigned to each fold (Jung et al., 2011). The training of each ML method was done using data from nine folds while predictions were made for the remaining one. This was repeated 10 times and each fold was used exactly once as a validation set, thus ensuring that the validation data were completely independent from the training data. Due to the computational expense of the RS+METEO setup, only one method representing each "family" – RF, MARS, ANN and KRR – was trained. ML methods hyperparameters (that account for regularization in order to avoid overfitting as well as for the shape and smoothness constraints) (see supplementary material S6 for details), were estimated in each fold.

### 2.3.4. Model evaluation

To highlight the differences between the RS and RS+METEO setups, the daily output from RS+METEO were aggregated to 8-day time steps; predictions from the same periods and sites were used for the comparison. Besides the statistical analysis of the individual ML cross-validation results, we focused on the ensemble median estimate, here defined as the median predicted value across all ML for a given setup and time step. We used a suite of metrics to evaluate the ML performance: the Nash and Sutcliffe model efficiency (MEF) (Nash and Sutcliffe, 1970); the root mean square error (RMSE); the empirical BIAS; the Pearson's linear correlation coefficient (ρ); the coefficient of determination ($R^2$); and the ratio of variance (ROV).

MEF is a measure of the capability of a model to estimate a target variable better than a reference, generally the mean value of the observations. In our study MEF was calculated as:

$$MEF = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

where $x_i$ and $y_i$ were the predicted and the observed values respectively and $\bar{y}$ is the mean value of the observations. MEF varied between -inf to 1; in the case of MEF > 0 the predictive capacity of the model was better than the mean (MEF = 1 for the ideal model), instead if MEF=0 the predictive capacity of the model was equivalent to the mean, finally if MEF < 0, the predictive capacity of the mean value of the target was better than the model.

The RMSE was estimated as the root square of the mean value of the squared residuals:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \quad (2)$$

The BIAS was evaluated as the mean value of model's residuals

$$BIAS = \frac{\sum_{i=1}^{n}(x_i - y_{i)}}{n} \quad (3)$$

Following Gupta et al. (2009) the importance of bias on the overall uncertainty was evaluated as the ratio between the square of BIAS and the Mean Square Error, the latter estimated as the square value of RMSE.

The Pearson's linear correlation coefficient (ρ) was the ratio between the covariance between the modeled and observed values ($\sigma_{xy}$) and the product of the standard deviation of modeled ($\sigma_x$) and observed ($\sigma_y$) values:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (4)$$

$R^2$ was estimated as the squared value of ρ; finally ROV was evaluated as the ratio between predicted and observed standard deviation.

We evaluated the overall predictive capacity and consistency of ML approaches—including the ML median estimate—by flux, by experimental setup and by site as well as grouped by Köppen climate zone and International Geosphere-Biosphere Programme (IGBP) plant functional types (PFT). In our evaluation we focused on site-specific means, the mean seasonal cycle (MSC), and anomalies (Jung et al., 2011). The MSC per site was calculated using the averaged values for each 8-day period across the years, but only when at least two values (i.e., years) for each 8-day period were available. To assess the mean values of the study sites, we calculated the mean of the MSC if at least 50% of the 46 8-day values were present, whereas the 8-day anomalies were calculated as the deviation of a flux value from the MSC. Finally, the mean site values were removed from the MSC to disentangle the seasonal variation from the mean site values, making them as complementary.

## 3. Results

### 3.1 Machine learning performance across fluxes

Prediction capability of the ensemble median estimate clustered into tiers whereby energy fluxes were better predicted than $CO_2$ fluxes: Rn > H/LE/GPP > TER > NEE (Table 3 and Table A1). The highest predictive capacity levels as exhibited by net radiation showed near perfect agreement; Rn displayed a model efficiency (MEF) of 0.91-0.92 and a correlation of 0.96. The decline in predictive capacity for the second tier fluxes was ca. 15% to 20%; MEF for H, LE, and GPP is 0.79, 0.75-0.76, and 0.71 respectively. The lowest two tiers exhibited 20% and 40% declines in MEF (0.57-0.64 and 0.43-0.46 for TER and NEE respectively). These relative rankings, consistent with previous studies (Jung et al., 2011; Yuan et al., 2010), were unchanged regardless the metric of the predictive capacity used in cross-validation—apart from RMSE where the difference in fluxes units and magnitude, confounded a direct comparison (Table 3).

There were only minor performance differences between the two $CO_2$ fluxes partitioning methods (Table 3), although for the RS setup, the performance of $TER_L$ were slightly lower than $TER_R$ (lower MEF, ρ and ROV). However, a similar pattern was not found in RS+METEO setup.

Accuracy metrics of median ensemble were similar, by flux, for both RS and RS+METEO approaches, showing that the spatiotemporal variability of remotely sensed land surface properties are appropriate to predict the top tier fluxes (Rn, H, LE, and GPP) (Jung et al., 2008; Tramontana et al., 2015; Xiao et al., 2010;.Yang et al., 2007). We found some minor differences for those fluxes which showed lower overall predictive capacity levels, in particular the NEE and $TER_L$ (Fig. 1, Table 3). MEF and correlation values were slightly larger for RS than RS+METEO but the differences in performances might be due to a different ensemble size, with the RS median ensemble composed of 11 MLs, whereas RS+METEO was based on only four. However, the output provided by MLs methods showed high overall consistency among them, that increased when predictions were obtained by different MLs trained with the same experimental setup (RS else RS+METEO; for more details see Appendix B and Table B1).

### 3.2 Capability to predict the across site variability, the mean seasonal cycle and the deviations from it.

Decomposing FLUXNET data into across-site variability, mean seasonal cycle, and interannual variability components (Sect. 2.3.4) revealed clear gradients in predictive capacity (Table 4 and Fig. 2, Fig. 3a and 3b). Across-site variability was in general well-captured by the ML ($R^2$ range: 0.61 to 0.81 except for NEE), suggesting that the ML methods are suitable to reproduce the spatial pattern of the mean annual fluxes. The variability in the mean seasonal cycle (at 8-day time scale) was also uniformly well predicted, in particular for LE and Rn ($R^2$ between 0.67-0.77 for GPP and TER, and between 0.86-0.98 for the energy fluxes). In contrast, the 8-day anomalies variability were generally poorly captured by all the ML approaches with only H and Rn showing an $R^2$ greater than 0.4. This low predictive skill was regardless of whether 8-day, monthly (Jung et al., 2011), or annual time steps were used (data not shown) and predicting interannual variability remains one of the largest challenges in the context of the empirical upscaling. NEE was confirmed to be the poorest predicted flux (Table 3). ML showed considerably lower predictive capability for NEE, by comparison with the other fluxes for across-sites variability ($R^2$ = 0.46), the mean seasonal cycle ($R^2$ = 0.59), and interannual variability ($R^2$ = 0.13, $TER_L$ was the lowest at 0.10).

### 3.3 Models performance for different climate zones and ecosystem types.

Climate zone and plant functional type (PFT) are important discriminating factor for ML predictive capacity for $CO_2$ fluxes. In general, the mixed forest (MF), the deciduous broadleaved forest (DBF) and the boreal sites (Bor) showed higher accuracy of prediction for the median ensembles (Fig. 4, Tables C1-C6 in Appendix C), even for NEE ($R^2$> 0.6). In contrast, relatively poor prediction capability was found in evergreen broadleaved forest (EBF), in the tropics (Trop), in the extreme environments for reduced water resource (Dry) or low temperature (Cold), and in managed sites such as croplands (Crop). This gradient largely reflects the mismatch between the seasonal dynamics of predicted fluxes and the models drivers. The absence of a clear seasonal cycle in evergreen broadleaf forest and in the tropical sites likely contributed to the low ML performance (in general) in these ecosystems (Sims et al., 2008; Yebra et al., 2015; Yuan et al., 2010). Similarly, cold and dry sites are characterized by both low magnitude and low variance of fluxes, making it difficult to explain the fluxes variability in these ecosystems types using empirical methods.-For the intensively managed croplands the seasonal dynamics of fluxes were highly constrained by management practices (e.g. irrigation, fertilization, tillage) which is not directly reflected in the explanatory variables used in training.

The gradient of prediction capability in different PFT and climate zone was less evident in the case of energy fluxes (not significant in the case of Rn) and the performance of ML were generally good. In fact the median $R^2$ between simulations and observations were greater than 0.7 for more than the 85% of the PFT and climate zone (in all sites for Rn). For comparison in the case of GPP and TER, the median $R^2$ between simulations and observations were greater than 0.6 for more than 75% of the PFT and climate zone.

## 4. Discussion

### 4.1 Comparison between experimental setups

In general the performance metrics across the two experimental setups were highly similar. Very few differences were found decomposing the fluxes variability into across-site variability, mean seasonal cycle, and interannual variability components. This suggests that $CO_2$ and energy fluxes can be mapped exclusively with remotely sensed inputs allowing for high-spatial resolution products without additional uncertainty introduced by gridded meteorological data products (Tramontana et al., 2015). However, differences between the two experimental setups are apparent at PFT and climate zone scales, particularly in the EBF PFT and in the tropics where RS+METEO performs better than RS for predicting $CO_2$ fluxes (e.g. in RS+METEO the decrease in RMSE was 0.10-0.68 $gCm^{-2}d^{-1}$ in comparison to RS). This might be due, from one side to the pattern of $CO_2$ fluxes that do not follow the seasonal pattern of the vegetation indices, from the other side to the increasing importance of meteorological drivers, in particular the ones accounting for the water stress/limitation (e.g. VPD or WAI). In addition, the larger sample size due to the daily resolution of the RS+METEO setup might have been beneficial. The RS setup might also suffer from poorer quality of remote sensing data in the tropics due to frequent cloud coverage. At cropland sites the RS has better predictive capacity than RS+METEO (in comparison to RS, RMSE of RS+METEO increase of 0.02-0.67 $gCm^{-2}d^{-1}$ for predicting $CO_2$ fluxes). This could be related to management (e.g. sowing and harvesting dates) which are partly captured by RS but not by RS+METEO which uses only the mean seasonal cycle of VI reflecting also the better performance of RS for predicting anomalies in the case of CRO, (although in general results for anomalies were not good, data not shown).

Another distinguishing element between RS and RS+METEO is the degree of uncertainty of the drivers. At the site level meteorological drivers (used only in RS+METEO) are generally measured with good quality while remote sensing data are generally affected by additional uncertainties. Scale mismatch between FLUXNET eddy covariance towers and satellite sensor footprints as well as satellite sensors limitations are important sources of uncertainty, not present in the in situ measured meteorological drivers. Furthermore, the quality of remote sensing data is affected by external factors such as the atmospheric condition, cloud cover and ground surface state. These issues were minimized in the RS+METEO by using only the smoothed mean seasonal cycle of satellite data which contains much less noise. We had expected that this would improve the performance of predicting anomalies (in general) with the RS+METEO setup, because anomalies area comparatively small signals that could be readily distorted by the comparatively large noise in remote sensing data. However, we found no clear indication for that in the cross-validation results.

### 4.2 Completeness of predictors

Certainly, the predictor variables used for the ML approaches do not capture all drivers of flux variability both across sites and temporally. For example, in managed sites, external factors such management practices and disturbances (Amiro et al., 2010; Thornton et al., 2002) are likely crucial. In addition, direct estimates of soil moisture would improve the prediction capability in dry environments. The absence of important drivers of flux variability in the predictor sets likely explains why NEE and TER is less well predicted compared to GPP. First order constraints of GPP such as radiation, temperature, and canopy properties are accounted for in the predictors. For TER some important factors like soil properties and carbon pools (Amiro et al., 2010) are not well presented in the predictor variables. For NEE, several studies have shown its dependence on long-term lag and memory effects (Bell et al., 2012; Frank et al., 2015, Papale et al., 2015; Paruelo et al., 2005) that are not accounted for by the drivers used in this study. Adding targeted variables (e.g. soil carbon stock, turnover of the soil organic matter, lagged drivers) among the candidate predictors and carrying out the driver selection for each $CO_2$ flux specifically, could improve the ML performance for each flux. At the same time it could be a detriment for the spatial upscaling of $CO_2$ fluxes generating less consistent global products with plausible bad effects on $CO_2$ uptake/release balance closure over the land ecosystems. Moreover, the choice of predictor variables for the ML approaches is limited in practical terms by the availability of consistent observations across all sites on the one hand, and on the availability of a corresponding consistent global gridded product for upscaling. Therefore, continued efforts of metadata collection at the sites in conjunction with large scale inventories and new Earth Observations are needed to improve the ML approach in the future.

### 4.3 Quality of the response variable

The predictive capacity of ML approaches also depends on the uncertainties of the flux variables themselves. Clearly, there is some variability in the target flux variables which is due to noise and measurement problems, and this portion of variability cannot (and should not!) be reproduced by the ML approaches. Interestingly, we obtained the best results for Rn and H which have lower measurement uncertainties than all other target fluxes. For example, for H only one sensor, the sonic anemometer, is used while other measured fluxes (LE and NEE) two sensors, a sonic anemometer and a $CO_2/H_2O$ trace gas analyser, are needed. GPP and TER estimates are additionally subject to uncertainties of the flux partitioning methods, and this might explain why LE as a direct measurement was better predicted than GPP. Random uncertainties of the fluxes is likely not a big issue because averaging at daily and 8-day time steps (as in this study) greatly reduces the random error (Hollinger and Richardson, 2005). Instead we hypothesize that site specific systematic uncertainties in the eddy covariance estimations (e.g. due to presence of strong advection not corrected by the standard methods) could play an important role because ML methods were trained across sites distributing uncertainties among them. Systematic uncertainties could also reduce the sensitivity of the models on the small signal explaining the comparatively poor predictive skill of ML for anomalies of eddy co-variance fluxes. We also hypothesize that the general tendency of better predictability of energy fluxes compared to carbon fluxes is at least partly related to their differences in data quality. To test these hypothesis improved ways of detecting and characterizing systematic uncertainties in eddy co-variance data are needed.

Another common issue with eddy covariance data is the gaps generated by the data exclusion rules. Data exclusion strike strongly the nighttime period (primarily for the low turbulence condition) affecting the representativeness of the diurnal cycle, hence the quality of the averaged daily/eight days eddy-covariance fluxes, in particular $CO_2$,. To reduce the risk biased estimates half hourly data gaps are filled by models. In our study NEE data were gap filled using site-specific empirical relationships between meteorological data and net $CO_2$ ecosystem exchange (the MDS method, Reichstein et al., 2005) that produce small biases when short gaps were encountered (Moffat et al., 2007). This has a limited effect in this study as only a very small percentage of high quality gap filled data are used. We also minimize the bias in estimates of gross $CO_2$ fluxes (GPP and TER) by using two different partitioning methods which yield very consistent results.

### 4.4 Data quantity and representativeness

The mismatch between prediction and eddy-covariance estimation were also influenced by data representativeness. FLUXNET sites are not uniformly distributed over the globe and not all climates and PFTs are well represented. Very few sites are currently distributed in tropical forest, and data availability over the record is fragmented. Similarly, very few sites are located in the poorly predicted extreme environments, e.g., cold and dry climates. There was a clear pattern in our cross-validation results where more accurate predictions were obtained for the better represented vegetation types and climates (e.g. temperate and boreal forests). Therefore increasing the number of study sites in less represented environments (e.g. the tropics and in the extreme climates), could improve the prediction by ML and models in general (Papale et al. 2015).

Data representativeness has also a temporal aspect. For example, remote sensing data discarded due to low quality occurs preferentially in the cold or wet season—due to snow, ice or cloud cover—by comparison with other seasonal periods.

## 5. Conclusions

The ML methods presented and evaluated in this study have shown high capability to predict $CO_2$ and energy fluxes, in particular the across site variability and the mean seasonal cycle, with a general tendency of increasing performance in the following order: NEE, TER, GPP, LE, H, and Rn. The relatively poor performance for NEE likely resulted from factors that cannot be easily accounted for in ML-based approaches, such as legacies of site history (e.g., disturbances, management, age, and stocks). Future progress in this direction requires the reconstruction of the relevant management and disturbance history and the integration of information from forest inventories, high resolution satellites such as LANDSAT, and high resolution biomass data from radar and LIDAR. We found no substantial bias in the predictions of the ML approaches for most vegetation types or biomes. However, there is less consistency with observations for evergreen broadleaf forests, croplands, the tropics, and extreme climates. The growing number of eddy covariance sites, in particular new sites in poorly represented regions, will improve the predictive capacity of ML methods in the future. This is particularly so for the tropics which account for a disproportionate share of global terrestrial water and carbon fluxes (Beer et al., 2010).

The predictions for ecosystem fluxes across FLUXNET by different explanatory variable sets (RS vs RS+METEO) were highly consistent, indicating that the extracted patterns by the trained models were robust, realistic and not subject to overfitting. We recommend using the ensemble median estimate for generating global flux products as extrapolation beyond the FLUXNET-sampled conditions may generate large differences among methods.

The ML based models presented and extensively evaluated here form the basis of an extensive archive of global gridded flux products, which is currently under development. The thorough cross-validation experiment presented in this paper helps users understand the products' strengths and weaknesses. The overall high skill of the ML methods, the planned archival of their ensemble median, and the detailed analysis of their uncertainties will make this product a sought-after data stream to study the global land-atmosphere exchange of $CO_2$, water and energy.

## Appendix A: Median performance of the methods.

In table A1 we reported, for both setups, the median value of predictive capacity metrics (MEF, RMSE, and absolute value of BIAS) realized across singular ML and their standard deviation estimated as reported in Jung et al., (2009).

## Appendix B: Consistency among ML algorithms

Pair-wise $R^2$ values among model outputs (Table B1) were close to unity ($R^2 \geq 0.90$), regardless of experimental setup, with NEE showing a slightly lower value ($R^2 = 0.84$). Among corresponding model residuals (Table B1), $R^2$ values ranged from 0.79 (Rn) to 0.89 ($TER_L$). Comparing the same ML technique but using different experimental setups (Table B1, RS vs. RS+METEO) showed similarly high, albeit somewhat diminished level of consistency ($R^2$ range ranged from 0.71 to 0.80

for model residuals). These results highlighted that the ML methods were mapping between explanatory variables and target fluxes both reliably and robustly. Across the all three consistency checks there was also a tendency for better predicted fluxes (e.g., H) to exhibit higher pair-wise $R^2$ values than poorly predicted fluxes (e.g., NEE).

## Appendix C Median value of site-by-site performance per vegetation and climate type.

## Data availability

Fluxes and in situ measured meteorological data are obtained by La Thuile dataset and they are freely available at the FLUXNET website (http://fluxnet.fluxdata.org/data/la-thuile-dataset/). Reanalyzed ERA-Interim meteorological data are produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and they are freely available at the ERA-Interim | ECMWF website (http://apps.ecmwf.int/datasets/). Satellite MODIS subset products are freely available at the MODIS subset website (https://daac.ornl.gov/cgi-bin/MODIS/GR_col5_1/mod_viz.html). The data used for the cross-validation analysis (e.g. machine learning output and/or their median ensemble) are available on request from the first (and second) author.

## Acknowledgments

G. Tramontana was supported by the GEOCARBON EU FP7 project (GA 283080). D. Papale, M. Jung and M. Reichstein acknowledge funding from the EU FP7 project GEOCARBON (grant agreement no. 283080) and the EU H2020 BACI project (grant agreement No 640176). G. Camps-Valls wants to acknowledge the support by an ERC Consolidator Grant with grant agreement 647423 (SEDAL). K. Ichii was supported by Environment Research and Technology Development Funds (2-1401) from the Ministry of the Environment of Japan and the JAXA Global Change Observation Mission (GCOM) project (#115). C. R. Schwalm was supported by National Aeronautics and Space Administration (NASA) Grants #NNX12AP74G, #NNX10AG01A, and #NNX11AO08A. M. A. Arain thanks the support of Natural Sciences and Engineering Research Council (NSREC) of Canada. P. Serrano Ortiz was partially supported by GEISpain project (CGL2014-52838-C2-1-R), funded by the Spanish Ministry of Economy and Competitiveness and the European Union ERDF funds. S. Wolf acknowledges support from a Marie Curie International Outgoing Fellowship (European Commission, grant 300083). The FLUXCOM initiative is coordinated by M.Jung, Max Planck Institute for Biogeochemistry (Jena, Germany). This work used Eddy Covariance data acquired by the FLUXNET community and in particular by the following networks: AmeriFlux (U.S. Department of Energy, Biological and Environmental Research, Terrestrial Carbon Program (DE-FG02-04ER63917 and DE-FG02-04ER63911)), AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont,ChinaFlux, Fluxnet-Canada (supported by CFCAS, NSERC, BIOCAP, Environment Canada, and NRCan), GreenGrass, KoFlux, LBA, NECC, OzFlux, TCOS-Siberia, USCCC. We acknowledge the financial support to the eddy covariance data harmonization provided by CarboEuropeIP, FAO-GTOS-TCO, iLEAPS, Max Planck Institute for Biogeochemistry, National Science Foundation, University of Tuscia and US Department of Energy and the databasing and technical support from Berkeley Water Center, Lawrence Berkeley National Laboratory, Microsoft Research eScience, Oak Ridge National Laboratory, University of California - Berkeley, University of Virginia.

## 6. References

Amiro, B. D., Barr, A. G., Barr, J. G.,Black, T. A., Gracho, R., Brown, M., Chen, J., Clark, K. L., Davis, K. J., Desai, A. R., Dore, S., Engel, V., Fuentes, J. D., Goldstein, A. H., Goulden, M. L., Kolb, T. E., Lavigne, M. B., Law, B. E., Margolis, H. A., Martin, T., McCaughey, J. H., Misson, L., Montes, Helu, M., Noormets, A., Randerson, J. T., Starr, G. and Xiao, J.: Ecosystem carbon dioxide fluxes after disturbance in forests of North America. J Geophys Res-Biogeo, 115, G00K02, doi:10.1029/2010JG001390, 2010.

Alonso Fernández, J.R., Díaz-Muñiza, C., Garcia, Nieto, P.J., de Cos, Juez, F.J, Sánchez, Lasheras, F. and Roqueñíc, M.N.: Forecasting the cyanotoxins presence in fresh waters: A new model based on genetic algorithms combined with the MARS technique, Ecol Eng, 53, 68–78. doi:10.1016/j.ecoleng.2012.12.015, 2013.

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R. and Zhu Z.: Evaluating the land and ocean components of the global carbon cycle in the cmip5 earth system models, J Climate, 26, 6801–6843, doi: http://dx.doi.org/10.1175/JCLI-D-12-00417.1, 2013.

Aubinet, M., Vesala, T. and Papale, D.: Eddy Covariance: A Practical Guide to Measurement and Data Analysis , Springer, Dordrecht Heidelberg London New York, 460, 2012.

Baldocchi, D.: Breathing of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems, Aust J Bot, 56, 1–26, http://dx.doi.org/10.1071/BT07151, 2008.

Baldocchi, D.: Measuring fluxes of trace gases and energy between ecosystems and the atmosphere – the state and future of the eddy covariance method, Global Change Biol, 20, 3600-3609, DOI: 10.1111/gcb.12649, 2014.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, A., M., Baldocchi, D., Bonan, B., G., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, W. K., Roupsard, O., Veenendaal, E., Viovy, N., Woodward, I. F. and Papale, D.: Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate, Science, 329, 834-838, doi: 10.1126/science.1184984, 2010.

Bell, T. W., Menzer, O., Troyo-Diéquez, E. and Oechel, W.: Carbon dioxide exchange over multiple temporal scales in an arid shrub ecosystem near La Paz, Baja California Sur, Mexico, Global Change Biol, 18, 2570–2582, doi:10.1111/j.1365-2486.2012.02720.x., 2012.

Bonan, G. B., Lawrence, P. J., Oleson, K. W., Levis, S., Jung, M., Reichstein, M., Lawrence, D. M. and Swenson, S. C.: Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data, J Geophys Res-Biogeo, 116, G02014, doi:10.1029/2010JG001593, 2011.

Breiman, L.: Random Forests, Mach. Learn., 45 (1), 5–32, doi:10.1023/A:1010933404324, 2001.

Camps-Valls, G. and Bruzzone, L. (Eds.): Kernel Methods for Remote Sensing Data Analysis, John Wiley & Sons, Ltd, Chichester, UK, doi: 10.1002/9780470748992.fmatte, 2009.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B., M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q.J.R. Meteorol Soc, 137, 553–597, doi: 10.1002/qj.828, 2011.

Frank, D., Reichstein, M., Bahn, M., Thonicke, K., Frank, D., Mahecha, M. D., Smith, P., Van Der Velde, M., Vicca, S., Babst, F., Beer, C., Buchmann, N., Canadell, J. G., Ciais, P., Cramer, W., Ibrom, A., Miglietta, F., Poulter, B., Ramming, A., Seneviratne, S. I., Walz, A., Wattenbach, M., Zavala, M. A. and Zscheischler, J.: Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts, Global Change Biol, 21, 2861–2880, doi: 10.1111/gcb.12916, 2015.

Friedman, J. H.: Multivariate Adaptive Regression Splines, Ann. Statist., 19, 1-67, doi:10.1214/aos/1176347963, 1991.

Fröhlich, B., Rodner, E., Kemmler, M. and Denzler, J.: Large-scale gaussian process classification using random decision forests, S. Mach. Perc., 22 (1), 113–120, DOI 10.1007/s00138-012-0480-y, 2012.

Gao, B. C.: NDWI-A Normalized difference water index for remote sensing of vegetation liquid water from space, Remote Sens Environ, 58, 257-266, doi:10.1016/S0034-4257(96)00067-3, 1996.

Garnaud, C., Sushama, L. and Arorab, V. K.: The effect of driving climate data on the simulated terrestrial carbon pools and fluxes over North America, Int J Climatol, 34, 1098-1110, DOI: 10.1002/joc.3748, 2014.

Gupta, H. V., Kling, H., Yilmaz, K. K. and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J Hydrol, 20, 80-91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Hammerle, A., Haslwanter, A., Schmitt, M., Bahn, M., Tappeiner, U., Cernusca, A. and Wohlfahrt, G.: Eddy covariance measurements of carbon dioxide, latent and sensible energy fluxes above a meadow on a mountain slope. Bound-lay meteorol, 122(2), 397–416. http://doi.org/10.1007/s10546-006-9109-x, 2007

Haykin, S.: Neural Networks – A Comprehensive Foundation (2nd ed.)., Prentice Hall., 1999.

Ho, T. K.: The Random Subspace Method for Constructing Decision Forests, IEEE T Pattern Anal, 20 (8), 832–844, doi:10.1109/34.709601, 1998.

Hollinger, D. Y. and Richardson, A. D.: Uncertainty in eddy covariance measurements and its application to physiological models, Tree Physiol, 25 (7), 873-885, doi:10.1093/treephys/25.7.873, 2005.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X. and Ferreira, L.G.: Overview of the radiometric and biophysical performance of the MODIS vegetation indices, Remote Sens Environ, 83, 195–213, doi:10.1016/S0034-4257(02)00096-2, 2002.

Ichii, K., Wang, W., Hashimoto, H., Yang, F., Votava, P., Michaelis, A. R. and Nemani, R. R.: Refinement of rooting depths using satellite-based evapotranspiration seasonality for ecosystem modeling in California, Agr Forest Meteorol, 149, 1907-1918, doi:10.1016/j.agrformet.2009.06.019, 2009.

IPCC: Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, Pachauri, R.K and Reisinger, A. (eds.)]. IPCC, Geneva, Switzerland, 104, 2007.

Jung, M., and Zscheischler, J.: A Guided Hybrid Genetic Algorithm for Feature Selection with Expensive Cost Functions, Procedia Computer Science, 18, 2337-2346, doi: 10.1016/j.procs.2013.05.405, 2013.

Jung, M., Verstraete, M., Gobronz, N., Reichstein, M., Papale, D., Bondeau, A., Robustelli, M. and Pinty, R.: Diagnostic assessment of European gross primary production, Global Change Biol, 14, 2349–2364, doi: 10.1111/j.1365-2486.2008.01647.x, 2008.

Jung, M., Reichstein, M. and Bondeau, A.: Towards global empirical upscaling of FLUXNET Eddy Covariance observations: validation of a model tree ensemble approach using a biosphere model, Biogeosciences, 6, 2001-2013, doi:10.5194/bg-6-2001-2009, 2009.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A., Chen, J., de Jeu, R., Dolman, A. J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J., Kimball, J., Law, B. E., Montagnani, L., Mu, Q., Mueller, B., Oleson, K., Papale, D., Richardson, A. D., Roupsard, O., Running, S., Tomelleri, E., Viovy, N., Weber, U., Williams, C., Wood, E., Zaehle, S. and Zhang, K.: Recent decline in the global land evapotranspiration trend due to limited moisture supply, Nature Letter, 467, 951-953, doi:10.1038/nature09396, 2010.

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, Journal of geophys res-Biogeo, 116, G00J07, doi:10.1029/2010JG001566, 2011.

Justice, C. O., Townshend, J. R. G., Vermote, E. F., Masuoka, E., Wolfe, R. E., Saleous, N., Roy, D. P. and Morisette, J. T.: An overview of MODIS Land data processing and product status, Remote Sens Environ, 83, 3–15, doi:10.1016/S0034-4257(02)00084-6, 2002.

Kondo, M., Ichii, K., Takagi, H. and Sasakawa, M.: Comparison of the data-driven top-down and bottom-up global terrestrial CO2 exchanges: GOSAT CO2 inversion and empirical eddy flux upscaling, Journal of geophys res-Biogeo, 120, 1226–1245, doi:10.1002/2014JG002866, 2015.

Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneth, A., Barr, A., Stoy, P. and Wohlfahrt, G.: Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation, Global Change Biol, 16, 187-208, doi:10.1111/j.1365-2486.2009.02041.x, 2010.

Myneni, R.B., Hoffman, S., Knyazikhin, Y., Privette, J.L., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., Smith, G.R., Lotsch, A., Friedl, M., Morisette, J.T., Votava, P., Nemani, R.R., and Running, S.W.: Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data, Remote Sens Environ, 83, 214–231, doi:10.1016/S0034-4257(02)00074-3, 2002.

Nash, J. E. and Sutcliffe J. V.: River flow forecasting through conceptual models part I: A discussion of principles, Journal Hydrol, 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Papale, D. and Valentini, R.: A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization, Global Change Biol, 9, 525–535, doi: 10.1046/j.1365-2486.2003.00609.x, 2003.

Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B., Rambal, S., Valentini, R., Vesala, T. and Yakir, D.: Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: Algorithms and uncertainty estimation, Biogeosciences, 3, 571–583, doi:10.5194/bg-3-571-2006, 2006.

Papale, D., Black, T. A., Carvalhais, N., Cescatti, A., Chen, J., Jung, M., Kiely, G., Lasslop, G., Mahecha, D. M., Margolis, H., Merbold, L., Montagnani, L., Moors, E., Olesen, J. E., Reichstein, M., Tramontana, G., van Gorsel, E., Wohlfahrt, G. and Ráduly, B.: Effect of spatial sampling from European flux towers for estimating carbon and water fluxes with artificial neural networks, Journal of geophys res-Biogeo, 120, 1941-1957,doi: 10.1002/2015JG00299, 2015.

Paruelo, J. M., Piñero, G., Oyonarte, C., Alcaraz, D., Cabello, J. and Escribano, P.: Temporal and spatial patterns of ecosystem functioning in protected arid areas in southeastern Spain, Appl Veg Sci, 8, 93–102, doi: http://dx.doi.org/10.1658/1402-2001(2005)008[0093:TASPOE]2.0.CO;2, 2005.

Piao, S., Sitch, S., Ciais, P., Friedlingstein, P., Peylin, P., Wang, X., Ahlström, A., Anav, A., Canadell, J. G., Cong, N., Huntingford, C., Jung, M., Levis, S., Levy, P. E., Li, J., Lin, X., Lomas, M. R., Lu, M., Luo, Y., Ma, Y., Myneni, R. B., Poulter, B., Sun, Z., Wang, T., Viovy, N., Zaehle, S. and Zeng, N.: Evaluation of terrestrial carbon cycle models for their response to climate variability and to CO2 trends, Glob Change Biol., 19, 2117–2132, doi:10.1111/gcb.12187, 2013.

Rasmussen C. E. and Williams C. K. I.: Gaussian Processes for Machine Learning, the MIT Press, ISBN 026218253X, 2006.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grünwald, T., Havránková, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J.-M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D., and Valentini R.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm, Global Change Biol, 11, 1424–1439, doi:10.1111/j.1365-2486.2005.001002.x, 2005.

Reich, P. B.: The carbon dioxide exchange, Science, 329, 774-775, http://dx.doi.org/10.1126/science.1194353, 2010.

Schwalm, C. R., Williams, C. A., Schaefer, K., Arneth, A., Bonal, D., Buchmann, N., Chen, J., Lindroth, A., Luyssaert, S., Reichstein, M. and Richardson, A. D.: Assimilation exceeds respiration sensitivity to drought: A FLUXNET synthesis. Global Change Biol, 16 (2), 657-670, doi: 10.1111/j.1365-2486.2009.01991.x, 2010.

Schwalm, C. R., Williams, C. A., Schaefer, K., Baldocchi, D., Black, T. A., Goldstein, A. H., Law, B. E., Oechel, W. C., Paw U, K. T., Scott, R. L.: Reduction in carbon uptake during turn of the century drought in western North America. Nature Geoscience, 5 (8), 551-556, 10.1038/NGEO1529, 2012.Schaaf, C. B., Gao, F., Strahler, A. H., Lucht, W., Li, X., Tsang, T., Strugnell, N. C., Zhang, X., Jin, Y., Muller, J-P., Lewis, P., Barnsley, M., Hobson, P., Disney, M., Roberts, G., Dunderdale, M., Doll, C., d'Entremont, R. P., Hu, B., Liang, S., Privette, J. L. and Roy, D.: First operational BRDF, albedo nadir reflectance products from MODIS, Remote Sens Environ, 83, 135–148, doi:10.1016/S0034-4257(02)00091-3, 2002.

Shawe-Taylor, J. and Cristianini, N.: Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

Shirmohammadi, R., Ghorbani, B., Hamedi, M., Hamedi, M. H., and Romeo, L. M.: Optimization of mixed refrigerant systems in low temperature applications by means of group method of data handling (GMDH), Journal of Natural Gas Science and Engineering, 26, 303-312, doi:10.1016/j.jngse.2015.06.028, 2015.

Sims, D. A., Rahman, A. F., Cordova, V. D., El-Masri, B. Z., Baldocchi, D. D., Bolstad, P. V., Flanagan, L. B., Goldstein, A. H., Hollinger, D. Y., Misson, L., Monson, R. K., Oechel, W. C., Schmid, H. P., Wofsy, S. C. and Xu, L.: A new model of gross primary productivity for North American ecosystems based solely on the enhanced vegetation index and land surface temperature from MODIS, Remote Sens Environ, 12, 1633–1646, doi:10.1016/j.rse.2007.08.004, 2008.

Thornton, P. E., Law, B. E., Gholz, H. L., Clark, K. L., Falge, E., Ellsworth, D. S., Goldstein, A. H., Monson, R. K., Hollinger, D., Falk, M., Chen, J., Sparks, J. P.: Modeling and measuring the effects of disturbance history and climate on carbon and water budgets in evergreen needleleaf forests, Agr Forest Meteorol, 113, 185-222, doi:10.1016/S0168-1923(02)00108-9, 2002.

Tramontana, G., Ichii, K., Camps-Valls, G., Tomelleri, E. and Papale, D.: Uncertainty analysis of gross primary production upscaling using Random Forests, remote sensing and eddy covariance data, Remote Sens Environ, 168, 360–373, doi:10.1016/j.rse.2015.07.015, 2015.

Ungaro, F., Calzolari, C. and Busoni, E: Development of pedotransfer functions using a group method of data handling for the soil of the Pianura Padano–Veneta region of North Italy: water retention properties, Geoderma, 124, 293–317, doi:10.1016/j.geoderma.2004.05.007, 2005.

Valentini, R., Arneth, A., Bombelli, A., Castaldi, S., Cazzolla Gatti, R., Chevallier, F., Ciais, P., Grieco, E., Hartmann, J., Henry, M., Houghton, R. A., Jung, M., Kutsch, W. L., Malhi, Y., Mayorga, E., Merbold, L., Murray-Tortarolo, G., Papale, D., Peylin, P., Poulter, B., Raymond, P. A., Santini, M., Sitch, S., Vaglio Laurin, G., van der Werf, G. R., Williams, C. A. and Scholes, R. J.: A full greenhouse gases budget of Africa: synthesis, uncertainties, and vulnerabilities, Biogeosciences, 11, 381-407, doi:10.5194/bg-11-381-2014., 2014.

Vapnik, V., Golowich, S. and Smola, A.: Support vector method for function approximation, regression estimation, and signal processing, Adv Neur In, 9, 281–287, 1997.

Verrelst, J., Camps-Valls, G., Muñoz, J., Rivera, J. P., Veroustraete, F., Clevers, J. P. G. W. and Moreno, J.: Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties – A review, ISPRS J Photogramm, doi:10.1016/j.isprsjprs.2015.05.005, 108, 273-290, 2015

Wan, Z., Zhang, Y., Zhang, Q. and Li, Z. L.: Validation of the land-surface temperature products retrieved from Terra Moderate Resolution Imaging Spectroradiometer data, Remote Sens Environ, 83, 163−180, doi: 10.1016/S0034-4257(02)00093-7, 2002.

Xiao, J., Zhuang, Q., Baldocchi, D. D., Law, B. E., Richardson, A. D., Chen, J., Oren, R., Starr, G., Noormets, A., Ma, S., Verma, S. B., Wharton, S., Wofsy, S. C., Bolstad, P. V., Burns, S. P., Cook, D. R., Curtis, P. S., Drake, B. G., Falk, M., Fischer, M. L., Foster, D. R., Gu, L., Hadley, J. L., Hollinger, D. Y., Katul, G. G., Litvak, M., Martin, T. A., Matamala, R., McNulty, S., Meyers, T. P., Monson, R. K., Munger, J. W., Oechel, W. C., Paw U, K. T., Schmid, H. P., Scott, R. L., Sun, G., Suyker, A. E., Torn, M. S.: Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data, Agr Forest Meteorol, 148, 1827-1847, doi:10.1016/j.agrformet.2008.06.015, 2008.

Xiao, J., Zhuang, Q., Law, B. E., Chen, J., Baldocchi, D. D., Cook, D. R., Oren, R., Richardson, A. D., Wharton, S., Ma, S., Martin, T. A., Verma, S. B., Suyker, A. E., Scott, R. L., Monson, R. K., Litvak, M., Hollinger, D. Y., Sun, G., Davis, K. J., Bolstad, P. V., Burns, S. P., Curtis, P. S., Drake, B. G., Falk, M., Fischer, M. L., Foster, D. R., Gu, L., Hadley, J. L., Katul, G. G., Matamala, R., McNulty, S., Meyers, T., P., Munger, J. W., Noormets, A., Oechel, W. C., Paw, K. T., Schmid, H. P., Starr, G. Torn, M. S. and Wofsy, S. C.: A continuous measure of gross primary production for the conterminous United States derived from MODIS and AmeriFlux data, Remote Sens Environ, 114, 576–591, doi: 10.1016/j.rse.2009.10.013, 2010.

Xiao, X., Boles, S., Liu, J. Y., Zhuang, D. F. and Liu, M. L.: Characterization of forest types in Northeastern China, using multi-temporal SPOT-4 VEGETATION sensor data, Remote Sens Environ, 82, 335–348, doi:10.1016/S0034-4257(02)00051-2, 2002.

Yang, F., White, M. A., Michaelis, A. R., Ichii, K., Hashimoto, H., Votava, P., Zhu, A. X. and Nemani, R. R.: Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support Vector machine, IEEE T. Geosci Remote, 44, 3452-3461, doi: 10.1109/TGRS.2006.876297, 2006.

Yang, F., Ichii, K., White, M. A., Hashimoto, H., Michaelis, A. R., Votava, P., Zhu, A-X., Huete, A., Running, S. W., and Nemani, R. R.: Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach, Remote Sens Environ, 110, 109–122, doi:10.1016/j.rse.2007.02.016, 2007.

Yebra, M., Van Dijk, A. I. J. M., Leuning, R. and Guerschman, J. P.: Global vegetation gross primary production estimation using satellite-derived light-use efficiency and canopy conductance, Remote Sens Environ, 163, 206–216, doi:10.1016/j.rse.2015.03.016, 2015.

Yuan, W., Liu, S., Yu, G., Bonnefond, J-M., Chen, J., Davis, K., Desai, A. R., Goldstein, A. H., Gianelle, D., Rossi, F., Suyker, A. E. and Verma, S. B.: Global estimates of evapotranspiration and gross primary production based on MODIS and global meteorology data, Remote Sens Environ, 114, 1416–1431, doi:10.1016/j.rse.2010.01.022, 2010.

Zhao, Y., Ciais, P., Pylin, P., Viovy, N., Longdoz, B., Bonnefond, J. M., Rambal, S., Klumpp, K., Olioso, A., Cellier, P., Maigna, F., Eglin, T. and Calvet, J. C.: How errors on meteorological variables impact simulated ecosystem fluxes: a case study for six French sites, Biogeosciences, 9, 2537-2564, doi:10.5194/bg-9-2537-2012, 2012.

**Table 1.** Distribution of flux tower sites across plant functional types (PFT) and climate zones.

| PFT | N° of sites | Climate zone | N° of sites |
|---|---|---|---|
| Evergreen needleleaf forest | 66 | Temperate | 111 |
| Grassland | 38 | Subtropical - Mediterranean | 47 |
| Cropland | 27 | Boreal | 34 |
| Deciduous broadleaf forest | 24 | Tropical | 14 |
| Evergreen broadleaf forest | 19 | Dry | 13 |
| Wetland | 17 | Artic | 5 |
| Shrubland | 12 | | |
| Mixed forest | 11 | | |
| Savannah | 10 | | |

**Table 2.** Selected predictors for both setup for $\text{CO}_2$ fluxes (GPP, TER and NEE) and energy fluxes (H, LE and Rn). List of acronyms: Enhanced Vegetation Index (EVI), fraction of Absorbed Photosynthetically Active Radiation (fAPAR), Leaf Area Index (LAI), daytime Land Surface Temperature ($\text{LST}_{\text{Day}}$) and nighttime Land Surface Temperature ($\text{LST}_{\text{Night}}$), Middle Infrared Reflectance (band 7) (MIR(1)), Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Plant Functional Type (PFT), incoming global Radiation (Rg), top of atmosphere potential Radiation (Rpot), Index of Water Availability (IWA), Relative humidity (Rh), Water Availability Index lower ($\text{WAI}_{\text{L}}$), and upper ($\text{WAI}_{\text{U}}$) (for details see supplementary material, Sect. S3), Mean Seasonal Cycle (MSC). The product between A and B (AxB) is shown as (A, B).

| Setup | Type of variability | $\text{CO}_2$fluxes | Energy fluxes |
|---|---|---|---|
| RS | Spatial | PFT | PFT |
| | | Amplitude of MSC of EVI | Maximum of MSC of (fAPAR, Rg) |
| | | Amplitude of MSC of MIR(1) | Minimum of MSC of Rg |
| | | Maximum of MSC of $\text{LST}_{\text{Day}}$ | |
| | Spatial & Seasonal | MSC LAI | MSC of (EVI, $\text{LST}_{\text{Day}}$) |
| | | | Rpot |
| | Spatial, Seasonal & Interannual | NDWI | Rg |
| | | $\text{LST}_{\text{Day}}$ | $\text{LST}_{\text{Day}}$ |
| | | $\text{LST}_{\text{Night}}$ | Anomalies of $\text{LST}_{\text{Night}}$ |
| | | (NDVI, Rg) | Anomalies of (EVI, $\text{LST}_{\text{Day}}$) |
| RS+METEO | Spatial | PFT | PFT |
| | | Amplitude of MSC of NDVI | Maximum of MSC of $\text{WAI}_{\text{U}}$ |
| | | Amplitude of MSC of band 4 BRDF reflectance(2) | Mean of MSC of band 6 BRDF reflectance(2) |
| | | Minimum of MSC of NDWI | Max of MSC of (fPAR, Rg) |
| | | Amplitude of MSC of $\text{WAI}_{\text{L}}$ | |
| | Spatial & Seasonal | MSC of $\text{LST}_{\text{Night}}$ | Rpot |
| | | MSC of (fPAR, $\text{LST}_{\text{Day}}$) | MSC of NDWI |
| | | MSC of (EVI, Rpot) | MSC of $\text{LST}_{\text{Night}}$ |
| | | | MSC of (EVI, Rg) |
| | Spatial & Seasonal & Interannual | Tair | Rain |
| | | (Rg, MSC of NDVI) | Rg |
| | | $\text{WAI}_{\text{L}}$ | Rh |
| | | | (MSC of NDVI, Rg, IWA) |

(1)derived from the MOD13 product; (2) derived from MCD43 product.

**Table** 3.Statistics of the accuracy of predictions of $CO_2$ and energy fluxes made by the ensemble median estimate based on RS and RS+METEO. For RMSE and BIAS, the reference units were gCm$^{-2}$d$^{-1}$ and MJm$^{-2}$d$^{-1}$ for $CO_2$ fluxes (GPP, TER and NEE) and energy fluxes (H, LE and Rn) respectively.

| Flux | RS | | | | | RS+METEO | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MEF | RMSE | ρ | ROV | BIAS | MEF | RMSE | ρ | ROV | BIAS |
| $GPP_R$ | 0.71 | 1.56 | 0.85 | 0.69 | -0.02 | 0.70 | 1.59 | 0.84 | 0.73 | 0.09 |
| $GPP_L$ | 0.71 | 1.53 | 0.84 | 0.68 | -0.02 | 0.71 | 1.54 | 0.84 | 0.74 | 0.09 |
| $TER_R$ | 0.64 | 1.14 | 0.80 | 0.61 | -0.01 | 0.64 | 1.15 | 0.80 | 0.69 | 0.09 |
| $TER_L$ | 0.60 | 1.18 | 0.77 | 0.56 | -0.01 | 0.63 | 1.14 | 0.79 | 0.66 | 0.08 |
| NEE | 0.46 | 1.24 | 0.68 | 0.39 | 0.04 | 0.43 | 1.28 | 0.65 | 0.40 | -0.02 |
| H | 0.79 | 1.36 | 0.89 | 0.71 | -0.02 | 0.79 | 1.37 | 0.89 | 0.75 | 0.02 |
| LE | 0.76 | 1.37 | 0.87 | 0.71 | -0.07 | 0.75 | 1.39 | 0.87 | 0.73 | -0.01 |
| Rn | 0.92 | 1.51 | 0.96 | 0.90 | -0.01 | 0.91 | 1.55 | 0.96 | 0.93 | 0.08 |

**Table 4**: R$^2$ and RMSE for the comparison across sites, mean seasonal cycle and anomalies. The last two columns showed the consistency between the median estimates of the two setups. For RMSE, the reference units were gCm$^{-2}$d$^{-1}$ and MJm$^{-2}$d$^{-1}$for CO$_2$ fluxes (GPP, TER and NEE) and energy fluxes (H, LE and Rn) respectively.

| Fluxes | RS *vs*. OBS | | RS+METEO *vs*. OBS | | RS *vs*. RS+METEO | |
|---|---|---|---|---|---|---|
| | R$^2$ | RMSE | R$^2$ | RMSE | R$^2$ | RMSE |
| Across site variability | | | | | | |
| GPP$_R$ | 0.78 | 0.80 | 0.77 | 0.82 | 0.95 | 0.34 |
| GPP$_L$ | 0.78 | 0.77 | 0.79 | 0.75 | 0.94 | 0.36 |
| TER$_R$ | 0.68 | 0.73 | 0.61 | 0.81 | 0.92 | 0.32 |
| TER$_L$ | 0.72 | 0.60 | 0.71 | 0.61 | 0.92 | 0.27 |
| NEE | 0.48 | 0.61 | 0.46 | 0.61 | 0.83 | 0.22 |
| H | 0.81 | 0.68 | 0.81 | 0.68 | 0.97 | 0.25 |
| LE | 0.79 | 0.74 | 0.75 | 0.80 | 0.93 | 0.33 |
| Rn | 0.80 | 0.93 | 0.79 | 0.96 | 0.96 | 0.38 |
| Mean seasonal cycle | | | | | | |
| GPP$_R$ | 0.76 | 1.03 | 0.77 | 1.02 | 0.93 | 0.48 |
| GPP$_L$ | 0.77 | 1.00 | 0.77 | 0.99 | 0.93 | 0.50 |
| TER$_R$ | 0.71 | 0.62 | 0.71 | 0.62 | 0.92 | 0.29 |
| TER$_L$ | 0.67 | 0.64 | 0.68 | 0.63 | 0.92 | 0.29 |
| NEE | 0.61 | 0.83 | 0.59 | 0.84 | 0.93 | 0.24 |
| H | 0.86 | 0.89 | 0.86 | 0.87 | 0.97 | 0.36 |
| LE | 0.87 | 0.79 | 0.87 | 0.79 | 0.95 | 0.45 |
| Rn | 0.98 | 0.74 | 0.98 | 0.74 | 0.99 | 0.43 |
| Anomalies | | | | | | |
| GPP$_R$ | 0.18 | 0.67 | 0.12 | 0.68 | 0.38 | 0.32 |
| GPP$_L$ | 0.16 | 0.67 | 0.11 | 0.68 | 0.37 | 0.31 |
| TER$_R$ | 0.14 | 0.48 | 0.15 | 0.48 | 0.36 | 0.17 |
| TER$_L$ | 0.10 | 0.58 | 0.13 | 0.57 | 0.35 | 0.18 |
| NEE | 0.13 | 0.56 | 0.13 | 0.55 | 0.43 | 0.20 |
| H | 0.43 | 0.81 | 0.41 | 0.81 | 0.77 | 0.34 |
| LE | 0.21 | 0.78 | 0.21 | 0.77 | 0.46 | 0.32 |
| Rn | 0.57 | 0.81 | 0.54 | 0.83 | 0.84 | 0.41 |

**Table A1**: Accuracy of $CO_2$ and energy fluxes predicted by machine learning method based on RS and RS+METEO setup. The median value and the standard deviation across methods (in brackets and estimated as reported in Jung et al., 2009) are shown.

| FLUXES | RS | | | RS+METEO | | |
|---|---|---|---|---|---|---|
| | MEF | RMSE | Abs BIAS | MEF | RMSE | Abs BIAS |
| GPP | 0.698 (±0.012) | 1.604 (±0.033) | 0.022 (±0.019) | 0.694 (±0.012) | 1.614 (±0.032) | 0.073 (±0.011) |
| $GPP_{HB}$ | 0.700 (±0.009) | 1.564 (±0.024) | 0.023 (±0.024) | 0.701 (±0.008) | 1.561 (±0.020) | 0.083 (±0.011) |
| TER | 0.612 (±0.022) | 1.183 (±0.033) | 0.026 (±0.025) | 0.623 (±0.005) | 1.166 (±0.008) | 0.089 (±0.033) |
| $TER_{HB}$ | 0.571 (±0.016) | 1.218 (±0.023) | 0.019 (±0.017) | 0.609 (±0.001) | 1.163 (±0.002) | 0.079 (±0.017) |
| NEE | 0.433 (±0.017) | 1.270 (±0.019) | 0.024 (±0.021) | 0.407 (±0.029) | 1.298 (±0.032) | 0.014 (±0.003) |
| H | 0.767 (±0.015) | 1.426 (±0.047) | 0.014 (±0.005) | 0.776 (±0.008) | 1.397 (±0.025) | 0.022 (±0.009) |
| LE | 0.739 (±0.015) | 1.418 (±0.042) | 0.052 (±0.046) | 0.734 (±0.003) | 1.434 (±0.009) | 0.023 (±0.008) |
| Rn | 0.909 (±0.009) | 1.589 (±0.082) | 0.030 (±0.025) | 0.908 (±0.008) | 1.600 (±0.070) | 0.073 (±0.015) |

**Table B1**: Mean values of the determination coefficient ($R^2$) by the pair-wise comparison of the models output and their residuals. We compared different ML and same drivers (RS and RS+METEO respectively) or the same ML and different drivers (RS vs RS+METEO). Numbers in brackets were the standard deviation of $R^2$. All correlations were statistically significant ($p < 0.001$).

| Fluxes | Correlation among models output | | | Correlation among models residuals | | |
|---|---|---|---|---|---|---|
| | RS | RS+METEO | RS vs RS+METEO | RS | RS+METEO | RS vs RS+METEO |
| $GPP_R$ | 0.95 (0.02) | 0.95 (0.02) | 0.89 (0.02) | 0.88 (0.04) | 0.87 (0.04) | 0.74 (0.04) |
| $GPP_L$ | 0.95 (0.02) | 0.94 (0.02) | 0.88 (0.02) | 0.88 (0.04) | 0.86 (0.04) | 0.72 (0.04) |
| $TER_R$ | 0.91 (0.03) | 0.94 (0.03) | 0.86 (0.04) | 0.86 (0.05) | 0.88 (0.05) | 0.75 (0.06) |
| $TER_L$ | 0.92 (0.03) | 0.93 (0.03) | 0.85 (0.03) | 0.89 (0.04) | 0.88 (0.05) | 0.77 (0.05) |
| NEE | 0.84 (0.06) | 0.84 (0.07) | 0.75 (0.08) | 0.88 (0.05) | 0.87 (0.06) | 0.80 (0.06) |
| H | 0.94 (0.02) | 0.96 (0.02) | 0.93 (0.03) | 0.80 (0.06) | 0.87 (0.05) | 0.76 (0.08) |
| LE | 0.94 (0.02) | 0.96 (0.01) | 0.90 (0.02) | 0.83 (0.05) | 0.88 (0.04) | 0.73 (0.04) |
| Rn | 0.98 (0.01) | 0.99 (0.00) | 0.97 (0.01) | 0.79 (0.08) | 0.86 (0.03) | 0.71 (0.12) |

**Table C1.** Median site-by-site $R^2$ and its standard deviation (in brackets and estimated as reported in Jung et al., 2009) for the $CO_2$ fluxes, per PFT and climate zones. List of acronyms: ENF, was evergreen needleleaf forest; DBF, deciduous broadleaf forest; EBF, Evergreen broadleaf forest; MF, mixed forest; SHR, shrubland; SAV, Savannah; GRA, Grassland; CRO, cropland; WET, Wetland; Trop, Tropical; SubTrop, subtropical; Dry, dry and arid land; Tmp, Temperate; TmpCont, Temperate-continental; Bor, boreal; Cold, cold and polar environment or covered by ice.

| CAT | $GPP_R$ | | $GPP_L$ | | $TER_R$ | | $TER_L$ | | NEE | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO |
| ENF | 0.87 (0.10) | 0.86 (0.10) | 0.85 (0.12) | 0.86 (0.12) | 0.81 (0.15) | 0.85 (0.11) | 0.75 (0.24) | 0.76 (0.20) | 0.50 (0.34) | 0.55 (0.30) |
| DBF | 0.89 (0.07) | 0.87 (0.09) | 0.87 (0.07) | 0.88 (0.08) | 0.81 (0.12) | 0.83 (0.13) | 0.76 (0.14) | 0.76 (0.14) | 0.72 (0.16) | 0.68 (0.17) |
| EBF | 0.50 (0.29) | 0.48 (0.20) | 0.48 (0.29) | 0.44 (0.28) | 0.34 (0.34) | 0.49 (0.35) | 0.15 (0.18) | 0.29 (0.20) | 0.26 (0.23) | 0.24 (0.26) |
| MF | 0.91 (0.06) | 0.95 (0.02) | 0.91 (0.03) | 0.95 (0.04) | 0.85 (0.10) | 0.90 (0.07) | 0.84 (0.10) | 0.86 (0.15) | 0.73 (0.10) | 0.75 (0.09) |
| SHR | 0.67 (0.30) | 0.71 (0.28) | 0.67 (0.36) | 0.72 (0.23) | 0.80 (0.13) | 0.78 (0.24) | 0.68 (0.18) | 0.66 (0.38) | 0.37 (0.38) | 0.41 (0.31) |
| SAV | 0.75 (0.13) | 0.70 (0.13) | 0.72 (0.05) | 0.67 (0.17) | 0.65 (0.07) | 0.72 (0.11) | 0.55 (0.16) | 0.61 (0.10) | 0.38 (0.20) | 0.34 (0.29) |
| GRA | 0.69 (0.27) | 0.62 (0.33) | 0.69 (0.25) | 0.60 (0.32) | 0.70 (0.25) | 0.73 (0.25) | 0.66 (0.20) | 0.72 (0.21) | 0.40 (0.29) | 0.36 (0.30) |
| CRO | 0.58 (0.41) | 0.44 (0.36) | 0.56 (0.41) | 0.45 (0.31) | 0.78 (0.17) | 0.76 (0.15) | 0.68 (0.22) | 0.65 (0.23) | 0.35 (0.46) | 0.33 (0.43) |
| WET | 0.87 (0.11) | 0.91 (0.07) | 0.85 (0.12) | 0.87 (0.09) | 0.78 (0.19) | 0.83 (0.14) | 0.65 (0.17) | 0.74 (0.20) | 0.64 (0.16) | 0.61 (0.24) |
| Trop | 0.32 (0.46) | 0.40 (0.39) | 0.63 (0.23) | 0.31 (0.32) | 0.25 (0.23) | 0.34 (0.47) | 0.11 (0.13) | 0.26 (0.14) | 0.28 (0.35) | 0.21 (0.30) |
| SubTrop | 0.64 (0.26) | 0.66 (0.28) | 0.65 (0.26) | 0.65 (0.24) | 0.64 (0.25) | 0.66 (0.26) | 0.52 (0.24) | 0.55 (0.28) | 0.39 (0.37) | 0.39 (0.26) |
| Dry | 0.47 (0.27) | 0.40 (0.33) | 0.50 (0.25) | 0.38 (0.30) | 0.62 (0.25) | 0.62 (0.38) | 0.55 (0.19) | 0.55 (0.39) | 0.21 (0.29) | 0.11 (0.14) |
| Tmp | 0.81 (0.19) | 0.74 (0.24) | 0.83 (0.14) | 0.78 (0.22) | 0.78 (0.13) | 0.77 (0.18) | 0.68 (0.20) | 0.72 (0.17) | 0.56 (0.28) | 0.47 (0.34) |
| TmpCont | 0.86 (0.09) | 0.82 (0.16) | 0.84 (0.11) | 0.80 (0.17) | 0.81 (0.12) | 0.78 (0.14) | 0.75 (0.17) | 0.76 (0.15) | 0.54 (0.42) | 0.53 (0.36) |
| Bor | 0.90 (0.07) | 0.90 (0.07) | 0.92 (0.06) | 0.89 (0.07) | 0.90 (0.05) | 0.91 (0.04) | 0.86 (0.08) | 0.89 (0.06) | 0.59 (0.31) | 0.59 (0.25) |
| Cold | 0.56 (0.57) | 0.50 (0.56) | 0.49 (0.62) | 0.46 (0.59) | 0.84 (0.20) | 0.86 (0.13) | 0.50 (0.38) | 0.55 (0.23) | 0.47 (0.56) | 0.45 (0.57) |

**Table C2.** Median site-by-site RMSE and its standard deviation (in brackets and estimated as reported in Jung et al., 2009) for the $CO_2$ fluxes per PFT and climate zones. List of acronyms: ENF, was evergreen needleleaf forest; DBF, deciduous broadleaf forest; EBF, Evergreen broadleaf forest; MF, mixed forest; SHR, shrubland; SAV, Savannah; GRA, Grassland; CRO, cropland; WET, Wetland; Trop, Tropical; SubTrop, subtropical; Dry, dry and arid land; Tmp, Temperate; TmpCont, Temperate-continental; Bor, boreal; Cold, cold and polar environment or covered by ice.

| CAT | $GPP_R$ ($gCm^{-2}d^{-}1$) | | $GPP_L$ ($gCm^{-2}d^{-}1$) | | $TER_R$ ($gCm^{-2}d^{-}1$) | | $TER_L$ ($gCm^{-2}d^{-}1$) | | NEE ($gCm^{-2}d^{-}1$) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO |
| ENF | 1.05 (0.60) | 1.12 (0.60) | 1.04 (0.59) | 1.14 (0.66) | 0.82 (0.50) | 0.80 (0.52) | 0.87 (0.60) | 0.91 (0.68) | 0.87 (0.51) | 0.86 (0.53) |
| DBF | 1.21 (0.78) | 1.35 (0.59) | 1.17 (0.68) | 1.36 (0.62) | 0.68 (0.26) | 0.76 (0.33) | 0.76 (0.33) | 0.93 (0.44) | 1.28 (0.39) | 1.28 (0.39) |
| EBF | 1.70 (0.55) | 1.64 (0.85) | 1.65 (0.70) | 1.46 (0.51) | 1.23 (0.69) | 1.48 (0.85) | 1.88 (1.23) | 1.71 (0.73) | 1.15 (0.48) | 1.15 (0.45) |
| MF | 0.87 (0.17) | 0.76 (0.45) | 0.89 (0.27) | 0.97 (0.56) | 0.65 (0.18) | 0.73 (0.42) | 0.79 (0.14) | 0.79 (0.18) | 0.91 (0.47) | 0.81 (0.29) |
| SHR | 0.73 (0.47) | 0.78 (0.46) | 0.69 (0.44) | 0.77 (0.37) | 0.50 (0.33) | 0.70 (0.41) | 0.50 (0.34) | 0.55 (0.36) | 0.57 (0.31) | 0.52 (0.15) |
| SAV | 0.83 (0.44) | 0.81 (0.18) | 0.87 (0.45) | 0.84 (0.18) | 0.80 (0.53) | 0.68 (0.41) | 0.86 (0.55) | 0.77 (0.38) | 0.71 (0.36) | 0.69 (0.31) |
| GRA | 1.22 (0.64) | 1.22 (0.60) | 1.18 (0.68) | 1.20 (0.62) | 1.00 (0.48) | 1.01 (0.54) | 0.99 (0.58) | 0.95 (0.52) | 0.76 (0.61) | 0.85 (0.49) |
| CRO | 1.69 (1.38) | 2.30 (1.02) | 1.57 (1.42) | 2.24 (1.10) | 0.87 (0.46) | 0.90 (0.57) | 0.80 (0.51) | 0.98 (0.57) | 1.42 (0.90) | 1.44 (0.70) |
| WET | 1.04 (0.95) | 0.93 (0.77) | 1.03 (0.96) | 0.78 (0.53) | 1.04 (0.87) | 0.98 (0.82) | 1.07 (0.51) | 1.02 (0.51) | 0.46 (0.19) | 0.64 (0.26) |
| Trop | 1.93 (0.46) | 1.74 (1.01) | 2.24 (0.62) | 1.56 (0.78) | 2.07 (0.69) | 1.55 (0.87) | 2.47 (0.74) | 2.05 (0.43) | 1.28 (0.29) | 1.17 (0.46) |
| SubTrop | 1.37 (0.55) | 1.40 (0.61) | 1.37 (0.56) | 1.38 (0.57) | 1.03 (0.46) | 1.00 (0.41) | 1.08 (0.36) | 1.11 (0.40) | 1.13 (0.63) | 1.15 (0.62) |
| Dry | 0.60 (0.24) | 0.78 (0.36) | 0.63 (0.16) | 0.74 (0.30) | 0.49 (0.10) | 0.54 (0.20) | 0.58 (0.26) | 0.67 (0.32) | 0.41 (0.13) | 0.46 (0.15) |
| Tmp | 1.73 (1.02) | 1.82 (0.99) | 1.73 (0.98) | 1.71 (1.03) | 1.09 (0.54) | 1.17 (0.67) | 1.24 (0.57) | 1.31 (0.59) | 1.43 (0.59) | 1.40 (0.61) |
| TmpCont | 1.01 (0.42) | 1.29 (0.59) | 1.00 (0.45) | 1.26 (0.57) | 0.71 (0.30) | 0.75 (0.38) | 0.74 (0.31) | 0.79 (0.34) | 0.95 (0.39) | 1.02 (0.43) |
| Bor | 0.66 (0.27) | 0.70 (0.36) | 0.66 (0.27) | 0.67 (0.33) | 0.48 (0.27) | 0.47 (0.27) | 0.48 (0.16) | 0.45 (0.21) | 0.50 (0.32) | 0.48 (0.22) |
| Cold | 0.44 (0.04) | 0.58 (0.42) | 0.51 (0.24) | 0.46 (0.32) | 0.41 (0.06) | 0.23 (0.06) | 0.57 (0.16) | 0.29 (0.12) | 0.51 (0.21) | 0.54 (0.35) |

**Table C3.** Median site-by-site absolute bias and its standard deviation (in brackets and estimated as reported in Jung et al., 2009) for the $CO_2$ fluxes per PFT and climate zones. List of acronyms: ENF, was evergreen needleleaf forest; DBF, deciduous broadleaf forest; EBF, Evergreen broadleaf forest; MF, mixed forest; SHR, shrubland; SAV, Savannah; GRA, Grassland; CRO, cropland; WET, Wetland; Trop, Tropical; SubTrop, subtropical; Dry, dry and arid land; Tmp, Temperate; TmpCont, Temperate-continental; Bor, boreal; Cold, cold and polar environment or covered by ice.

| CAT | GPP$_R$ (gCm$^{-2}$d$^{-}$1) | | GPP$_L$ (gCm$^{-2}$d$^{-}$1) | | TER$_R$ (gCm$^{-2}$d$^{-}$1) | | TER$_L$ (gCm$^{-2}$d$^{-}$1) | | NEE (gCm$^{-2}$d$^{-}$1) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO |
| ENF | 0.53 (0.46) | 0.54 (0.56) | 0.45 (0.42) | 0.48 (0.50) | 0.47 (0.47) | 0.50 (0.54) | 0.42 (0.40) | 0.41 (0.43) | 0.39 (0.44) | 0.32 (0.36) |
| DBF | 0.43 (0.38) | 0.56 (0.59) | 0.42 (0.36) | 0.50 (0.52) | 0.29 (0.32) | 0.35 (0.35) | 0.39 (0.33) | 0.42 (0.34) | 0.60 (0.28) | 0.55 (0.30) |
| EBF | 0.82 (0.91) | 0.77 (0.50) | 0.75 (0.81) | 0.76 (0.48) | 0.88 (0.98) | 0.84 (0.72) | 0.76 (0.81) | 0.93 (0.65) | 0.36 (0.45) | 0.46 (0.44) |
| MF | 0.47 (0.20) | 0.34 (0.38) | 0.38 (0.29) | 0.57 (0.29) | 0.39 (0.28) | 0.41 (0.13) | 0.37 (0.15) | 0.30 (0.35) | 0.34 (0.49) | 0.32 (0.36) |
| SHR | 0.38 (0.37) | 0.54 (0.49) | 0.38 (0.44) | 0.39 (0.47) | 0.36 (0.38) | 0.50 (0.43) | 0.31 (0.40) | 0.32 (0.23) | 0.27 (0.27) | 0.28 (0.24) |
| SAV | 0.42 (0.40) | 0.36 (0.21) | 0.35 (0.40) | 0.23 (0.15) | 0.43 (0.41) | 0.35 (0.23) | 0.42 (0.37) | 0.31 (0.10) | 0.23 (0.21) | 0.19 (0.10) |
| GRA | 0.60 (0.59) | 0.48 (0.49) | 0.60 (0.56) | 0.52 (0.55) | 0.38 (0.29) | 0.36 (0.37) | 0.44 (0.39) | 0.38 (0.38) | 0.17 (0.20) | 0.31 (0.31) |
| CRO | 0.47 (0.37) | 0.66 (0.44) | 0.36 (0.33) | 0.56 (0.47) | 0.29 (0.32) | 0.25 (0.22) | 0.29 (0.32) | 0.30 (0.29) | 0.41 (0.31) | 0.56 (0.55) |
| WET | 0.54 (0.64) | 0.28 (0.41) | 0.55 (0.62) | 0.29 (0.25) | 0.72 (0.35) | 0.48 (0.52) | 0.69 (0.29) | 0.50 (0.51) | 0.24 (0.19) | 0.30 (0.25) |
| Trop | 1.66 (1.31) | 0.67 (0.79) | 1.71 (1.23) | 0.77 (0.86) | 1.73 (0.88) | 1.16 (1.19) | 1.94 (0.81) | 1.21 (0.67) | 0.52 (0.57) | 0.38 (0.55) |
| SubTrop | 0.54 (0.45) | 0.55 (0.43) | 0.50 (0.38) | 0.52 (0.55) | 0.46 (0.44) | 0.53 (0.47) | 0.47 (0.35) | 0.42 (0.37) | 0.34 (0.44) | 0.37 (0.34) |
| Dry | 0.31 (0.20) | 0.33 (0.26) | 0.33 (0.38) | 0.36 (0.29) | 0.24 (0.21) | 0.32 (0.35) | 0.34 (0.21) | 0.43 (0.26) | 0.14 (0.08) | 0.22 (0.14) |
| Tmp | 0.72 (0.55) | 0.77 (0.71) | 0.66 (0.59) | 0.63 (0.56) | 0.50 (0.46) | 0.47 (0.50) | 0.51 (0.55) | 0.41 (0.45) | 0.46 (0.43) | 0.51 (0.41) |
| TmpCont | 0.45 (0.35) | 0.60 (0.52) | 0.39 (0.35) | 0.57 (0.47) | 0.37 (0.28) | 0.29 (0.25) | 0.37 (0.33) | 0.38 (0.37) | 0.35 (0.40) | 0.55 (0.55) |
| Bor | 0.36 (0.30) | 0.32 (0.34) | 0.32 (0.24) | 0.27 (0.31) | 0.32 (0.40) | 0.32 (0.33) | 0.31 (0.35) | 0.26 (0.32) | 0.27 (0.26) | 0.23 (0.26) |
| Cold | 0.07 (0.00) | 0.08 (0.09) | 0.08 (0.12) | 0.15 (0.06) | 0.34 (0.04) | 0.12 (0.06) | 0.34 (0.06) | 0.15 (0.01) | 0.37 (0.15) | 0.27 (0.27) |

**Table C4.** Median site-by-site R$^2$ and its standard deviation (in brackets and estimated as reported in Jung et al., 2009) for the energy fluxes per PFT and climate zones. List of acronyms: ENF, was evergreen needleleaf forest; DBF, deciduous broadleaf forest; EBF, Evergreen broadleaf forest; MF, mixed forest; SHR, shrubland; SAV, Savannah; GRA, Grassland; CRO, cropland; WET, Wetland; Trop, Tropical; SubTrop, subtropical; Dry, dry and arid land; Tmp, Temperate; TmpCont, Temperate-continental; Bor, boreal; Cold, cold and polar environment or covered by ice.

| CAT | H | | LE | | Rn | |
|---|---|---|---|---|---|---|
| | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO |
| ENF | 0.87 (0.10) | 0.86 (0.10) | 0.83 (0.10) | 0.84 (0.11) | 0.97 (0.02) | 0.97 (0.02) |
| DBF | 0.76 (0.18) | 0.74 (0.12) | 0.87 (0.05) | 0.87 (0.07) | 0.97 (0.01) | 0.97 (0.02) |
| EBF | 0.85 (0.13) | 0.82 (0.17) | 0.56 (0.30) | 0.52 (0.42) | 0.95 (0.05) | 0.96 (0.03) |
| MF | 0.85 (0.06) | 0.82 (0.10) | 0.91 (0.07) | 0.89 (0.06) | 0.97 (0.02) | 0.96 (0.02) |
| SHR | 0.83 (0.15) | 0.83 (0.17) | 0.73 (0.29) | 0.77 (0.23) | 0.98 (0.01) | 0.97 (0.01) |
| SAV | 0.74 (0.25) | 0.77 (0.26) | 0.85 (0.06) | 0.78 (0.11) | 0.86 (0.05) | 0.88 (0.10) |
| GRA | 0.72 (0.22) | 0.71 (0.22) | 0.85 (0.11) | 0.83 (0.16) | 0.96 (0.02) | 0.96 (0.02) |
| CRO | 0.70 (0.16) | 0.66 (0.18) | 0.79 (0.14) | 0.80 (0.14) | 0.97 (0.02) | 0.96 (0.02) |
| WET | 0.81 (0.06) | 0.78 (0.14) | 0.86 (0.10) | 0.84 (0.06) | 0.94 (0.02) | 0.92 (0.06) |
| Trop | 0.52 (0.18) | 0.60 (0.32) | 0.56 (0.38) | 0.50 (0.44) | 0.86 (0.14) | 0.89 (0.13) |
| SubTrop | 0.81 (0.18) | 0.82 (0.18) | 0.78 (0.13) | 0.80 (0.13) | 0.96 (0.03) | 0.96 (0.02) |
| Dry | 0.87 (0.07) | 0.86 (0.13) | 0.80 (0.07) | 0.79 (0.14) | 0.90 (0.06) | 0.93 (0.05) |
| Tmp | 0.78 (0.14) | 0.78 (0.13) | 0.86 (0.11) | 0.83 (0.13) | 0.97 (0.02) | 0.96 (0.02) |
| TmpCont | 0.72 (0.18) | 0.69 (0.18) | 0.83 (0.08) | 0.84 (0.09) | 0.97 (0.02) | 0.96 (0.02) |
| Bor | 0.90 (0.07) | 0.89 (0.08) | 0.92 (0.05) | 0.92 (0.03) | 0.98 (0.01) | 0.97 (0.02) |
| Cold | 0.83 (0.12) | 0.57 (0.19) | 0.83 (0.08) | 0.82 (0.07) | 0.94 (0.03) | 0.85 (0.13) |

**Table C5.** Median site-by-site RMSE and its standard deviation (in brackets and estimated as reported in Jung et al., 2009) for the energy fluxes per PFT and climate zones. List of acronyms: ENF, was evergreen needleleaf forest; DBF, deciduous broadleaf forest; EBF, Evergreen broadleaf forest; MF, mixed forest; SHR, shrubland; SAV, Savannah; GRA, Grassland; CRO, cropland; WET, Wetland; Trop, Tropical; SubTrop, subtropical; Dry, dry and arid land; Tmp, Temperate; TmpCont, Temperate-continental; Bor, boreal; Cold, cold and polar environment or covered by ice.

| CAT | H (MJm$^{-2}$d$^{-1}$) | | LE (MJm$^{-2}$d$^{-1}$) | | Rn (MJm$^{-2}$d$^{-1}$) | |
|---|---|---|---|---|---|---|
| | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO |
| ENF | 1.09 (0.25) | 1.16 (0.25) | 1.00 (0.56) | 1.02 (0.55) | 1.27 (0.68) | 1.26 (0.57) |
| DBF | 1.30 (0.43) | 1.31 (0.38) | 1.22 (0.26) | 1.14 (0.46) | 1.11 (0.42) | 1.24 (0.41) |
| EBF | 1.14 (0.60) | 1.29 (0.76) | 1.55 (0.39) | 1.60 (0.46) | 1.33 (0.43) | 1.14 (0.56) |
| MF | 1.18 (0.44) | 1.12 (0.42) | 0.82 (0.37) | 1.15 (0.54) | 1.14 (0.45) | 1.09 (0.43) |
| SHR | 1.21 (0.46) | 1.14 (0.28) | 1.12 (0.41) | 1.11 (0.56) | 1.37 (0.80) | 1.01 (0.43) |
| SAV | 1.23 (0.25) | 1.20 (0.22) | 1.32 (0.56) | 1.35 (0.30) | 1.10 (0.33) | 1.19 (0.60) |
| GRA | 1.14 (0.35) | 1.08 (0.47) | 1.09 (0.34) | 1.32 (0.54) | 1.48 (0.83) | 1.48 (0.90) |
| CRO | 1.24 (0.45) | 1.36 (0.33) | 1.51 (0.61) | 1.54 (0.35) | 1.24 (0.52) | 1.23 (0.26) |
| WET | 0.97 (0.36) | 1.22 (0.60) | 0.88 (0.13) | 0.90 (0.18) | 1.42 (0.51) | 1.65 (0.71) |
| Trop | 0.98 (0.51) | 1.19 (0.63) | 1.60 (0.52) | 1.62 (0.41) | 1.33 (0.73) | 1.03 (0.48) |
| SubTrop | 1.28 (0.38) | 1.32 (0.46) | 1.36 (0.62) | 1.36 (0.53) | 1.40 (0.40) | 1.33 (0.49) |
| Dry | 1.07 (0.24) | 1.05 (0.50) | 1.21 (0.33) | 1.27 (0.52) | 1.61 (0.75) | 2.02 (0.93) |
| Tmp | 1.18 (0.23) | 1.15 (0.33) | 1.18 (0.43) | 1.17 (0.49) | 1.10 (0.36) | 1.14 (0.47) |
| TmpCont | 1.30 (0.42) | 1.35 (0.37) | 1.25 (0.41) | 1.47 (0.37) | 1.17 (0.65) | 1.16 (0.54) |
| Bor | 0.98 (0.23) | 1.05 (0.26) | 0.70 (0.26) | 0.61 (0.20) | 0.88 (0.31) | 1.08 (0.50) |
| Cold | 1.03 (0.36) | 1.50 (0.55) | 1.00 (0.23) | 1.03 (0.45) | 1.47 (0.18) | 2.04 (0.19) |

**Table C6.** Median site-by-site absolute bias and its standard deviation (in brackets and estimated as reported in Jung et al., 2009) for the energy fluxes. List of acronyms: ENF, was evergreen needleleaf forest; DBF, deciduous broadleaf forest; EBF, Evergreen broadleaf forest; MF, mixed forest; SHR, shrubland; SAV, Savannah; GRA, Grassland; CRO, cropland; WET, Wetland; Trop, Tropical; SubTrop, subtropical; Dry, dry and arid land; Tmp, Temperate; TmpCont, Temperate-continental; Bor, boreal; Cold, cold and polar environment or covered by ice.

| CAT | H (MJm$^{-2}$d$^{-1}$) | | LE (MJm$^{-2}$d$^{-1}$) | | Rn (MJm$^{-2}$d$^{-1}$) | |
|---|---|---|---|---|---|---|
| | RS | RS+METEO | RS | RS+METEO | RS | RS+METEO |
| ENF | 0.44 (0.40) | 0.40 (0.33) | 0.42 (0.41) | 0.44 (0.49) | 0.78 (0.63) | 0.64 (0.61) |
| DBF | 0.60 (0.35) | 0.66 (0.35) | 0.57 (0.56) | 0.49 (0.50) | 0.38 (0.28) | 0.61 (0.49) |
| EBF | 0.38 (0.48) | 0.55 (0.46) | 0.97 (0.79) | 0.88 (0.70) | 0.88 (0.51) | 0.62 (0.43) |
| MF | 0.48 (0.40) | 0.26 (0.31) | 0.34 (0.40) | 0.64 (0.52) | 0.56 (0.45) | 0.56 (0.57) |
| SHR | 0.34 (0.43) | 0.47 (0.52) | 0.41 (0.41) | 0.50 (0.43) | 0.62 (0.76) | 0.44 (0.52) |
| SAV | 0.68 (0.35) | 0.56 (0.15) | 0.63 (0.80) | 0.40 (0.15) | 0.27 (0.22) | 0.63 (0.55) |
| GRA | 0.51 (0.39) | 0.40 (0.24) | 0.38 (0.38) | 0.57 (0.50) | 0.97 (0.81) | 0.81 (1.03) |
| CRO | 0.23 (0.21) | 0.24 (0.24) | 0.36 (0.38) | 0.41 (0.50) | 0.66 (0.58) | 0.68 (0.39) |
| WET | 0.47 (0.51) | 0.67 (0.37) | 0.54 (0.41) | 0.38 (0.21) | 0.34 (0.34) | 0.83 (0.78) |
| Trop | 0.37 (0.51) | 0.67 (0.47) | 0.97 (0.79) | 1.24 (0.82) | 0.94 (1.10) | 0.63 (0.60) |
| SubTrop | 0.58 (0.59) | 0.50 (0.39) | 0.62 (0.58) | 0.58 (0.56) | 0.83 (0.71) | 0.70 (0.55) |
| Dry | 0.68 (0.62) | 0.55 (0.56) | 0.21 (0.14) | 0.30 (0.26) | 1.06 (0.55) | 1.61 (0.91) |
| Tmp | 0.38 (0.23) | 0.34 (0.31) | 0.49 (0.46) | 0.56 (0.54) | 0.65 (0.49) | 0.68 (0.58) |
| TmpCont | 0.49 (0.41) | 0.40 (0.46) | 0.44 (0.51) | 0.53 (0.50) | 0.69 (0.72) | 0.61 (0.58) |
| Bor | 0.33 (0.32) | 0.38 (0.24) | 0.22 (0.16) | 0.23 (0.24) | 0.38 (0.27) | 0.50 (0.47) |
| Cold | 0.43 (0.46) | 0.71 (0.11) | 0.56 (0.31) | 0.39 (0.18) | 0.30 (0.29) | 0.86 (0.58) |

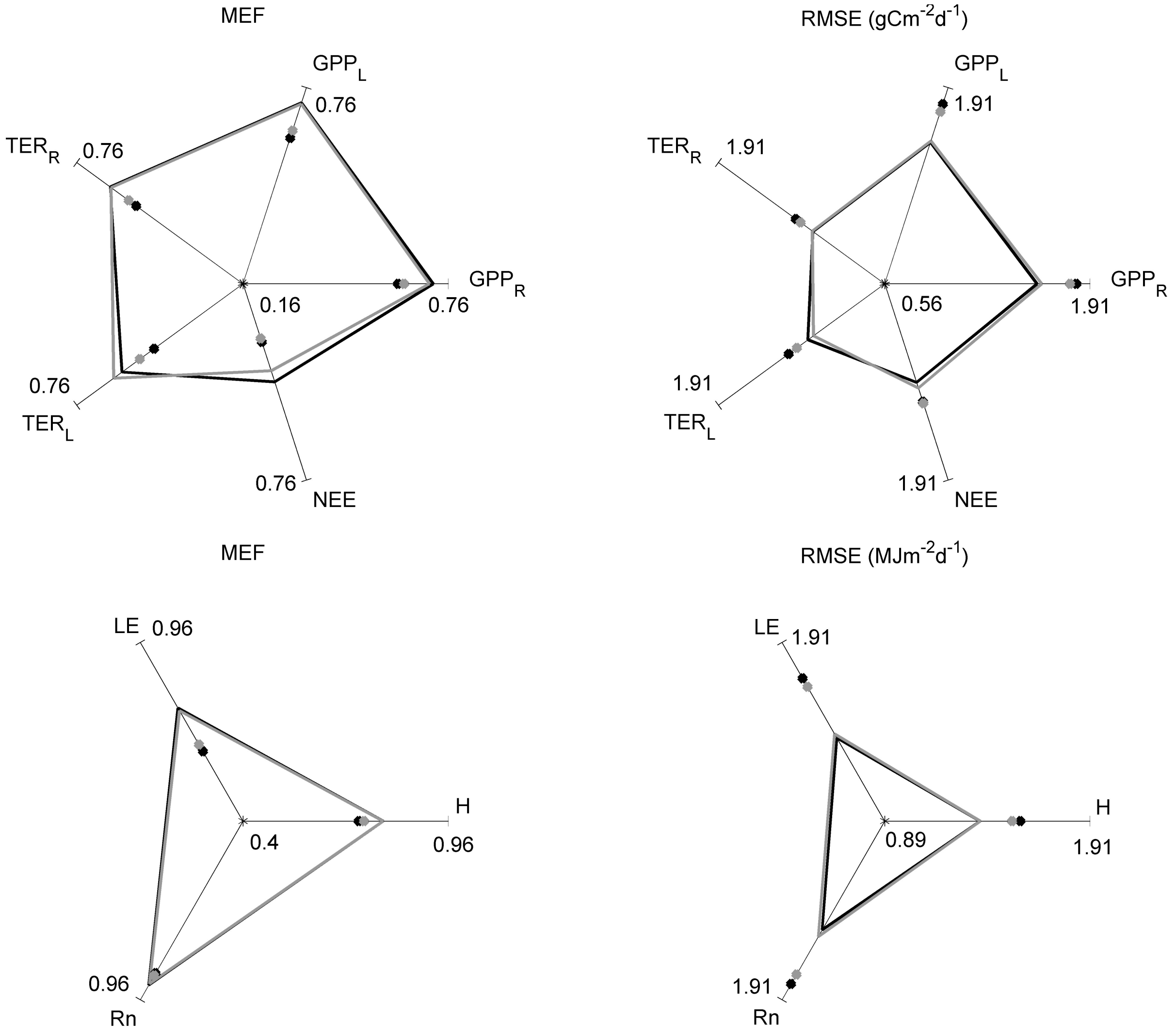


**Figure 1.** Spider plot of MEF (first column) and RMSE (second column) for $CO_2$ (first row) and energy fluxes (second row) showing the consistency of prediction made by RS (black line) and RS+METEO (grey lines) setups. The lines were the ensemble median estimate of ML; we also showed the performance of multiple regressions trained with RS (black point) and RS+METEO (gray points). $GPP_R$ and $GPP_L$ were respectively the gross primary production estimated following Reichstein et al. (2005) and Lasslop et al. (2010), $TER_R$ and $TER_L$ the total ecosystem respiration estimated following Reichstein et al. (2005) and Lasslop et al. (2010), NEE net ecosystem exchange, H the sensible heat, LE the latent heat and Rn the net radiation.

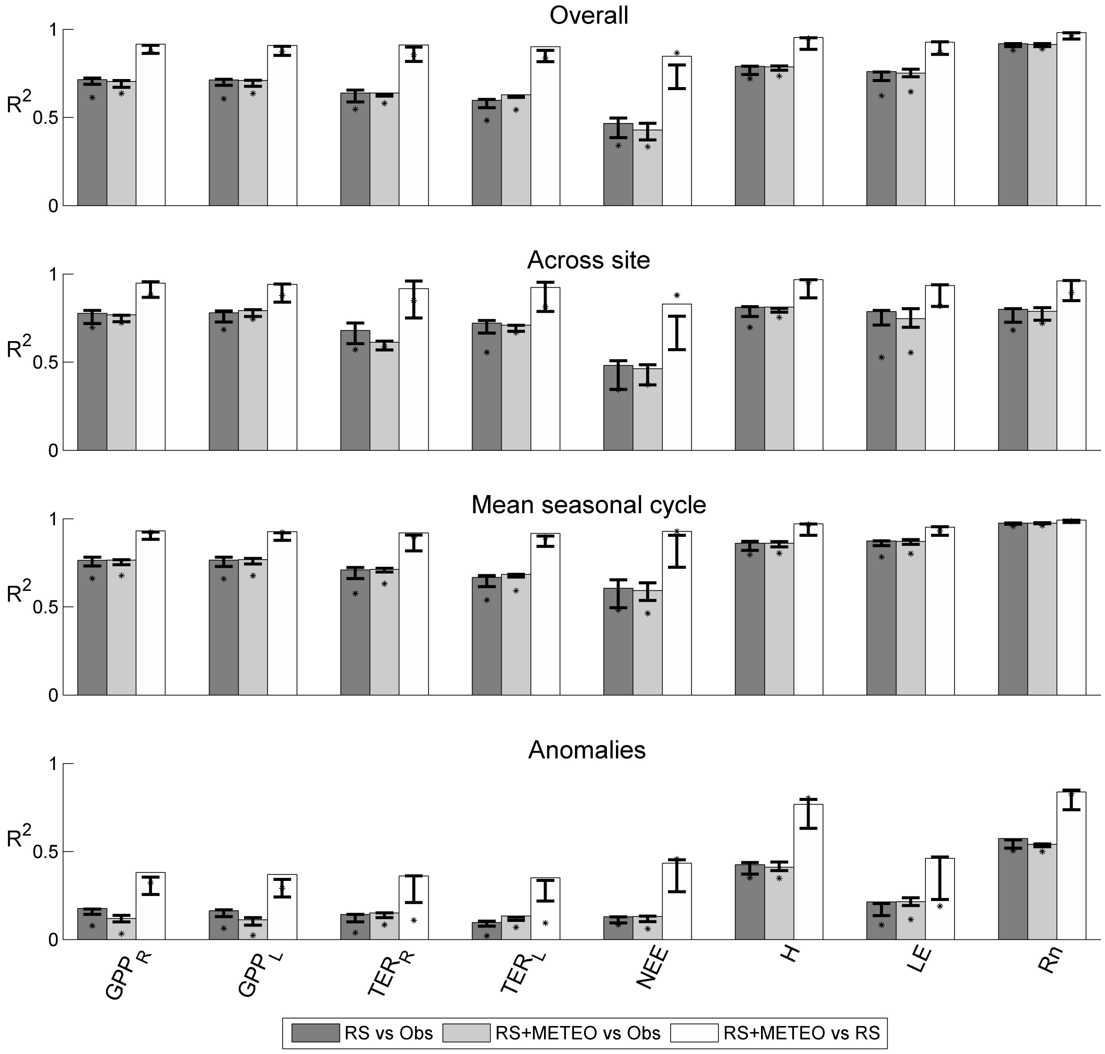


**Figure 2.** Coefficients of determination ($R^2$) from the comparison of overall time series, across-sites, mean seasonal cycle, and the anomalies, in particular: the determination coefficients between predictions by the ensemble median estimate of RS setup and observation (dark grey bars), between predictions by the ensemble median estimate of RS+METEO setup and observation (light grey bars), and between the two ensembles median estimate (white bars). Whiskers were the higher and lower $R^2$ when the comparisons were made among the singular ML. The comparison of output by the multiple regressions was also shown (black points). $GPP_R$ and $GPP_L$ were respectively the gross primary production estimated following Reichstein et al. (2005) and Lasslop et al. (2010), $TER_R$ and $TER_L$ the total ecosystem respiration estimated following Reichstein et al. (2005) and Lasslop et al. (2010), NEE net ecosystem exchange, H the sensible heat, LE the latent heat and Rn the net radiation.

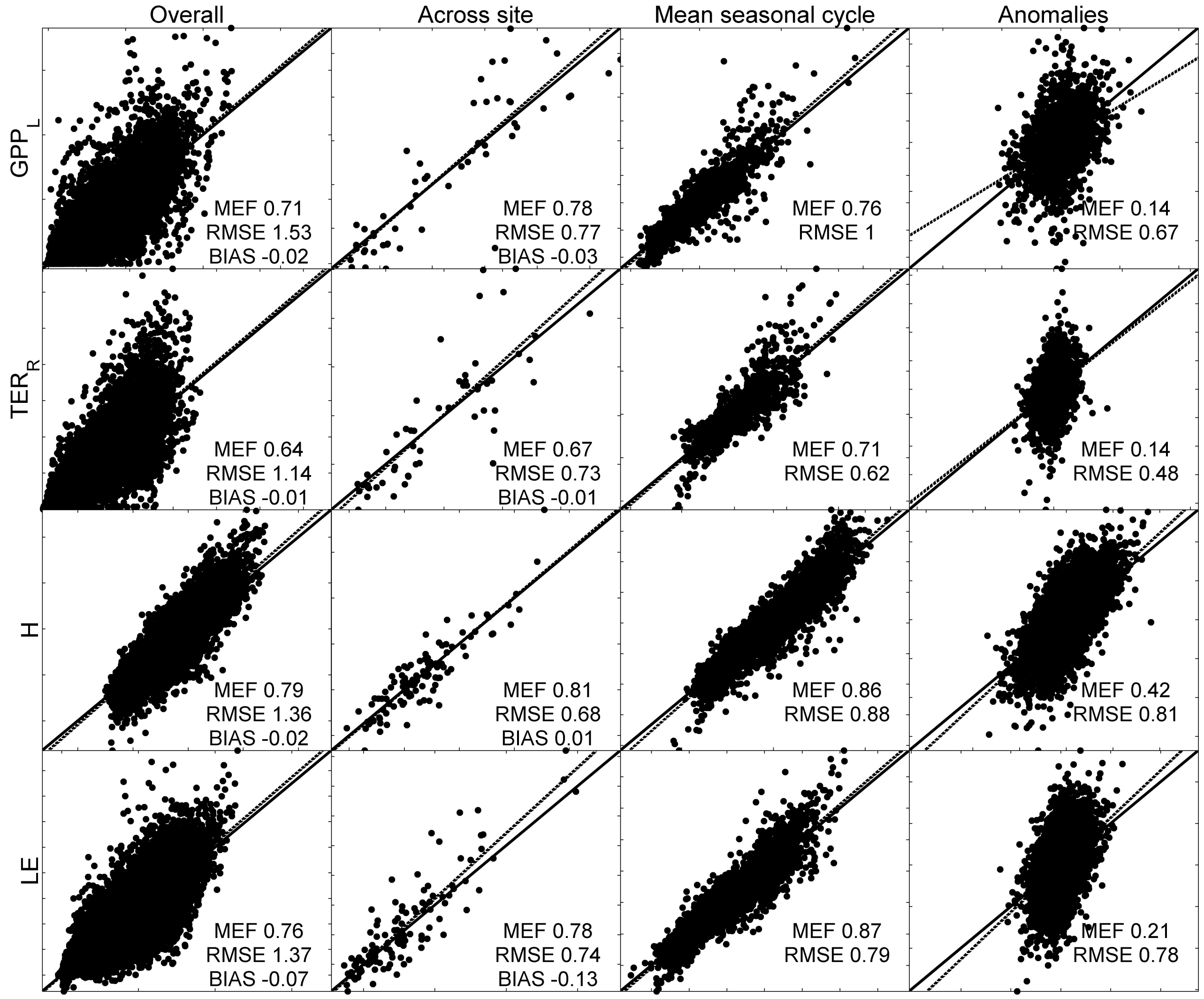


**Figure 3a**. Scatterplots of observed data by eddy covariance (y-axis) and the median ensemble of modeled fluxes by RS setup (x-axis). The panels from left to right were the 8-day predictions, the across sites variability, the mean seasonal cycle and the 8-day anomalies. The fluxes considered here were: the gross primary production estimated following Lasslop et al (2010), GPP$_L$ (first row); the total ecosystem respiration estimated following Reichstein et al., (2005), TER$_R$ (second row); the sensible heat, H (third row); the latent heat, LE (fourth row). The reference units were gCm$^{-2}$d$^{-1}$ and MJm$^{-2}$d$^{-1}$ for CO$_2$ fluxes (GPP$_L$ and TER$_R$) and energy fluxes (H and LE) respectively.

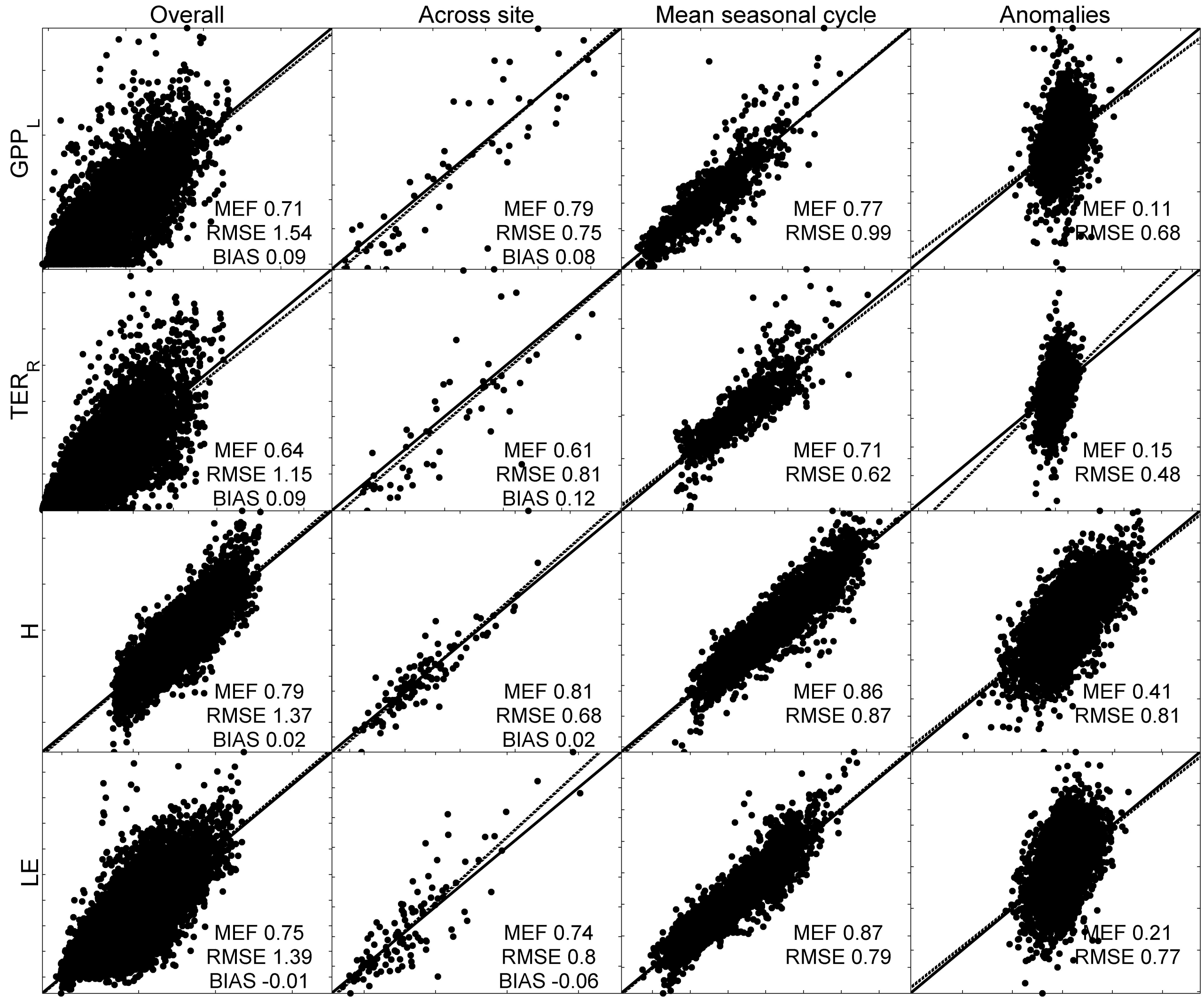
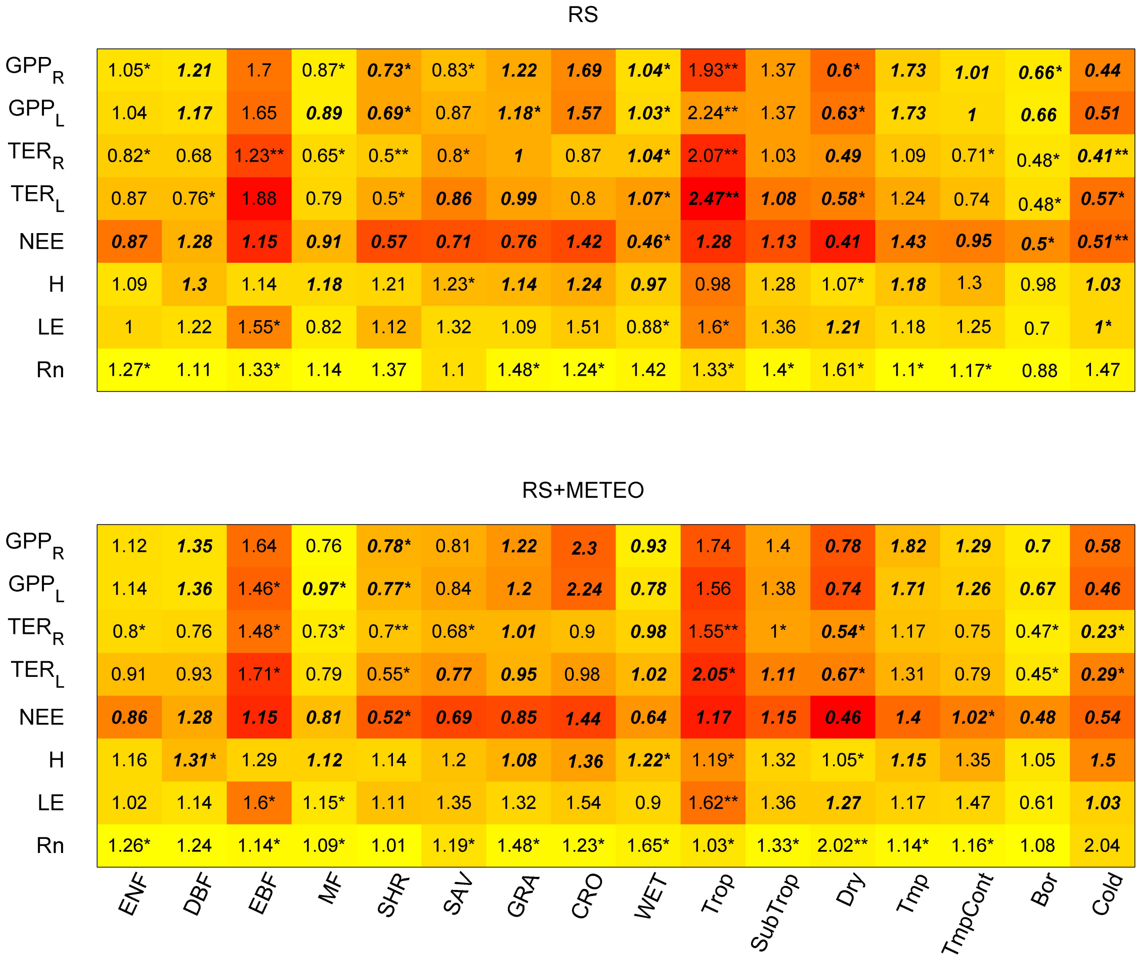


**Figure 3b**. As in Figure 3a but the predictions (x-axis) were obtained by the RS+METEO setup.



RS

| | ENF | DBF | EBF | MF | SHR | SAV | GRA | CRO | WET | Trop | SubTrop | Dry | Tmp | TmpCont | Bor | Cold |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $GPP_R$ | 1.05* | ***1.21*** | 1.7 | 0.87* | ***0.73**** | 0.83* | ***1.22*** | ***1.69*** | ***1.04**** | 1.93** | 1.37 | ***0.6**** | ***1.73*** | ***1.01*** | ***0.66**** | ***0.44*** |
| $GPP_L$ | 1.04 | ***1.17*** | 1.65 | ***0.89*** | ***0.69**** | 0.87 | ***1.18**** | ***1.57*** | ***1.03**** | 2.24** | 1.37 | ***0.63**** | ***1.73*** | ***1*** | ***0.66*** | ***0.51*** |
| $TER_R$ | 0.82* | 0.68 | 1.23** | 0.65* | 0.5** | 0.8* | ***1*** | 0.87 | ***1.04**** | 2.07** | 1.03 | ***0.49*** | 1.09 | 0.71* | 0.48* | ***0.41***** |
| $TER_L$ | 0.87 | 0.76* | 1.88 | 0.79 | 0.5* | ***0.86*** | ***0.99*** | 0.8 | ***1.07**** | ***2.47***** | ***1.08*** | ***0.58**** | 1.24 | 0.74 | 0.48* | ***0.57**** |
| NEE | ***0.87*** | ***1.28*** | ***1.15*** | ***0.91*** | ***0.57*** | ***0.71*** | ***0.76*** | ***1.42*** | ***0.46**** | ***1.28*** | ***1.13*** | ***0.41*** | ***1.43*** | ***0.95*** | ***0.5**** | ***0.51***** |
| H | 1.09 | ***1.3*** | 1.14 | ***1.18*** | 1.21 | 1.23* | ***1.14*** | ***1.24*** | ***0.97*** | 0.98 | 1.28 | 1.07* | ***1.18*** | 1.3 | 0.98 | ***1.03*** |
| LE | 1 | 1.22 | 1.55* | 0.82 | 1.12 | 1.32 | 1.09 | 1.51 | 0.88* | 1.6* | 1.36 | ***1.21*** | 1.18 | 1.25 | 0.7 | ***1**** |
| Rn | 1.27* | 1.11 | 1.33* | 1.14 | 1.37 | 1.1 | 1.48* | 1.24* | 1.42 | 1.33* | 1.4* | 1.61* | 1.1* | 1.17* | 0.88 | 1.47 |

RS+METEO

| | ENF | DBF | EBF | MF | SHR | SAV | GRA | CRO | WET | Trop | SubTrop | Dry | Tmp | TmpCont | Bor | Cold |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $GPP_R$ | 1.12 | ***1.35*** | 1.64 | 0.76 | ***0.78**** | 0.81 | ***1.22*** | ***2.3*** | ***0.93*** | 1.74 | 1.4 | ***0.78*** | ***1.82*** | ***1.29*** | ***0.7*** | ***0.58*** |
| $GPP_L$ | 1.14 | ***1.36*** | 1.46* | ***0.97**** | ***0.77**** | 0.84 | ***1.2*** | ***2.24*** | ***0.78*** | 1.56 | 1.38 | ***0.74*** | ***1.71*** | ***1.26*** | ***0.67*** | ***0.46*** |
| $TER_R$ | 0.8* | 0.76 | 1.48* | 0.73* | 0.7** | 0.68* | ***1.01*** | 0.9 | ***0.98*** | 1.55** | 1* | ***0.54**** | 1.17 | 0.75 | 0.47* | ***0.23**** |
| $TER_L$ | 0.91 | 0.93 | 1.71* | 0.79 | 0.55* | ***0.77*** | ***0.95*** | 0.98 | ***1.02*** | ***2.05**** | ***1.11*** | ***0.67**** | 1.31 | 0.79 | 0.45* | ***0.29**** |
| NEE | ***0.86*** | ***1.28*** | ***1.15*** | ***0.81*** | ***0.52**** | ***0.69*** | ***0.85*** | ***1.44*** | ***0.64*** | ***1.17*** | ***1.15*** | ***0.46*** | ***1.4*** | ***1.02**** | ***0.48*** | ***0.54*** |
| H | 1.16 | ***1.31**** | 1.29 | ***1.12*** | 1.14 | 1.2 | ***1.08*** | ***1.36*** | ***1.22**** | 1.19* | 1.32 | 1.05* | ***1.15*** | 1.35 | 1.05 | ***1.5*** |
| LE | 1.02 | 1.14 | 1.6* | 1.15* | 1.11 | 1.35 | 1.32 | 1.54 | 0.9 | 1.62** | 1.36 | ***1.27*** | 1.17 | 1.47 | 0.61 | ***1.03*** |
| Rn | 1.26* | 1.24 | 1.14* | 1.09* | 1.01 | 1.19* | 1.48* | 1.23* | 1.65* | 1.03* | 1.33* | 2.02** | 1.14* | 1.16* | 1.08 | 2.04 |

**Figure 4**. Performance of FLUXCOM median estimates per climate zone and plant functional type (PFT). The colored matrices show the median values of $R^2$ (red pixels for low $R^2$, yellow pixels for high $R^2$). Numbers indicate the RMSE (units of $CO_2$ fluxes are $gCm^{-2}d^{-1}$ and $MJm^{-2}d^{-1}$ in the case of energy fluxes). Oblique and bold fonts are used when the relative RMSE (normalized for the mean observed fluxes per PFT and climate zone) was greater than 0.5. The symbols '**' after RMSE were used when the weight of bias (estimated as the ratio between the square of median absolute bias and the MSE) was greater than 0.5, instead '*' symbols were used if the weight of bias was between 0.25 than 0.5. No symbols were used if the weight of bias is less than 0.25. List of acronyms: ENF, was evergreen needleleaf forest; DBF, deciduous broadleaf forest; EBF, Evergreen broadleaf forest; MF, mixed forest; SHR, shrubland; SAV, Savannah; GRA, Grassland; CRO, cropland; WET, Wetland; Trop, Tropical; SubTrop, subtropical; Dry, dry and arid land; Tmp, Temperate; TmpCont, Temperate-continental; Bor, boreal; Cold, cold and polar environment or covered by ice; $GPP_R$ and $GPP_L$ were respectively the gross primary production estimated following Reichstein et al. (2005) and Lasslop et al. (2010); $TER_R$ and $TER_L$ the total ecosystem respiration estimated following Reichstein et al. (2005) and Lasslop et al. (2010); NEE, net ecosystem exchange; H, sensible heat; LE, latent heat; Rn net radiation.