AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Biogeosciences

RESEARCH ARTICLE
10.1002/2015JG002997

Key Points:
- Uncertainty due to spatial sampling is evaluated using ANNs and FLUXNET data
- GPP and LE budgets and IAV are analyzed with different site networks
- The uncertainty in upscaling due to spatial sampling is highly heterogeneous

Supporting Information:
- Supporting Information S1
- Figure S1
- Figure S2
- Figure S3
- Figure S4
- Figure S5
- Table S1

Correspondence to:
D. Papale,
darpap@unitus.it

Citation:
Papale, D., et al. (2015), Effect of spatial sampling from European flux towers for estimating carbon and water fluxes with artificial neural networks, *J. Geophys. Res. Biogeosci.*, *120*, 1941–1957, doi:10.1002/2015JG002997.

Received 20 MAR 2015
Accepted 9 SEP 2015
Accepted article online 14 SEP 2015
Published online 9 OCT 2015

Effect of spatial sampling from European flux towers for estimating carbon and water fluxes with artificial neural networks

Dario Papale[1,2], T. Andrew Black[3], Nuno Carvalhais[4,5], Alessandro Cescatti[6], Jiquan Chen[7], Martin Jung[4], Gerard Kiely[8], Gitta Lasslop[9], Miguel D. Mahecha[4], Hank Margolis[10], Lutz Merbold[11], Leonardo Montagnani[12,13], Eddy Moors[14,15], Jørgen E. Olesen[16], Markus Reichstein[4], Gianluca Tramontana[1], Eva van Gorsel[17], Georg Wohlfahrt[18,19], and Botond Ráduly[20]

[1]Department for Innovation in Biological, Agro-Food and Forest Systems, University of Tuscia, Viterbo, Italy, [2]CzechGlobe, Global Change Research Centre AS CR, Brno, Czech Republic, [3]Faculty of Land and Food Systems, University of British Columbia, Vancouver, British Columbia, Canada, [4]Department Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany, [5]Departamento de Ciências e Engenharia do Ambiente, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisbon, Portugal, [6]Institute for Environment and Sustainability, Joint Research Centre, European Commission, Ispra, Italy, [7]CGCEO/Department of Geography, Michigan State University, East Lansing, Michigan, USA, [8]Civil and Environmental Engineering Department, and Environmental Research Institute, University College Cork, Cork, Ireland, [9]Fire in the Earth System, Land in the Earth System, Max Planck Institute for Meteorology, Hamburg, Germany, [10]Centre d'étude de la forêt, Université Laval, Quebec, Quebec, Canada, [11]Department of Environmental Systems Science, Institute of Agricultural Sciences, ETH Zurich, Zurich, Switzerland, [12]Forest Services, Autonomous Province of Bolzano, Bolzano, Italy, [13]Faculty of Sciences and Technology, Free University of Bolzano, Bolzano, Italy, [14]Alterra Wageningen UR, Wageningen, Netherlands, [15]VU University Amsterdam, Amsterdam, Netherlands, [16]Department of Agroecology, Aarhus University, Tjele, Denmark, [17]CSIRO Oceans and Atmosphere Flagship, Yarralumla, Australia, [18]Institute for Ecology, University of Innsbruck, Innsbruck, Austria, [19]European Academy Bolzano, Bolzano, Italy, [20]Department of Bioengineering, Sapientia Hungarian University of Transylvania, Miercurea Ciuc, Romania

Abstract Empirical modeling approaches are frequently used to upscale local eddy covariance observations of carbon, water, and energy fluxes to regional and global scales. The predictive capacity of such models largely depends on the data used for parameterization and identification of input-output relationships, while prediction for conditions outside the training domain is generally uncertain. In this work, artificial neural networks (ANNs) were used for the prediction of gross primary production (GPP) and latent heat flux (LE) on local and European scales with the aim to assess the portion of uncertainties in extrapolation due to sample selection. ANNs were found to be a useful tool for GPP and LE prediction, in particular for extrapolation in time (mean absolute error MAE for GPP between 0.53 and 1.56 gC m^{-2} d^{-1}). Extrapolation in space in similar climatic and vegetation conditions also gave good results (GPP MAE 0.7–1.41 gC m^{-2} d^{-1}), while extrapolation in areas with different seasonal cycles and controlling factors (e.g., the tropical regions) showed noticeably higher errors (GPP MAE 0.8–2.09 gC m^{-2} d^{-1}). The distribution and the number of sites used for ANN training had a remarkable effect on prediction uncertainty in both, regional GPP and LE budgets and their interannual variability. Results obtained show that for ANN upscaling for continents with relatively small networks of sites, the error due to the sampling can be large and needs to be considered and quantified. The analysis of the spatial variability of the uncertainty helped to identify the meteorological drivers driving the uncertainty.

1. Introduction

Rapid development of the FLUXNET network [*Baldocchi*, 2008; *Papale et al.,* 2012], along with its first example of liberal open-data use, has greatly promoted collaboration, enhancing our understanding of carbon and water fluxes in terrestrial ecosystems in recent years. With the eddy covariance method, fluxes of CO_2, water, and energy, since recently as well CH_4 and N_2O, are directly measured using fast response gas analyzers and a 3-D sonic anemometer following well-established procedures [*Aubinet et al.,* 2012]. There are more than 600 active sites globally, many of which are organized into regional and continental networks and contribute to global synthesis activities in the context of the FLUXNET initiative, like for example, the LaThuile 2007 data set (www.fluxdata.org) that has been used in this study. The scientific applications of these data are further

advanced by combining the information from remotely sensed products, such as Moderate Resolution Imaging Spectroradiometer-derived fraction of absorbed photosynthetically active radiation (FAPAR) and vegetation indices. These applications have stimulated an increasing interest in empirical, semiempirical, and process-oriented modeling approaches to estimate regional to global carbon, water, and energy budgets [e.g., *Beer et al.,* 2007, 2009, 2010; *Jung et al.,* 2010, 2011; *Verma et al.,* 2014; *Xiao et al.,* 2010; *Yang et al.,* 2007].

Ecosystem ecology traditionally relies on concepts derived from first principles that lead to parametric description of process responses to environmental drivers. This leads to the large family of process-based modeling approaches, where ecosystem functional dependencies are reproduced as much as possible based on the mechanistic knowledge of the processes involved [*Haxeltine and Prentice*, 1996; *Krinner et al.,* 2005; *Sitch et al.,* 2003], but often also requires some data for parameter calibration. In semiempirical approaches, a limited number of basic assumptions and parameters are predefined and then estimated from observations with optimization techniques, such as in radiation-use-efficiency models [e.g., *Lin et al.,* 2011; *Monteith,* 1972; *Running et al.,* 2000; *Xiao et al.,* 2004]. Empirically derived models follow a very different path by describing functional dependencies between input and output based on self-defined data adaptive functions, parameterized with a subset of observations during the training process. Hence, empirical models do not impose predefined relations between drivers and outputs of the simulated processes [*Hastie et al.,* 2001]. Clearly, obvious overlaps among these three groups (empirical, semiempirical, and process-oriented models) exist.

When empirical models are applied to estimate regional ecosystem fluxes of greenhouse gases or energy using gridded drivers (also known as empirical upscaling), observed fluxes and potential driving variables are required to parameterize the model or even derive the relationship between driving variables and the target variables—the fluxes. The best data available today for applying this approach are the fluxes measured at ecosystem scale using the eddy covariance technique, since they are the only direct continuous measurement integrated over a sufficiently large area. Different types of empirical models have been used for upscaling CO_2 and water vapor fluxes starting from eddy covariance data, remotely sensed vegetation indexes, and meteorological gridded variable [*Reichstein et al.,* 2007; *Vetter et al.,* 2008; *Beer et al.,* 2010; *Jung et al.,* 2010, 2011]. Artificial neural networks (ANN) are one of these empirical upscaling techniques that were applied to continental and global scale [see, for example, *Papale and Valentini*, 2003]. We used a set of ANNs in this study to assess the most critical issues in the empirical upscaling processes: the uncertainty in predictions due to the size, length, and distribution of the training data set.

For the empirical models, the training data set used to build and parameterize the functional relationships between inputs and outputs plays a crucial role [see, as example, *Jung et al.,* 2011]. Generally, these models have excellent predictive capacity for interpolation within the parameterization domain (e.g., similar land use type and climate conditions). On the contrary, extrapolation to conditions outside the domain used for model training (e.g., in a different land use or to predict effects of future climate change) remains highly uncertain [e.g., *Jung et al.,* 2009]. For example, training a simple empirical model such as a linear regression using only flux data measured between 11:00 and 13:00 h during the growing season would be inapplicable to nighttime conditions, because it cannot predict nighttime net ecosystem CO_2 exchange (NEE) (nighttime fluxes are determined by respiration, while growing season daytime fluxes are mostly driven by photosynthesis). Similarly, an empirical model parameterized on only high latitude sites will likely fail when applied to tropical conditions, due to different functioning of the vegetation, different drivers, and different ecosystem characteristics. Therefore, the coverage of all conditions in the training data set is crucial for deriving a successful model.

For this reason, the widespread spatial distribution of the existing eddy covariance sites is an important factor to be considered in the context of empirical upscaling [*Carvalhais et al.,* 2010]. Although FLUXNET networks cover a large variety of regions, the measurement sites are not uniformly distributed, with higher density in Europe and North America and large undersampled areas in Africa, South America, and parts of Asia [e.g., *Valentini et al.,* 2014; *Schimel et al.,* 2015]. This nonuniform distribution of sites can significantly affect the parameterization and consequently the results of empirical upscaling models. For example, uneven distribution of flux sites and skewed or misrepresented ecosystem types may lead to systematic errors or biases in the estimated fluxes at regional and global scales. In addition to the spatial representativeness, the temporal extent of the training data is also important. A long time series has more chances to include a variety of climatic conditions and biotic stress events that are needed in order to correctly parameterize the model and to reproduce ecosystems functioning.

The aim of this paper is to assess the effect of the sampling selection of observations on the upscaled fluxes and quantify model uncertainty when the network used for model training has a limited representativeness of the spatial domain. In particular, we used an ANN and tested the following: (1) its efficiency in reproducing carbon and water fluxes, and (2) its efficiency in extrapolating in space and time as a function of the spatial distribution of the training sites.

To address these aspects, after optimization and validation of the ANN, we did a number of experiments where the training data sets were reduced or subsampled to assess the effect on the uncertainty. In particular, extrapolation in time was tested by splitting the data set into two periods while extrapolations in space (for both the total European budget and interannual variability simulations) were checked using the training subsampled networks of increasing size.

Since the amount of information available for empirical model training is directly related to the number of sites collecting measurements in a given region or continent, we expect that artificially reducing the size of the network the uncertainty component due to the training data set representativeness will increase. However, it is worth mentioning that the concepts of "representativeness" and "network size" are, in general, conceptually different. In fact, a smaller network created by following a proper sampling design could bring more information for a specific analysis than a larger but randomly (or biased) established network [see, for example, *Sulkava et al.,* 2011]. The existing network of sites is not designed or optimized following this line of thought, but is instead the result of a number of different scientific questions, often more strongly related to a national requirement than a global strategy. Therefore, in this study it was assumed that a larger network is more representative than a smaller network because it increases the likelihood of sampling different ecosystems under different climatic conditions.

2. Materials and Methods

2.1. Training of the Artificial Neural Network

ANNs can be set up to become purely empirical, nonlinear regression models characterized by a set of nodes (i.e., a simple linear or nonlinear processing unit), often organized in layers and connected by weights that are equivalent to the regression parameters. The first step in using an ANN is the network parameterization process called "training": the ANN is trained by presenting it with sets of input data (drivers) and associated output data. In this phase, a training algorithm minimizes the error between predicted and observed outputs by modifying the connection weights. Once the ANN is trained, the underlying dependencies of the output on the driver variables are mapped onto the weights and the ANN can be then used to predict the variable of interest, starting from new and unseen input values [*Bishop,* 1995].

To avoid the problem of overfitting that results in reduced generalization capacity, "early stopping" was adopted in this study. This was achieved by splitting the data set available for the ANN parameterization into three subsets: (1) the training data set (50% of the data) for adjusting the weights in order to minimize the error between predicted and observed values; (2) a test set of 25% of the data to evaluate the ANN performance during the training and to "early stop" the training process if the errors in the test set started to increase, since this means loss of generalization power, and (3) a validation set (25% of the data) to evaluate the ANN performance on unused data after completion of the training process.

The ANNs used here are the same type used in *Beer et al.* [2010]: a feedforward backpropagation ANN with one hidden layer and sigmoidal transfer functions, trained with the Levenberg-Marquardt algorithm and random initialization of weights. One ANN characteristic that needs to be defined before the training is the number of nodes in the hidden layer. A larger number of nodes not only translate in larger flexibility and adaptation capacity but also bear high risk of overfitting, which needs sufficient training data (increasing the number of nodes automatically leads to increases in the number of parameters). Since there is no standard procedure to define the optimal number of hidden nodes, the different ANN architectures were tested and the ANN with the least complex structure (i.e., lowest number of nodes) and with similar performances is selected [see, for example, *Scardi et al.,* 1999]. In this study, six different architectures (25, 19, 16, 12, 8, and 5 nodes in the hidden layer) were tested.

2.1.1. Data

These ANNs were trained separately to simulate NEE and LE starting from eddy covariance measurements as well as GPP derived via a flux partitioning method [*Reichstein et al.,* 2005]. Fluxes were aggregated at a

monthly time resolution and were acquired at 164 sites across the globe (Table S1 in the supporting information). The sites used are part of the LaThuile collection (www.fluxdata.org) and were processed using standard methods as described in *Reichstein et al.* [2005] and *Papale et al.* [2006]. To better evaluate the model performances relative to other empirical approaches, the same sites and monthly time resolution of *Jung et al.* [2011] were used. The use of different time resolutions were also tested but results did not improve when daily of weekly frequencies were used (despite a larger number of data points being available), rather results slightly decreased probably due to processes not captured by the drivers selected (data not shown).

The drivers used as input for all the fluxes include monthly gap-filled air temperature, incoming shortwave radiation, vapor pressure deficit (VPD), precipitation, the fraction of absorbed photosynthetically active radiation (FAPAR) from the SeaWiFs products available at a 1 km resolution [*Jung et al.,* 2011], and two transformations of top-of-atmosphere radiation (one obtained by rescaling it between 0 and 1 and the other calculating its first derivative for each month) to represent time and seasonality. These inputs were selected on the basis of prior experience [*Beer et al.,* 2010], as well as their availability at site level and in gridded format for Europe, where we performed the spatial upscaling tests. To evaluate the best set of input variables, different combinations of the five drivers (excluding the top-of-atmosphere radiation transformations that were always present) were tested, using two differently organized ANNs: one ANN per plant functional type (PFT) and one ANN trained for all the PFTs. The results indicated that the difference between the prediction accuracy of the two structures at monthly time resolution is minimal (data not shown), and based on this we used only one ANN trained with data from all the sites. The small differences could be due to the high variability among sites in the same PFT, which is comparable to the variability among different PFTs [*Groenendijk et al.,* 2011; *Yuan et al.,* 2014]. Furthermore, the use of a single ANN has the advantage of more available data for training and being independent of the PFT maps (and the related uncertainty) [e.g., *Jung et al.,* 2007; *Giri et al.,* 2005]. As for the inputs, different combinations of the five variables were tested but the best results were obtained using all of them together.

The training and validation of ANNs were accomplished using site-level measurements for the meteorological variables and the tower-pixel value for the retrieved FAPAR; for the spatial application the meteorological variables were used from the ERA-Interim collection (http://apps.ecmwf.int/datasets/data/interim_full_daily/). The use of site-level measurements in the training and validation phase was to ensure that the microclimatic conditions, which have a direct impact on the fluxes, were correctly considered. The agreement between site-level and ERA-Interim variables differed among the different variables [*Balzarolo et al.,* 2014] but should have a limited effect on this study because we focused on the uncertainty resulting from the sampling of the training data set and because the same strategy was followed in all the simulations.

The data set was split into training (Tr), test (Ts), and validation (Va) subsets 10 times through a random extraction from the available data. Each of the 10 Tr-Ts-Va, combined data sets were used to train the six ANNs with different architectures (i.e., different number of hidden nodes) and the best ANN identified on the basis of performances (mean absolute error, MAE; root-mean-square error, RMSE; and coefficient of determination, R^2) with the validation set was selected as our reference ANN. This approach led, at the end of the training process, to 10 selected ANNs (i.e., one for each Tr-Ts-Va extracted data set), and the final output was calculated as the median of the predictions of these 10 ANNs.

2.1.2. ANN Selection

In order to assess the error in the simulation of sites that were not used in the training process, we followed a tenfold scheme similar to *Jung et al.* [2011]: ten groups of sites were randomly selected with each group including 90% of the sites and the 10% excluded, which were always different between the ten folds. This approach achieved the exclusion of each site once. The ANNs were trained following the procedure explained above and then applied to the 10% of the sites excluded to evaluate the real efficiency of the spatial extrapolation. This error assessment forms part of the evaluation of the general capacity of the model to predict the fluxes, which is similar to another widely used empirical approach such as the Model Tree Ensemble (MTE) [*Jung et al.,* 2011]. Because these results are not the primary objective of our study, they are presented in the supporting information (Figure S1 in the supporting information) and discussed here as part of the methodology.

The performance of the ANNs was evaluated by directly comparing both the monthly values and the mean annual value of NEE, GPP, and LE for each site in order to assess the capacity to correctly simulate the intersite variability. The indexes used to analyze the results were the Pearson correlation coefficient (R), the RMSE, and

the Nash-Sutcliffe model efficiency (MEF) coefficient [*Nash and Sutcliffe,* 1970]. Our results are in agreement with those of *Jung et al.* [2011] who used the MTE, with an R of 0.88 between modeled and observed data (RMSE = 1.35 gC m^{-2} d^{-1}; MEF = 0.77) and 0.91 (RMSE = 1.11 MJ m^{-2} d^{-1}; MEF = 0.83) for GPP and LE, respectively, when the monthly values are compared, and 0.82 (RMSE = 0.76 gC m^{-2} d^{-1}; MEF = 0.65) and 0.93 (RMSE = 0.67 MJ m^{-2} d^{-1}; MEF = 0.86) among sites, respectively. For NEE, the model predictions were less accurate (R = 0.69; RMSE = 1.08 gC m^{-2} d^{-1}, MEF = 0.47 at a monthly time scale, and R = 0.53, RMSE = 0.56 gC m^{-2} d^{-1}, MEF = 0.28 among sites), but consistent with previous studies [*Jung et al.,* 2011; *Xiao et al.,* 2014], likely due to the lack of discriminating information among the model predictors that are significant in influencing NEE such as stand age, standing and dead biomass and soil carbon pools, management practices, and past disturbance events. In fact, although these ecosystem characteristics are also important for GPP and LE, NEE is the difference between two large quantities (GPP and ecosystem respiration) which are driven by these factors in different ways; for example, ecosystem respiration is strongly linked to substrate availability, total biomass, and soil conditions, which are not included in the drivers but have an important effect on NEE. For this reason, the remaining analysis of uncertainty in spatial representativeness focused only on GPP and LE.

2.2. Uncertainty Analysis

When empirical models are applied to estimate fluxes at continental or global scales for multiple years, there are two main potential sources of uncertainty linked to the training data: (i) extrapolation in time (i.e., extrapolation to climatic and environmental conditions not present in the training phase) and (ii) extrapolation in space (i.e., sites not used in the training phase). To examine these uncertainties, a specific model experiment was designed to evaluate the errors in the different extrapolations. We focused on GPP predictions using the ANNs trained according to the methodology described above.

2.2.1. Extrapolation in Time

To evaluate the extrapolation in time, we split the data set into two periods. The ANNs were trained on one of the two data sets and used to predict the other data set, and vice versa. The following two different splitting criteria were used: (a) each single site time series with at least 18 months of data was divided into two equally long consecutive periods (58 sites in total, 1026 monthly values), and (b) same as criterion (a) but using only the European sites (25 sites, 459 months).

The two splitting options were applied to ensure the same number of data points in the two subsets.

2.2.2. Extrapolation in Space

Spatial extrapolation was tested through three different exercises where GPP was simulated at sites not used in the training of the ANN. The three tests were designed with increasing levels of extrapolation complexity on the basis of the differences between the training data set and the simulated examples. First, the ANNs were tested in the extrapolation within Europe with a leave-one-out approach, where one site at a time was removed from the training data set (67 sites, see Table S1in the supporting information) and then simulated. To evaluate the extrapolation across continents, the ANNs were trained using all 67 European sites and then applied to simulate 69 sites in North America and the 19 sites in South America, Australia, Africa, and Asia (Table S1in the supporting information), where in most of the these cases the seasonal trend is not present (equatorial sites) or is opposite to those at the training sites (Southern Hemisphere).

2.3. Uncertainty Due to Network Size and Distribution

To estimate the impact of the network size and site distribution on the results of the empirical upscaling, we used the ANNs to estimate annual GPP and LE of Europe with different network configurations. Europe was chosen for this activity because it has the densest network (number of sites per unit area). Subnetworks with an increasing number of sites (5-10-15-20-30-40-50-60) were created by randomly extracting from the 67 available sites (only sites with at least 3 months of quality data in a single year were used, i.e., with at least 75% of original measurements or high-quality gap filled, Table S1 in the supporting information). For each subnetwork size, 50 random extractions were made to evaluate different possible distributions. The ANNs trained using data from the extracted subnetworks were then applied at continental level using monthly gridded inputs from ERA-Interim and SeaWiFs at 0.5° resolution to calculate the two fluxes at the annual time scale, obtaining 50 estimates per subnetwork size.

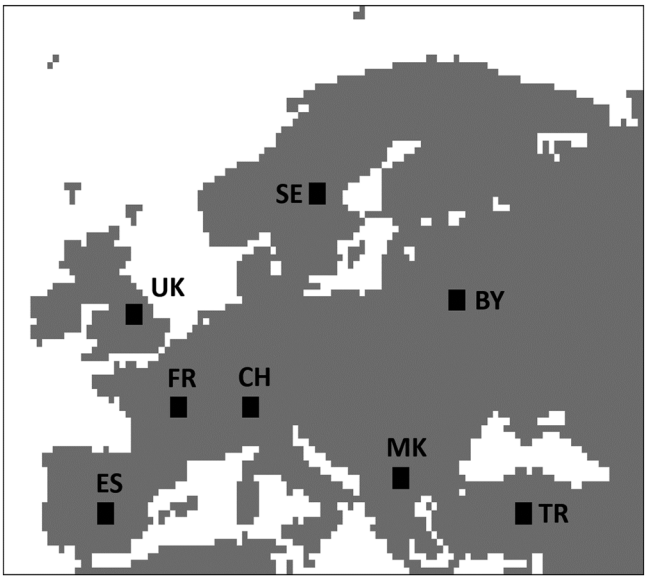

Figure 1. Location of 3 × 3 pixel areas where the local IAV was analyzed. Areas were randomly selected in different European regions: Spain (ES), France (FR), United Kingdom (UK), Switzerland (CH), Sweden (SE), Macedonia (MK), Belarus (BY), and Turkey (TR).

To quantify the effect of the network size and distribution on the model outputs, a reference value for comparison was required. Given that the objective of the analysis is to evaluate the subsampling effect and not to estimate the "real" fluxes, the reference value used in this study was the result predicted by the ANN trained with all the 67 sites available (maximum quantity of information). The reference ANN was also trained 50 times, each time randomly splitting the data set into training, test, and validation sets; the differences in the results can be considered as the uncertainty from the model parameterization. The reference values, GPP_REF and LE_REF, were calculated as the median of the 50 simulations.

2.3.1. Sampling Effect on the Annual Budget

The uncertainty in the annual GPP and LE of Europe was estimated as the difference between the 50 extractions per network size class and the reference values (GPP_REF and LE_REF). In order to quantify the spatial variability of the uncertainty, a comparison was made using the gridded results obtained for GPP and LE in 2005. In this case, the uncertainty was estimated for each pixel as the median of the 50 absolute differences between the extractions and the reference values.

Following this approach, the ANN results are affected at the same time not only by a different spatial representativeness but also by a larger number of data points used in the training when the ANN is parameterized using more sites. This, however, also reflects the reality of what a larger network represents (more sites that provide more data and more ecosystems and climate conditions sampled). In order to disentangle the roles of both effects (i.e., number of sites and amount of data), an additional test was performed following the same network subsampling scheme, while keeping the number of data points constant for the different sampling size. In this test, the number of sites was 10-15-20-25-30 for a total of 300 data points for each extraction (from 10 months × 30 sites to 30 months × 10 sites).

2.3.2. Sampling Effect on Interannual Variability

A reference interannual variability (IAV_REF) time series was calculated as the difference between the annual GPP_REF and LE_REF (for each year in the period 1999–2008) and their 10 year averages. For each network subsample, the value of R between the interannual variability of the extraction and IAV_REF was calculated. A value close to 1 means that the interannual variability obtained with the subsampled network had a similar pattern relative to the one obtained with the full data set. To better analyze how variations in space affects the IAV estimation when the network is subsampled, we performed the same analysis at the European scale for eight locations, focusing on 3 × 3 pixel areas extracted in different countries (Figure 1).

To analyze how much the uncertainties of upscaled GPP and LE are due to a low-level representativeness of the drivers, the similarity between each pixel in Europe and all the sites used in the training was analyzed. We counted, for each pixel and each of the five drivers, the number of sites which had a Nash-Sutcliffe model efficiency (MEF) index between the pixel and site time series larger than zero. In other words, we calculated the MEF between FAPAR, incoming solar radiation, air temperature, VPD, and precipitation extracted at each pixel and the one measured and all 67 sites, obtaining 67 MEF values per pixel and per driver and then counted how many times it was higher than zero (maximum 67). The MEF was originally proposed to compare modeled and observed time series where a value larger than zero indicates that the model results are better than the mean of the observed data. *Carvalhais et al.* [2010] followed a similar approach to identify similarities between eddy covariance site and grid cells characteristics using meteorological and phenological variables.

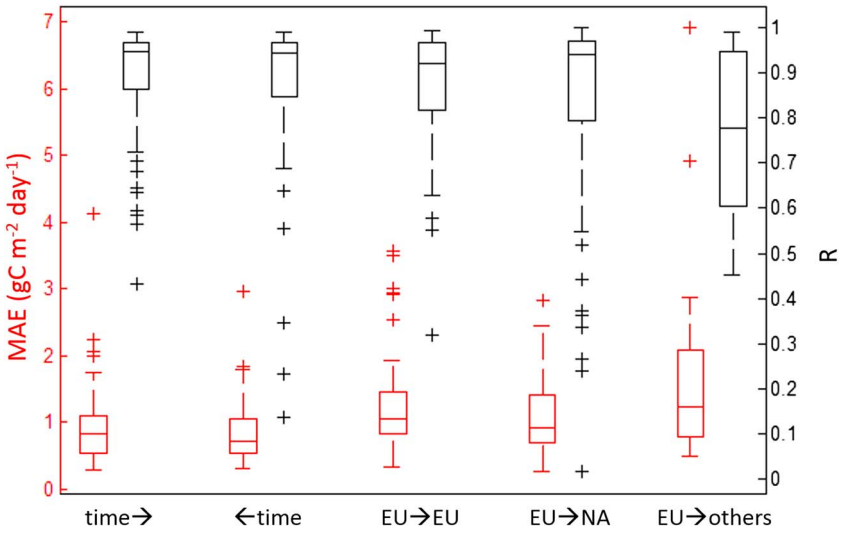

Figure 2. Mean absolute error (MAE gC m^{-2} day^{-1}, in red) and correlation coefficient (*R*, in black) obtained in the extrapolation in time and space in estimating GPP. Each boxplot represents median, interquartile range (IQR), full range, and outliers (indicated with "plus" and defined as points higher or lower, respectively, than the 75th percentile + [1.5*IQR] or 25th percentile − [1.5*IQR]) of the results obtained from simulating each site. The ANNs were trained on a subset of data: "time ➔" and "time ←" (58 sites): all sites time series were split into two parts, one used to train the ANN and the other for validation; time ➔ indicates the use of the oldest part of the time series to predict the newest, time ← indicates the opposite; "EU ➔ EU" (67 sites): trained using the leave-one-out strategy with European sites; "EU ➔ NA" (69 sites): trained using European sites and applied to predict North American sites; "EU ➔ others" (19 sites): trained using European sites and applied to non-European and non-North American sites (Australia, Asia, South America, and Africa).

3. Results and Discussion

3.1. Uncertainty and Errors in Extrapolation in Time and Space

The error in extrapolation in time was lower than that for the extrapolation in space, which may be expected since the ANNs had the possibility to "see" all the sites. Figure 2 shows the distribution of the MAE and the *R* in the extrapolation of GPP in time and space (extrapolation in time calculated by splitting all the sites into two equally long subsets, i.e., option "a" of the four splitting criteria for temporal extrapolation). The other temporal splitting test (only European sites) gave very similar results (Figure S2 in the supporting information) with a slight increase in the extrapolation error when only European data were used. This confirms that, despite the presence of interannual variability and anomalous years, the extrapolation by the ANN for relatively short periods leads to lower errors than for the extrapolation in space. Interestingly, the error in the spatial extrapolation from Europe to North America is on average lower than extrapolation inside Europe. This could be due to the higher level of landscape fragmentation in Europe which, for example, affects the representativeness of the FAPAR retrieval [*Cescatti et al.,* 2012] and the heterogeneity of the land use and management practices applied in Europe. The latter, including their history, are most likely landing to higher differences in the fluxes for the same climate-FAPAR combinations than in North America. As expected, errors increased for extrapolation to areas such as tropical, desert, and equatorial ecosystems where climatic and vegetation conditions (e.g., the seasonal cycle is inverse or not present) differ significantly from the European sites used, although FAPAR and the indexes derived from the top-of-atmosphere radiation used as an input helped to take into account these differences.

3.2. Uncertainty Due to Sample Selection

GPP and LE estimations at the European scale showed remarkable variability as a function of the number of sites used to train the ANN, represented by the total range and interquartile range of the simulation results (Figure 3). This was quantified as the relative change in the error when the network increased in size. It is evident that a reduction in the uncertainty was found relative to its reference (note that this indicates only the uncertainty reduction relative to a simulation considered as the reference). The rate of decrease in uncertainty appeared similar for all the fluxes with an initial fast decrease with networks smaller than 15–20 sites,

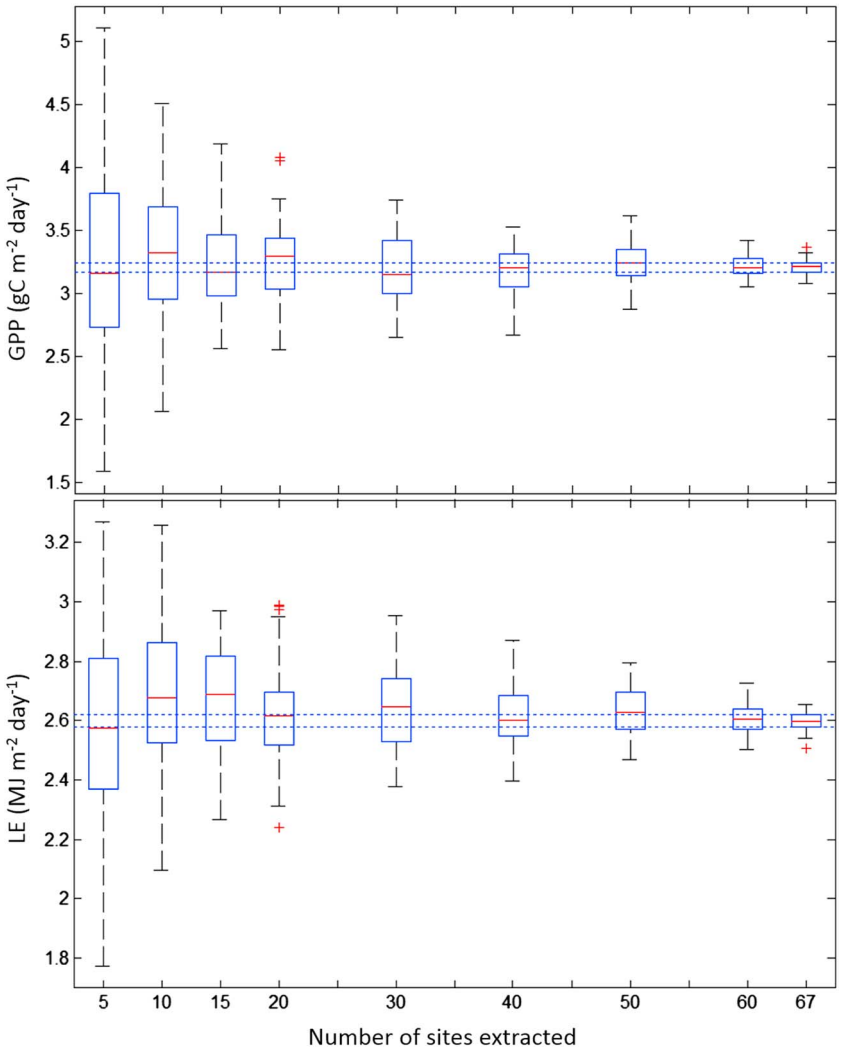

Figure 3. The annual mean flux (gC m^{-2} day^{-1} for GPP and MJ m^{-2} day^{-1} for LE) of Europe in 2005, estimated by ANNs trained with a subsample of the available sites. The boxplots show the distribution of simulated GPP and LE obtained with 50 random extractions of the subset of sites; red crosses indicate outliers. The distribution of results obtained using the whole network (67 sites, last boxplot) is due to the uncertainty linked to the ANN parameterization because, in this case, 50 ANNs have been trained using all the sites and only the splitting between Tr-Ts-Vl sets changes (dashed blue lines). It is important to note that this is not the total uncertainty, but the only uncertainty relative to an ANN realization taken as reference.

then with a smaller rate, and finally, the uncertainty stabilizes in the reference runs when the network size is approximately between 30 and 50 sites. With this approach, it is possible to identify the upper limit of the method applied in the use of the information from the observations. Clearly, the fact that additional sites (for networks larger than 30 sites) did not lead to significantly better estimates is linked to the model and drivers used (maximum use of the information reached). A change in the modeling approach or different explanatory variables would probably lead to different results. Median values were similarly independent of the network size because the 50 simulations with a reduced set of sites covered all the same range of possible values and conditions. These results show that to assess the average yearly continental budget with the proposed modeling approach, in an area with a spatial heterogeneity (in terms of drivers) like the European continent, a network larger than 50 sites does not substantially improve the performance of the model. In contrast, with smaller networks, built without a previous network design strategy (see *Sulkava et al.* [2011] as an example), the uncertainty increases more than proportionally. Inductively, it could be expected that in undersampled continents (such as Africa or South America) the error due to the network size would

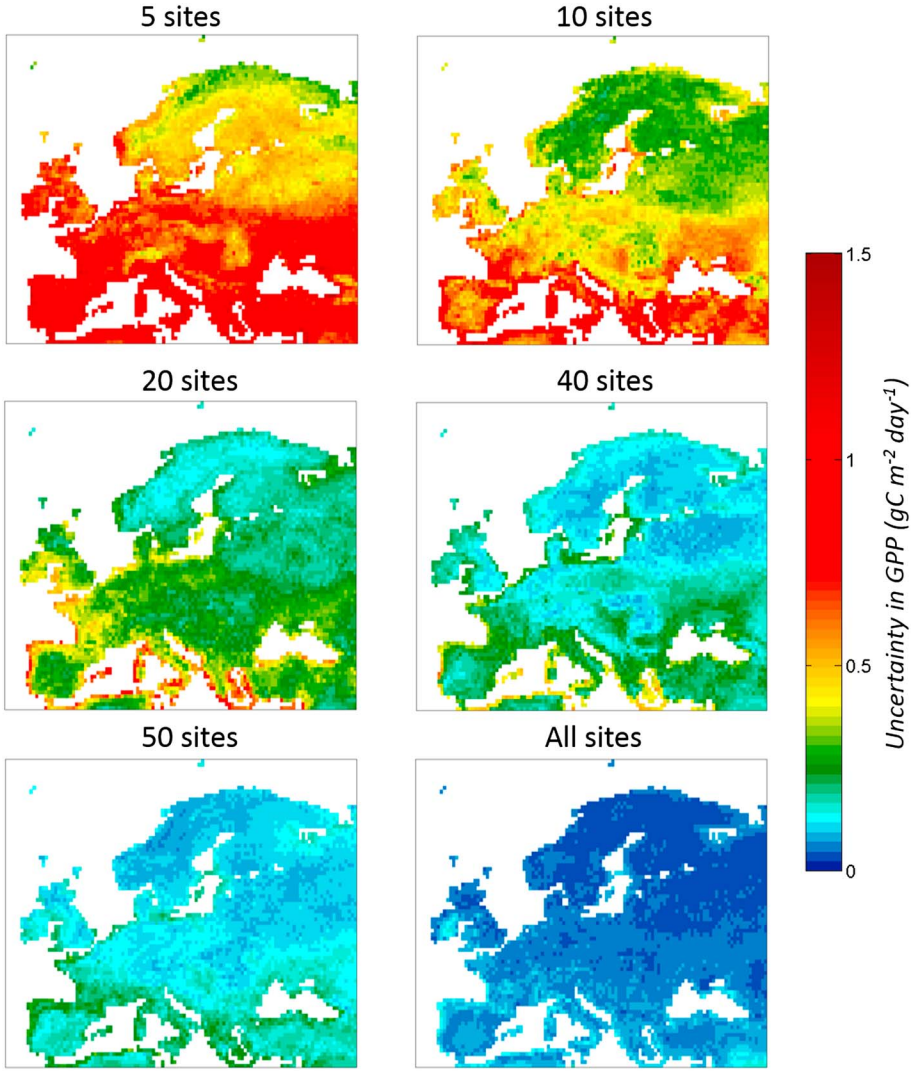

Figure 4. Uncertainty component in the estimation of annual GPP (gC m^{-2} d^{-1}) due to the size of the training data set used to parameterize an ANN. For each map, the values reported are the median of the absolute difference between each of the 50 extractions and the reference value estimated using all the 67 sites available. The bottom right map (All sites) reports the uncertainty in the model parameterization when all sites are used. Note that the color scale is not linear. See Figure S3 in the supporting information for the same map expressed as relative uncertainty (%) with respect to the median.

follow a similar pattern if the spatial heterogeneity of the drivers were similar. Other modeling approaches or areas with lower spatial heterogeneity than Europe may differ in these thresholds, but in general other empirical modeling approaches are expected to give comparable results because they are all based on the extraction of information from the observations. For LE, the results were similar, with the main difference being that, based on the present result, the rate of decrease in uncertainty seen with more than 30 sites is expected to continue.

A similar pattern in the uncertainty of estimates, with a decrease of variability when the number of sites sampled increased, was found with the second test, where the number of points was kept constant for the different extractions (Figure S3 in the supporting information), confirming that spatial representativeness is a key aspect for the ANN parameterization when applied in upscaling activities.

The spatial variation in the uncertainty of GPP and LE was calculated as a median of the absolute differences from the median of the reference simulations for each pixel. The uncertainty in GPP was consistently larger with respect to the average in Southern and Western Europe (Figure 4) and along the Mediterranean and Atlantic coasts. This may be an indication of the underrepresentation of these areas in the training

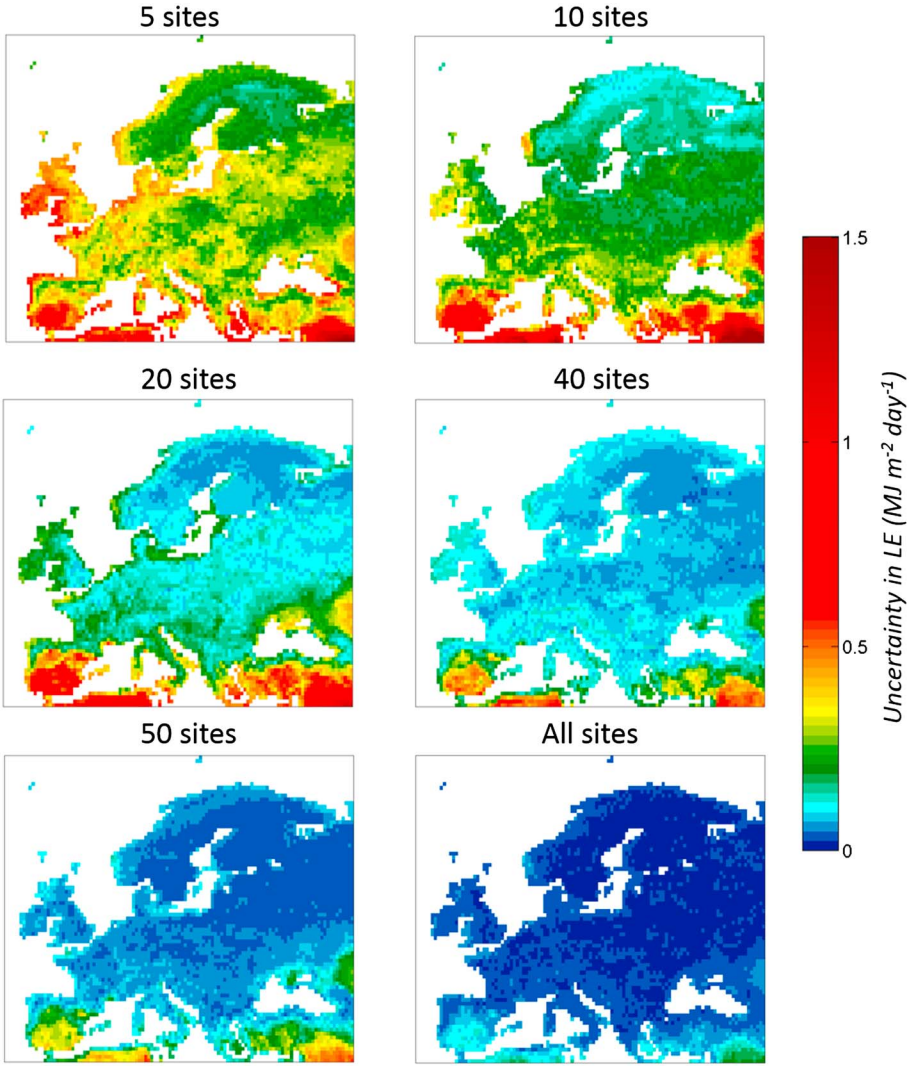

Figure 5. Uncertainty component in the estimation of annual LE (MJ m^{-2} d^{-1}) due to the size of the training data set used to parameterize an ANN. For each map, the values reported are the median of the absolute difference between each of the 50 extractions and the reference value estimated using all the 67 sites available. The bottom right map (All sites) reports the uncertainty in the model parameterization when all sites are used. Note that the color scale is not linear. See Figure S4 in the supporting information for the same map expressed as relative uncertainty (%) with respect to the median.

data set. These areas are characterized by dry and warm climatic conditions, particularly in summer (for the Mediterranean area), and primarily evergreen plant functional types. Other areas of relatively high uncertainty, particularly when less than 20 sites were used, were central and southern France, Italy, Ireland, the western part of the United Kingdom and Poland. The general patterns and hot spots of high and low uncertainty are similar also when quantified as relative uncertainty values (i.e., percentage of uncertainty in respect to the median value, Figure S4 in the supporting information) with the exception of the areas with low fluxes where the relative values are, as expected, high.

The same regions (southern Europe, Mediterranean, and Atlantic coasts) were also identified as hot spots of uncertainty for LE (Figure 5) with a more marked north-south trend. The same map but with uncertainty expressed in relative units (Figure S5 in the supporting information, percentage of uncertainty with respect to the median value) shows areas with generally low absolute fluxes, such as the Scandinavian Peninsula and the Alps, as uncertainties hot spot with respect to the rest of Europe. This is in agreement with results published by *Sulkava et al.* [2011] who, in using a different approach, found higher uncertainty in the same areas because of the network design.

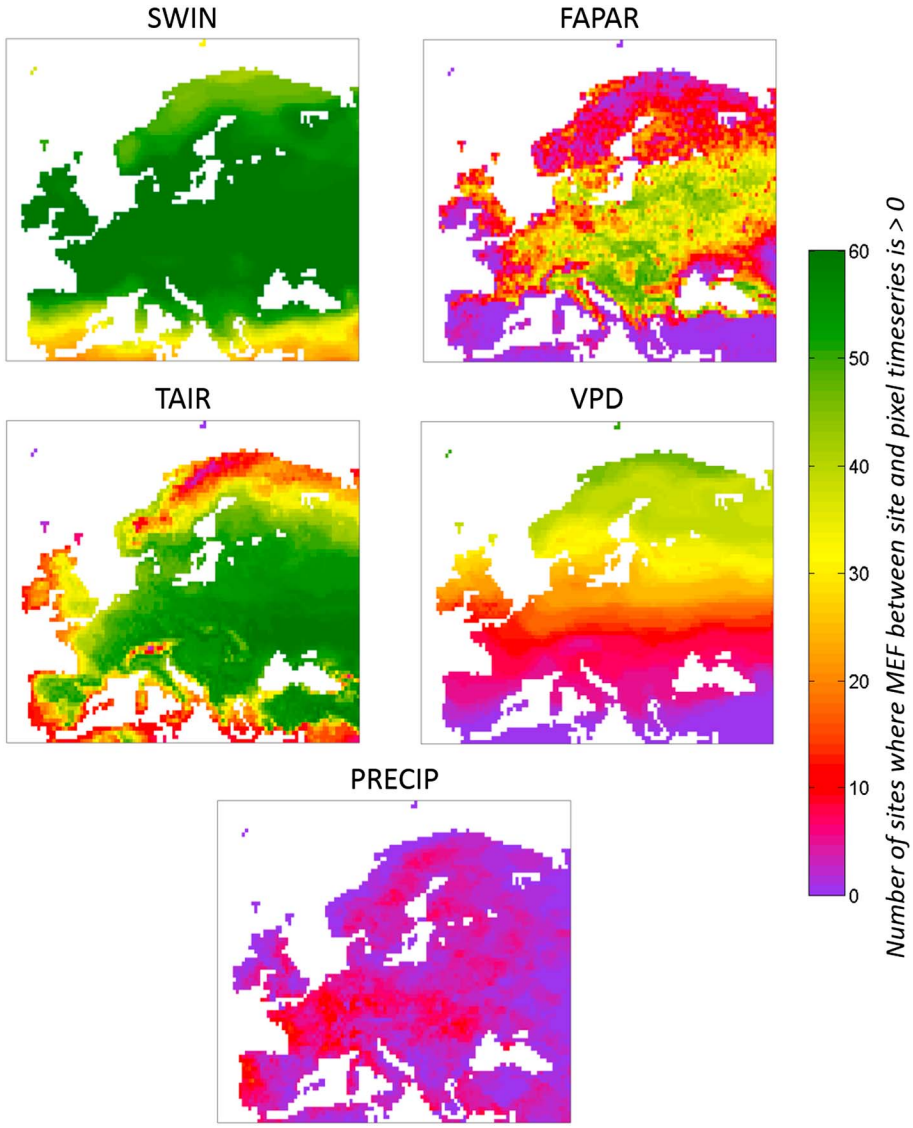

Figure 6. The representativeness in the training data set that has been estimated for each pixel and each of the five drivers. Pixel values are the number of sites that have a model efficiency (MEF, calculated between pixel and site time series) greater than zero. Pixels represented by more than 60 sites were grouped in this last class.

There may be several reasons for these patterns: given the nature of the data-driven model, the main issue is probably linked to the representativeness of the network used in the ANN parameterization. In this case, the areas with higher uncertainty had vegetation and climatic characteristics (in one or more of the five drivers used) that were not well represented in the data set used for the ANN training. In fact, the possible effect of the sea and ocean on the climate, the very dry conditions in South Europe and North Africa, and the cold climate in mountains regions (Alps) and the Scandinavian Peninsula are all underrepresented in the network of eddy covariance sites used. This was confirmed in the analysis of the training set representativeness for each pixel in the drivers domain based on the MEF (Figure 6).

Incoming shortwave radiation was the better represented variable, particularly in the central Europe latitudinal band, due to the high number of sites. A clear north-south trend exists for the representativeness of VPD, where only few sites deliver information on the dry regions in South Europe and North Africa. The same areas were also the less represented in terms of FAPAR, which also showed underrepresented areas north of the Black Sea, the Scandinavian Peninsula, and along the coasts of Italy, France, and the UK. It is also important to remark that part of the areas less represented in terms of VPD and FAPAR are in areas with bare soil or sparse vegetation

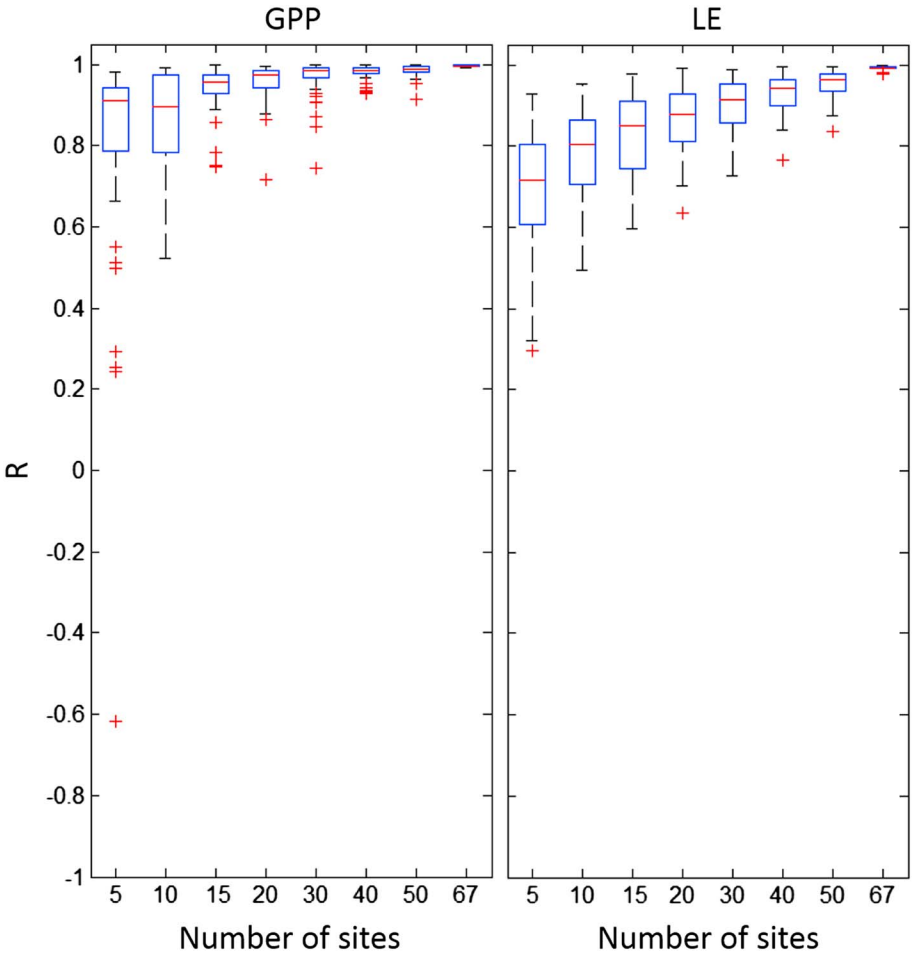

Figure 7. Evaluation of the interannual variability estimates. Each boxplot represents the distribution of the correlation coefficient (*R*) calculated between the 50 ten years annual time series (1999–2008) using an ANN trained with subsamples of the sites and the reference time series estimated with training ANN with all the 67 sites.

(in particular North Africa and Anatolia peninsula). Coastal areas, the Alps, and the Scandinavian region were similarly less represented in terms of air temperature while precipitation showed in general the lowest degree of similarity which is probably due to its high spatial variability; this could further indicate a minor role of this variable in the ANNs parameterization.

By comparing "representativeness of drivers" and uncertainties in the fluxes, we detected some common patterns: South Europe and North African regions characterized by larger uncertainty in both fluxes (both in absolute and relative terms) were less represented for VPD and FAPAR, while air temperature is less well represented for coastal areas, western UK, and Ireland. This could be the reason for the high uncertainty, particularly in GPP, in these regions. Air temperature is also a less well represented driver in the inland areas of the Scandinavian Peninsula, but its effect on the uncertainty in GPP and LE is low in absolute terms while significant when it is expressed as relative quantities (Figures S3 and S4 in the supporting information). FAPAR is the driver with higher fine-scale variability in terms of representativeness, which was also reflected in some of the flux uncertainty patterns (e.g., GPP in Poland, western UK and France, LE east of the Black Sea, Ireland, and Norway coasts).

In addition to the continental budget, the effect of network size and site distribution on the interannual variability of the estimated fluxes (IAV) was investigated. Similar to the annual budget, there was a clear convergence as the network size increased for both GPP and LE. Interestingly, IAV trends for GPP were in agreement with a network size of 40 sites while the improvement in *R* for LE was more linear with the increasing number of sites, showing a significant difference even between 50 and 67 sites (Figure 7). Given that the prediction capability

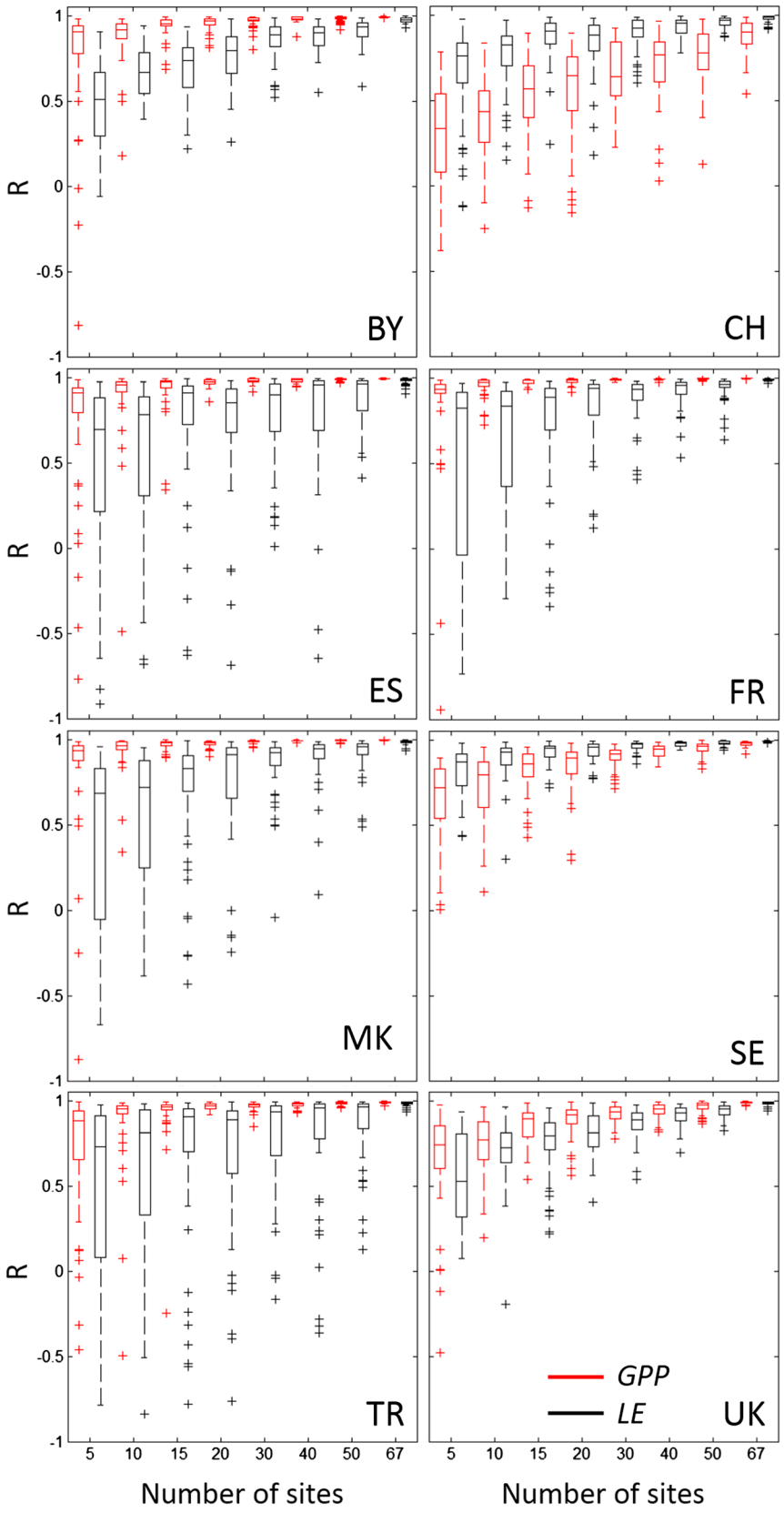

Figure 8. Distribution of the correlation coefficient (*R*) between the interannual variability in the 10 year period calculated using the 50 estimates per sampling size and the reference estimated by training the ANN with all the 67 available sites, for the eight regions reported in Figure 1. Red boxplots are for GPP, and black for LE.

Table 1. Summary of the Uncertainties in GPP[a]

Uncertainty Type	Method	Results
Model validation	Tenfold cross validation using FLUXNET sites	RMSE of 0.76 gC m^{-2} d^{-1} in this study (Figure S1), RMSE of 0.74 gC m^{-2} d^{-1} in *Jung et al.* [2011]
Errors in the extrapolation in time	Training and validation splitting the data set in two equally long subsets	MAE of 0.78 (IQR 0.53–1.1) gC m^{-2} d^{-1} (Figure 2). MAE of 0.97 (IQR 0.76–1.36) gC m^{-2} d^{-1} when only European sites are used (Figure S2)
Errors in the extrapolation in space	Training and validation based on different sets of sites according to different geographical regions	MAE of 0.91 (IQR 0.7–1.41) gC m^{-2} d^{-1} in extrapolation from Europe (training) to North America (validation), MAE of 1.24 (IQR 0.8–2.09) gC m^{-2} d^{-1} in extrapolation from Europe (training) to other regions (validation) (this study, Figure 2)
Effect of network size and sampling in extrapolation in space	Subsampling the full European network using a bootstrapping approach and using the estimate obtained using all the sites available as reference	Interquartile range for the reduced networks varying between 0.12 and 1.06 gC m^{-2} d^{-1}, full range varying between 0.38 and 3.4 gC m^{-2} d^{-1} (this study, Figure 3)
Overall GPP estimation uncertainty	Uncertainty in the estimation of average GPP at global level derived from different modeling approaches	Different modeling approaches were compared at global scale including two empirical models, four offline land carbon process models, and four online Earth system process models. GPP ranged between 112 and 168 PgC yr^{-1} (equivalent to 2.28–3.5 gC m^{-2} d^{-1}), with median 141.5 PgC yr^{-1} (equivalent to 2.87 gC m^{-2} d^{-1}) and IQR 130–148 PgC yr^{-1} (equivalent to 2.64–3.03 gC m^{-2} d^{-1}). Results are from *Anav et al.* [2015]

[a]The table summarizes the different uncertainties addressed in this paper and also in other papers for comparison. IQR = Interquartile range.

of the ANN for LE is higher than that for the other fluxes (Figure S1 in the supporting information), this result is somehow unexpected and indicates a pronounced spatial heterogeneity in the interannual pattern of evapotranspiration. One possible reason is that a driver exclusively linked to the IAV of water fluxes is missing (e.g., soil water content is not used because it is not available as high-quality gridded product for areas with dense vegetation cover for the upscaling). Other explanations are that precipitation—linked to the water fluxes—is the driver with a higher spatial heterogeneity (i.e., least represented by the 67 sites used), or simply that the water budget IAV has a larger spatial variability in Europe relative to GPP (i.e., affected more by the network distribution).

The analysis of the effect of spatial heterogeneity on GPP and LE IAV in relatively small areas of Europe was done following the same approach presented for the continental analysis but for the eight regions (defined in Figure 1) separately. The different areas present contrasting results of the interannual variability modeled for the two fluxes (Figure 8). Even though the result confirmed that IAV of GPP does not change significantly when the network size becomes larger than 40 sites, some deviations from this pattern were identified. For example, Switzerland, and to a lesser extent UK and Sweden, showed a stronger effect of the number of sites used (Figure 8, plots CH, UK, and SE). These three areas are characterized by a low degree of representativeness of FAPAR (Figure 6) that could explain the uncertainty in the IAV simulations. In other words, areas less similar to the 67 sites available have more chance to not be adequately represented in the reduced training set extracted in the simulations. Switzerland, in particular, is characterized by a unique combination of drivers due to its mountainous environment (e.g., low temperature similar to northern Europe but higher incoming solar radiation since it is at a lower latitude), which is one of the less represented driver combinations that could explain the high uncertainty in the IAV estimates even when all the sites are used.

The effect of network size on IAV of LE in general was also similar to the one at the continental scale but presented a less pronounced decrease of uncertainty with the growth of the network in most areas. The trend was, however, different for the eight locations, with southern regions (Spain, Turkey, Macedonia, and France) characterized by lower values of R (even negative) when the network was small (5–10 sites) and a slower improvement with the growth of the network (Figure 8, plots ES, TR, MK, and FR). This can be due to the lower level of representativeness of the key variables affecting LE (e.g., VPD and precipitation) in southern Europe (Figure 6), making the estimation of fluxes in this area more difficult for both the annual budgets and IAV.

The different tests performed, also in comparison with some other studies, are summarized in Table 1. The magnitude of the uncertainty in temporal extrapolation is relatively small in particular when compared to the errors in extrapolation in space, but it can still play an important role. Differences among models used

to estimate global GPP have comparable values with respect to the uncertainties addressed in this analysis [*Anav et al.*, 2015], confirming that underrepresentativeness is an important issue, in particular in the tropics and equatorial areas where a limited number of eddy covariance sites exists despite the important contribution of these regions in the global GPP.

4. Conclusions

This study helped to improve our understanding of uncertainties and limitations in the use of machine learning methods for the upscaling of water and carbon fluxes by provided insights into the effect of network size and distribution on model parameterization, which can be used as guidance for the future development of the observation networks. ANNs were confirmed as a useful tool for the GPP and LE estimation at continental scale, while for NEE the results were less accurate. This is probably due to the marked effect of management practices and disturbances as well as site history in NEE (e.g., through their effects on ecosystem respiration). These variables, however, could not be included as drivers in the ANNs, suggesting a future direction for the development of the method. This was confirmed in *Xiao et al.* [2014] where they found in a similar upscaling activity that the inclusion of aboveground biomass and stand age as drivers reduced the error in both GPP and NEE by about 15%, and an additional 3–4% of error reduction was found when adding leaf nitrogen content as an input. The results suggest that there are key drivers that could substantially improve upscaling results if available in gridded format and at a global scale. These missing drivers are particularly the ones related to ecosystem respiration and NEE such as biomass, soil carbon and water content, nutrient availability, and history of the management and disturbances. The validation activity performed using data acquired on different continents showed that extrapolation to similar climatic and vegetation conditions is possible (e.g., between Europe and North America), though the errors increased when the extrapolation was to areas with different seasonal cycles and regulating factors. This further highlights the need for additional research sites to provide the essential data to reliably estimate GPP, NEE, and LE fluxes in regions that are currently underrepresented while at the same time critical for global C cycling (Tropical, Equatorial, Mediterranean, and dry regions).

We conclude that site network size plays an important role even in a simple model with five variables as drivers. By using all available sites in ANN parameterization and considering the results as references (a large simplification of reality), the uncertainty component just due to the sites sampling and data set size in the continental budget varied up to ±50% for GPP and ±25% for LE when the network was particularly small (5 or 10 sites). The same effect was found for the interannual variability where results showed a stronger dependency for LE than GPP and a higher degree of spatial variability. This suggests that if a similar upscaling method is applied to continents with small networks (e.g., Africa or South America), compared to Europe or the U.S., the error due to the sampling could be large.

The analysis of the spatial variability of the uncertainty in the flux estimates and a quantification of the representativeness of the network of sites in terms of the domain of drivers both highlight the importance of variables such as FAPAR, VPD, and air temperature, which were less represented in some regions (dry areas, coastal zones, mountains, and cold environments). These same areas exhibited a higher level of uncertainty, confirming the crucial role of the training data set in upscaling exercises using empirical models: biases in the distribution of measurements could lead to high uncertainties in the undersampled areas where results are extrapolated. Our investigations here suggest that future developments of the European flux network should improve the coverage of areas poorly sampled in the climate space including south-east and south-west Europe and coastal and mountainous regions.

Eddy covariance networks across the globe provide unique data sets that are used across different disciplines and for a large number of applications. Globally, the number of eddy covariance sites is still growing (more than 700 registered in FLUXNET in June 2015—http://fluxnet.ornl.gov—although not all of them are still active), and each site could contribute to better-parameterized models applied in upscaling exercises. However, most of the data are still not shared and there are sites not yet registered in the continental and global networks. To better understand carbon and water cycles, we need a joint collaboration across communities that can only start with data integration, standardization, harmonization, and open access initiatives, similar to what was accomplished with the LaThuile data set in 2007 (www.fluxdata.org). Measurement networks and research infrastructures such as ICOS, NEON, Ameriflux, and OzFlux have started to move in this direction and the hope is that more sites and networks will join these initiatives with the aim to find answers to the global issues.

Acknowledgments
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, and USCCC. We acknowledge the financial support to the eddy covariance data harmonization provided by CarboEuropeIP, COST ES0804 ABBA, FAO-GTOS-TCO, iLEAPS, Max Planck Institute for Biogeochemistry, National Science Foundation, University of Tuscia, Université Laval, Environment Canada, and the U.S. Department of Energy, as well as the database development and technical support from Berkeley Water Center, Lawrence Berkeley National Laboratory, Microsoft Research eScience, Oak Ridge National Laboratory, University of California Berkeley, and the University of Virginia. The data used in this work are available at the FLUXNET database (www.fluxdata.org). G.T., M.R., and M.D.M. were supported by the EU project GEOCARBON under grant agreement 6283080; D.P. thanks the support of the EU project ICOS-INWIRE under grant agreement 313169. D.P., M.R., and M.D.M. thank the EU for supporting the BACI project funded by the EU's Horizon 2020 Research and Innovation Programme under grant agreement 640176. We thank Gabriela Shirkey for her careful language edits in the later version of this manuscript.

References

Anav, A., et al. (2015), Spatiotemporal patterns of terrestrial gross primary production: A review, *Rev. Geophys.*, *53*, doi:10.1002/2015RG000483.
Aubinet, M., T. Vesala, and D. Papale (2012), *Eddy Covariance - a Practical Guide to Measurement and Data Analysis*, Springer, doi:10.1007/978-94-007-2351-1.
Baldocchi, D. (2008), 'Breathing' of the terrestrial biosphere: Lessons learned from a global network of carbon dioxide flux measurement systems, *Aust. J. Bot.*, *56*, 1–26, doi:10.1071/BT07151.
Balzarolo, M., et al. (2014), Evaluating the potential of large-scale simulations to predict carbon fluxes of terrestrial ecosystems over a European eddy covariance network, *Biogeosciences*, *11*, 2661–2678, doi:10.5194/bg-11-2661-2014.
Beer, C., M. Reichstein, P. Ciais, G. D. Farquhar, and D. Papale (2007), Mean annual GPP of Europe derived from its water balance, *Geophys. Res. Lett.*, *34*, L05401, doi:10.1029/2006GL029006.
Beer, C., et al. (2009), Temporal and among-site variability of inherent water use efficiency at the ecosystem level, *Global Biogeochem. Cycles*, *23*, GB2018, doi:10.1029/2008GB003233.
Beer, C., et al. (2010), Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate, *Science*, *329*, 834–838, doi:10.1126/science.1184984.
Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford Univ. Press, New York.
Carvalhais, N., M. Reichstein, G. J. Collatz, M. Mahecha, M. Migliavacca, C. S. R. Neigh, E. Tomelleri, A. A. Benali, D. Papale, and J. Seizas (2010), Deciphering the components of regional net ecosystem fluxes following a bottom-up approach for the Iberian Peninsula, *Biogeosciences*, *7*, 3707–3729, doi:10.5194/bg-7-3707-2010.
Cescatti, A., et al. (2012), Intercomparison of MODIS albedo retrievals and in situ measurements across the global FLUXNET network, *Remote Sens. Environ.*, *121*, 323–334, doi:10.1016/j.rse.2012.02.019.
Giri, C., Z. Zhu, and B. Reed (2005), A comparative analysis of the Global Land Cover 2000 and MODIS land cover data sets, *Remote Sens. Environ.*, *94*, 123–132, doi:10.1016/j.rse.2004.09.005.
Groenendijk, M., et al. (2011), Seasonal variation of photosynthetic model parameters and leaf area index from global Fluxnet eddy covariance data, *J. Geophys. Res.*, *116*, G04027, doi:10.1029/2011JG001742.
Hastie, T., R. Tibshirani, and J. Friedman (2001), *The Elements of Statistical Learning*, *Springer Ser. Stat.*, Springer, New York.
Haxeltine, A., and I. C. Prentice (1996), BIOME3: An equilibrium terrestrial biosphere model based on ecophysiological constraints, resource availability and competition among plant functional types, *Global Biogeochem. Cycles*, *10*, 693–709, doi:10.1029/96GB02344.
Jung, M., et al. (2007), Uncertainties of modelling gross primary productivity over Europe: A systematic study on the effects of using different drivers and terrestrial biosphere models, *Global Biogeochem. Cycles*, *21*, GB4021, doi:10.1029/2006GB002915.
Jung, M., M. Reichstein, and A. Bondeau (2009), 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.
Jung, M., et al. (2010), Recent decline in the global land evapotranspiration trend due to limited moisture supply, *Nature*, *467*, 951–954, doi:10.1038/nature09396.
Jung, M., et al. (2011), Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, *J. Geophys. Res.*, *116*, G00J07, doi:10.1029/2010JG001566.
Krinner, G., N. Viovy, N. de Noblet-Ducoudre, J. Ogee, J. Polcher, P. Friedlingstein, P. Ciais, S. Stitch, and C. Prentice (2005), A dynamic global vegetation model for studies of the coupled atmosphere biosphere system, *Global Biogeochem. Cycles*, *19*, GB1015, doi:10.1029/2003GB002199.
Lin, J. C., M. R. Pejam, E. Chan, S. C. Wofsy, E. W. Gottlieb, H. A. Margolis, and J. H. McCaughey (2011), Attributing uncertainties in simulated biospheric carbon fluxes to different error sources, *Global Biogeochem. Cycles*, *25*, GB2018, doi:10.1029/2010GB003884.
Monteith, J. L. (1972), Solar radiation and productivity in tropical ecosystems, *J. Appl. Ecol.*, *9*(3), 747–766.
Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I—A discussion of principles, *J. Hydrol.*, *10*(3), 282–290.
Papale, D., and R. Valentini (2003), 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.
Papale, D., et al. (2006), 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.
Papale, D., D. A. Agarwald, D. Baldocchi, R. B. Cook, J. B. Fisher, and C. van Ingen (2012), Database maintenance, data sharing policy, collaboration, in *Eddy Covariance - a Practical Guide to Measurement and Data Analysis*, edited by M. Abinet, T. Vesala, and D. Papale, pp. 399–424, Springer.
Reichstein, M., et al. (2005), 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.
Reichstein, M., et al. (2007), Reduction of ecosystem productivity and respiration during the European summer 2003 climate anomaly: A joint flux tower, remote sensing and modelling analysis, *Global Change Biol.*, *13*(3), 634–651, doi:10.1111/j.1365-2486.2006.01224.x.
Running, S. W., P. E. Thornton, R. Nemani, and J. M. Glassy (2000), Global terrestrial gross and net primary productivity from the earth observing system, in *Methods in Ecosystem Science*, edited by O. E. Sala, R. B. Jackson, and H. A. Mooney, pp. 44–57, Springer, New York.
Scardi, M., W. Lawrence, and W. Harding Jr. (1999), Developing an empirical model of phytoplankton primary production: A neural network case study, *Ecol. Modell.*, *129*, 213–223.
Schimel, D., R. Pavlik, J. B. Fischer, G. P. Asner, S. Saatchi, P. Townsend, C. Miller, C. Frankenberg, K. Hibbard, and P. Cox (2015), Observing terrestrial ecosystems and the carbon cycle from space, *Global Change Biol.*, *32*, 1762–1776, doi:10.1111/gcb.12822.
Sitch, S., et al. (2003), Evaluation of ecosystem dynamics plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model, *Global Change Biol.*, *9*, 161–185, doi:10.1046/j.1365-2486.2003.00569.x.
Sulkava, M., S. Luyssaert, S. Zaehle, and D. Papale (2011), Assessing and improving the representativeness of monitoring networks: The European flux tower network example, *J. Geophys. Res.*, *116*, G00J04, doi:10.1029/2010JG001562.
Valentini, R., et al. (2014), A full greenhouse gases budget of Africa: Synthesis, uncertainties, and vulnerabilities, *Biogeosciences*, *11*, 381–407, doi:10.5194/bg-11-381-2014.
Verma, M., et al. (2014), Remote sensing of annual terrestrial gross primary productivity from MODIS: An assessment using the FLUXNET La Thuile data set, *Biogeosciences*, *11*(8), 2185–2200, doi:10.5194/bg-11-2185-2014.
Vetter, M., et al. (2008), Analyzing the causes and spatial pattern of the European 2003 carbon flux anomaly using seven models, *Biogeosciences*, *5*, 561–583, doi:10.5194/bg-5-561-2008.
Xiao, J., et al. (2010), 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.

Xiao, J., et al. (2014), Data-driven diagnostics of terrestrial carbon dynamics over North America, *Agric. For. Meteorol.*, *197*, 142–157, doi:10.1016/j.agrformet.2014.06.013.

Xiao, X. M., Q. Y. Zhang, B. Braswell, S. Urbanski, S. Boles, S. Wofsy, M. Berrien, and D. Ojima (2004), Modeling gross primary production of temperate deciduous broadleaf forest using satellite images and climate data, *Remote Sens. Environ.*, *91*(2), 256–270.

Yang, F., K. Ichii, M. A. White, H. Hashimoto, A. R. Michaelis, P. Votava, A.-X. Zhu, A. Huete, S. W. Running, and R. R. Nemani (2007), 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.

Yuan, W., et al. (2014), Vegetation-specific model parameters are not required for estimating gross primary production, *Ecol. Modell.*, *292*, 1–10, doi:10.1016/j.ecolmodel.2014.08.017.