Uncertainty analysis of gross primary production upscaling using Random Forests, remote sensing and eddy covariance data

Gianluca Tramontana, Kazuito Ichii, Gustau Camps-Valls, Enrico Tomelleri, Dario Papale

1. Introduction

The accurate estimation of spatially explicit carbon fluxes is an important goal to improve the understanding of the feedbacks between the terrestrial biosphere and the atmosphere in the context of global change and facilitation of climate policy decisions (Running et al., 1999).

The carbon, water, and energy fluxes of land ecosystems are intimately connected (Beer, Reichstein, Ciais, Farquhar, & Papale, 2007; Beer et al., 2009; Schimel, Braswell, & Parton, 1997). The in-situ estimations of carbon, water and energy fluxes can be obtained by the eddy covariance technique (Aubinet, Vesala, & Papale, 2012), a well-developed method for measuring trace flux quantities between the biosphere and the atmosphere (Running et al., 1999). Using this technique, net ecosystem carbon exchange (NEE) is directly measured, whereas gross primary production (GPP) and total ecosystem respiration are estimated using different partitioning methods (Desai et al., 2008; Lasslop et al., 2010; Reichstein et al., 2005).

From the site level measurements, the regional, continental and global estimates of carbon fluxes are obtained by spatial extrapolation conducted with models in which the spatial variability is mostly driven by earth observation data (Jung, Le Maire, et al., 2007; Jung, Reichstein, & Bondeau, 2009; Jung, Vetter, et al., 2007; Running et al., 1999).

Both process-based and empirical approaches are commonly used to estimate spatially explicit carbon fluxes. Process-based models such as ORCHIDEE (Krinner et al., 2005), BIOME3 (Haxeltine & Prentice, 1996) and...
and LPJ-DGVM (Sitch et al., 2003) explicitly describe the physical processes that regulate energy, carbon and water cycles. These models are useful for predicting future scenarios under global climate change. However, the use of these models has limitations due to inherent assumptions such as the complexity of the model structure and ad-hoc parameters. The empirical models are established differently and use statistics to find the best possible relation between a set of explanatory variables (inputs) and one or more target (outputs) without including an explicit parametric description of the physical processes relating them. In general, machine learning (ML) techniques are applied for data-driven models that use empirical data (measured examples) to develop quantitative predictive models (Hastie, Tibshirani, & Friedman, 2001). Several ML algorithms that are based on different statistical or computational principles, such as the Artifical Neural Networks (ANN, Papale & Valenti, 2003), the Model Tree Ensemble (MTE, Jung et al., 2009) and the Support Vector Machine (SVM, Yang et al., 2007) are applied to upscale fluxes.

Because of the basic premise, the application of empirical models is strictly dependent on the variables used as drivers and on the representativeness of all primary ecosystem characteristics that affect carbon fluxes (i.e., vegetation type, age, health, abiotic and biotic stress, seasonality, and phenology). Additionally, empirical models generally predict outcomes for samples that have similar characteristics to training data, but typically fail when they are applied to situations not observed during the training phase (extrapolation). The ability of the model to correctly estimate the output when applied to new examples is the “generalisation” and it is affected by many factors, such as model complexity, missing of important drivers, data quality and the representativeness of the training examples.

In the spatial upscaling by empirical models, the choice of the drivers is crucial and often it is a compromise between usefulness for the upscaling purposes and availability in gridded format with sufficient quality. As an example, the importance of soil characteristics in ecosystem carbon flux dynamics is well known, but these data are generally not used because their limited availability as spatially explicit databases and high uncertainty. In contrast, meteorological data are often used as drivers because these key variables are measured at the sites and are available as spatially explicit fields from reanalysis products. Moreover, meteorological data provide information both for seasonal conditions and for daily stress factors but not for the green biomass and the vegetation health, which can be inferred from earth observation data.

Remote sensing variables, particularly vegetation indices, do not directly represent carbon fluxes processes (Jung et al., 2008), but as shown previously, they are statistically related to ecosystem fluxes (Olofsson et al., 2008; Rahman, Sims, Cordova, & El-Masri, 2005). Vegetation indices are calculated using measured reflectances in specific spectral bands that are related to some chemical and physical properties of the vegetation. For example, greenness indices such as the Normalised Difference Vegetation Index (NDVI) or the Enhanced Difference Vegetation Index (EVI) (Olofsson et al., 2008; Sims et al., 2008) are related to the amount of green biomass (e.g., leaf area index, LAI), whereas water indices such as the Normalised Difference Water Index (NDWI) (Gao, 1996) provide information on the canopy water content. Remote sensing data are also used as the basis to derive the land cover maps that are used in modelling exercises when the model parameterisation is specific for a Plant Functional Type (PFT).

Generally, ML methods use both meteorological data and measured or derived remote sensing data as inputs to estimate carbon fluxes (Jung et al., 2011). At the site level, this strategy provided satisfactory results (Moffat, Beckstein, Churkina, Mund, & Heimann, 2010), though the model parameters could be affected by the uncertainty of the measurements. When the models are applied at larger spatial scales, gridded versions of the inputs are necessary, and the uncertainties must be considered an additional source of errors that affect simulated outputs.

The spatially gridded inputs necessary to apply the models can be measured (e.g., the remotely sensed spectral reflectances), obtained by other models or interpolation techniques (e.g., the gridded meteorological data) or be obtained from classification schemes such as the land cover or PFT maps. If a ML model only uses spatially explicit variables that are directly measured as inputs (e.g., vegetation indices or spectral reflectances), the uncertainty associated with the production of the derived spatial data is removed. Moreover, although remotely sensed spectral reflectance and land surface temperature provide a great amount of useful information, if the modelling exercise is performed without meteorological or land cover data, important information may be missing. For example, during drought, an immediate effect occurs on the fluxes caused by stomata closure, but reflectance is generally affected later when the stress conditions persist (e.g., when the leaf tissue chlorophyll contents changes).

In this study, a diagnostic machine learning method called Random Forest (RF) (Breiman, 2001), was used to predict the eight days GPP and the mean European annual carbon budget, with the aim of analysing the impacts of different sources of uncertainty on the predictions. RF methods were used with the GPP derived from the eddy-covariance measurements of NEE. At site and ecosystem levels, the effects of the missing key drivers on the accuracy of GPP predictions were evaluated. At European scale it has been analysed the effects of the uncertainty in gridded drivers that are obtained by other models (meteorological variables and land cover maps) on the mean European annual GPP.

2. Materials and methods

2.1. Site level data

In this study, the time series of meteorological variables, GPP, and remote sensing measured and derived data coming from 44 European study sites were used (Table 1). GPP and meteorological in-situ data, in particular the incoming solar radiation, air temperature, vapour pressure deficit (VPD) and precipitation, were obtained by the European database of flux data (www.europe-fluxdata.eu), while the satellite data were obtained by the MODIS sensor on board of the TERRA satellite.

The measurements of net CO2 fluxes (NEE) were quality filtered using standard methods (Papale et al., 2006) and gap-filled using Artificial Neural Network (ANN) method (Papale & Valenti, 2003). The GPP was calculated from NEE measurements using the partitioning method described by Reichstein et al. (2005). Half hourly estimations of GPP were aggregated to eight days time resolution to be consistent with the satellite data (Sims et al., 2008; Xiao et al., 2004, 2010). An additional quality filtering was applied to the data using only eight days periods in which the GPP was calculated starting with at least 80% of original or high quality gap filled NEE measurements (quality flag in the data equal to 0.8). This last filtering method was also applied to each meteorological variable.

The remotely sensed data were obtained by the MODIS sensor on board of the TERRA satellite. For the site level analysis, the MODIS cutout data, (available at the FLUXNET website (http://daac.ornl.gov/cgi-bin/MODIS/GR_cofs_1/mdb_viz.html)), were used. These data cover 7×7 km2 centred on each site. The time series of spectral reflectances (MOD09A1; Xiao et al., 2010; Vermote & Vermuelen, 1999), daily land surface temperature (LST; Xiao et al., 2010; Wan, Zhang, Zhang, & Li, 2002), fraction of absorbed photosynthetically active radiation (IPAR, MOD15A2 product; Xiao et al., 2010; Myneni et al., 2002), and GPP (MOD17A2 product; Xiao et al., 2010; Heinsch et al., 2006; Zhao, Heinsch, Nemani, & Running, 2005) were acquired. The MOD09A1 spectral reflectances were further used to calculate NDVI, EVI and NDWI vegetation indices. The MODIS products had specific quality flags that
were used to select only high or good quality pixel values at the eddy tower positions.

2.2. Spatial data

Two meteorological gridded products, two land use cover maps and one IPAR gridded dataset were used to evaluate the variability of the mean annual GPP due to the uncertainty of the modelled spatially explicit inputs. The analysis of the spatial grid was conducted for the period 2006–2008.

The meteorological grids were provided by the Modern-Era Retrospective Analysis For Research and Applications (MERRA; http://disc.gsfc.nasa.gov/SSW/) and the European Centre for Medium-Range Weather Forecasts (ECMWF) Interim Reanalysis (ERA-Interim; http://data-portal.ecmwf.int/data/d/interim_full_daily). These products were applied using the nearest neighbour algorithm. The land cover maps used were the MCD12Q1 product and the Global Land Cover 2000 (GLC 2000). The MCD12Q1 was land cover data with a spatial resolution of 500 m and a yearly time step obtained by the MODIS sensor on board of the TERRA and AQUA satellites (Friedl et al., 2010). Among the available classification schemes, the 17-class International Geosphere Biosphere Programme (IGBP) was adopted. The GLC-2000 project was implemented by the Joint Research Centre (JRC) of the European Commission (EC) in partnership with more than 30 partners institutions using the Satellite Pour l’Observation de la Terre (SPOT) VEGETATION one km satellite data (Giri, Zhu, & Reed, 2005). The two land cover maps (GLC2000 and MODIS) were reclassified following Giri et al. (2005) for comparison with the categories in the eddy covariance sites ancillary data.

The IPAR gridded dataset (eight days time step and one km spatial resolution) was obtained with MCD15A2, in which the IPAR values were derived through the inversion of a radiative transfer model using the MODIS data acquired by satellites TERRA and AQUA as inputs. For each MODIS product (MCD15A2 and MCD12), quality control was conducted with specific quality flag layers, and low-quality pixels were discarded. The IPAR and land cover data were aggregated to a 0.5 degree spatial resolution per PFT to maintain information related to each land cover type. More specifically, for each land cover class, the fractional cover of that class, the mean value of IPAR and its variability determined by the standard deviation, were derived at a 0.5 degree spatial resolution.
2.3. Models

2.3.1. Random Forest

The modelling approach used in this study was based on the Random Forest (RF) concept. The RF concept is an improved ensemble method of decorrelated tree models (e.g., Breiman, 2001; Hastie et al., 2001; Ho, 1998). Roughly speaking, RF combines several decision trees built on different combinations of input explanatory variables (drivers), and provide as output the mean prediction of the individual trees. This strategy is very beneficial to alleviate the often reported overfitting problem of simple decision trees. Let us review the main aspects of RF. The primary property of tree models is a partitioning of the space into smaller regions to manage phenomena characterised by very complex interactions among drivers. In particular, in tree models, partitioning is recursive. This phenomenon occurs when the sub-divisions are divided again until the partitioning reduces the appropriate cost function. Recursive partitioning is terminated when the cost function cannot be further minimised; hence, a simple model, usable only for the partitioned sub-region, can be estimated.

The tree models use a tree scheme for representing recursive partitioning. Each node of the tree is a splitting rule for the inputs, except for the terminals ones (the leaves), to which a simple model is attached and applied in that cell only. For regression trees, a mean value of the target is in each leaf.

Different techniques are used to realise an ensemble of tree models. For the RF method, tree models are independently realised by “bagging”, a bootstrap aggregating technique (Hastie et al., 2001) by which m training sets of samples are randomly extracted from the overall dataset and used to parameterise m tree models. The fraction of observations not used for the parameterisation, “Out-Of-Bag” (OOB), is used to evaluate the ensemble. The decorrelation of trees is achieved in the tree-growing process through the random selection of the input variables (Hastie et al., 2001) by bootstrap methods. Thus, many of the issues with regression trees, such as overfitting (Breiman, 2001; Cutler et al., 2007), are corrected, and the risk of overfitting is reduced.

For each observation, the output of RF is the average of the outputs of the trees, and hence RFs typically yield a reduced bias of the estimations and in general good accuracies.

The Random Forest method has been applied in different areas of concern in forest ecology, such as modelling the gradient of coniferous species (Evans & Cushman, 2009), the occurrence of fire in Mediterranean regions (Oliveira, Oehler, San-Miquel-Ayanz, Camia, & Pereira, 2012), the classification of species or land cover type (Cutler et al., 2007; Gislason, Benediktsson, & Sveinsson, 2006) and the analysis of the relative importance of the proposed drivers (Cutler et al., 2007) or for the drivers selection (Genuer, Poggi, & Tuleau-Malot, 2010; Gislason et al., 2006; Jung & Zscheischler, 2013).

2.3.2. Semiempirical GPP model: MOD17 and MOD17 +

The results provided by the RF were compared to those obtained by the well-known radiation use efficiency model MOD17 (Heinsch et al., 2006; Zhao et al., 2005). In MOD17, (1) the GPP is calculated as the product of fPAR, incoming PAR and actual light use efficiency (ε):

\[
GPP = PAR \times fPAR \times \varepsilon
\]

The actual light use efficiency was obtained multiplying a maximum light use efficiency (\(\varepsilon_{\text{max}}\)), specific to each plant functional type (PFT) with two reducing factors (2) that were calculated using the minimum air temperature (\(T_{\text{min}}\)) and the vapour pressure deficit (VPD):

\[
\varepsilon = \varepsilon_{\text{max}} \times T_{\text{min}} \times \text{VPD}
\]

For this study, we used two types of MOD17 outputs. The first was the standard distributed product MOD17A2 collection 5 (Zhao et al., 2005), and the second was the output produced by a version of the model in which parameters were optimised for the site level data used in this study (MOD17 +), following the example of Beer et al. (2010).

The optimisation of MOD17 + was conducted in the R environment by “R Based Genetic Algorithm” (rbga) on “genalg” package. The optimisation procedure selected a population of parameters to minimise the sum of squares of the model residuals. The MOD17 + parameters were estimated per PFT; more specifically, after 50 iterations, 100 best vectors of MOD17 + parameters were selected, and the median values were calculated.

Other models based on remote sensing input were available such for example the Vegetation Photosynthesis Model (VPM, Xiao et al., 2004) and the Temperature Greenness model (TG, Sims et al., 2008). For the scope of this paper, it was sufficient to select only one model to be used as reference; however, a cross-validation activity based on optimised version of these additional models was performed and results showed very similar performances (data not shown).

2.4. Experimental setup

Ten versions of RF were trained with different combinations of four groups of drivers: vegetation indices (VIs), meteorological data (METEO), fraction of absorbed photosynthetically active radiation (fPAR) and plant functional type (PFT). The VIs dataset was composed of vegetation indices NDVI, EVI, and NDWI, daily land surface temperature and the top of atmosphere (TOA) potential radiation, which was calculated based on the geographical position of each site. The METEO dataset was composed of the measured incoming solar radiation, air temperature, VPD and precipitation.

A binary code was associated to each version of RF to indicate which group of drivers was used. The order of the drivers in the binary code was bit four (‘1000’ for VIs, bit three (‘0100’) for METEO, bit two (‘0010’) for fPAR and bit one (‘0001’) for PFT, with ‘1’ indicating the use of the drivers. For example, the model RF1000 considered as input only vegetation indices, whereas the model RF0111 used meteorological data, fPAR and PFT. The models codes and the input data groups are presented in Table 2.

2.4.1. Analysis of model performances and uncertainties at the site level

The site level cross-comparison of RF was conducted with the leave-one-out strategy: more specifically one site at time was excluded by the training of RF and then used for the models evaluations by predictions. The capability of the models to predict the overall time series, the differences among sites (i.e., the mean GPP of each site), the seasonal patterns

<table>
<thead>
<tr>
<th>Model code</th>
<th>VIs</th>
<th>METEO</th>
<th>fPAR</th>
<th>PFT</th>
<th>Input datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF1000</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>VIs</td>
</tr>
<tr>
<td>RF0100</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>METEO</td>
</tr>
<tr>
<td>RF0110</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>METEO, fPAR</td>
</tr>
<tr>
<td>RF1100</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>VIs, METEO</td>
</tr>
<tr>
<td>RF1110</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>VIs, METEO, fPAR</td>
</tr>
<tr>
<td>RF0101</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>VIs, PFT</td>
</tr>
<tr>
<td>RF0111</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>METEO, fPAR, PFT</td>
</tr>
<tr>
<td>RF1101</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>VIs, METEO, PFT</td>
</tr>
<tr>
<td>RF1111</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>VIs, METEO, fPAR, PFT</td>
</tr>
</tbody>
</table>
and the anomalies was analysed following Jung et al. (2011). More specifically, the mean seasonal cycle (MSC) per site was calculated with the averaged values for each eight days period across all available years, but only when at least two values (i.e., years) for each eight days period were available. To assess the among sites variability, the mean value for each site was calculated as the mean of MSC if at least 23 of the 46 values of the MSC were present, whereas the anomalies were calculated as the deviation of a flux value from the MSC. Finally, the mean site values were removed from the MSC.

For each analysed time series component (overall, among sites, seasonality and anomalies), a ranking of the different RF approaches (different input datasets) was realised using Pearson’s linear correlation coefficient between the predicted values and the ones estimated by eddy covariance. Then, RF performances using only remotely sensed data were compared with the best RF result from the ranking analysis and with the ones of MOD17, both the standard product and that optimised at the site level (MOD17 +). This specific focus was useful to estimate the distance between the accuracy of the best predictions and that obtained using only remote sensing measurements.

The evaluation of eight days GPP predicted by RF at the site level was completed with the analysis of the accuracy of the prediction per ecosystem type. This evaluation was conducted because the capability to predict GPP for some ecosystems could be affected by the absence of important drivers, particularly if they were representative of specific seasonality and anomalies. A ranking of the different RF approaches was realised with Pearson’s linear correlation coefficient between the predicted values and the ones estimated by eddy covariance. Then, RF performances using only remotely sensed data were compared with the best RF result from the ranking analysis and with the ones of MOD17, both the standard product and that optimised at the site level (MOD17 +). This specific focus was useful to estimate the distance between the accuracy of the best predictions and that obtained using only remote sensing measurements.

The estimated values of RMSE per model and PFT were used as the input to a one-way analysis-of-variance (ANOVA) with a blocking effect (the driver datasets were considered “treatments”, and the PFT was considered a “blocking factor”) to establish if the differences between models and PFT were statistically significant and to determine which models were the best for each ecosystem type.

### 2.4.2. Analysis of uncertainties at a European spatial scale

When an empirical or statistical model such as the one applied in this work uses data that are not directly measured but coming from another modelling exercise (e.g., meteorological data or land cover maps), the uncertainty associated with the production should be considered. This effect is typically referred in statistics and remote sensing and geosciences as uncertainty propagation (O’Hagan, 2012).

The uncertainty of the up-scaled mean annual GPP of Europe was analysed for the uncertainties introduced by meteorological grids and land cover maps. RF was applied to different meteorological and land cover gridded products combinations (Table 3). The prediction of GPP was performed for the period 2006–2008 at an eight days time step then aggregated at yearly time step: the yearly outputs, produced by different gridded product combinations, were used to estimate both the mean GPP over the European domain and the uncertainties originated from the modelled grids. A two-way analysis of variance (ANOVA) was used to find the pixels where the use of different land cover maps and meteorological gridded products determined statistically significant differences in predictions.

The results were also compared with those obtained by nonparametric Friedman’s and Kruskal–Wallis statistical tests but not important differences were observed (see Appendix A).

Finally, the differences between the European GPP predicted by each one-product combination and their mean (used as reference), were analysed and compared with the differences between modelled drivers (meteorological and land-use) in order to explore their roles and relationships with the uncertainty of output.

### 3. Results and discussions

In this section we analysed the main findings of our work. We first provided an analysis of the accuracy at site level and at ecosystem type. Then, we investigated the uncertainty of prediction at European spatial scale and the relationships between uncertainty of predictions and the ones of the modelled drivers’ dataset.

#### 3.1. Analysis of model performances and uncertainties at the site level

The ranking of Pearson’s correlation coefficients for the different tests (among site comparisons, anomalies, seasonality and overall time series) is presented in Fig. 1. The models that used VI as input were among the best (Fig. 1, red bars) when they were also associated with meteorological data.

In contrast, when VI were not used, the models were among those with a lower ρ (Fig. 1, yellow bars), particularly RF100 and RF101, in which the fPAR was not used (Fig. 1, green bar). The RF driven by only remotely sensed measured data (Fig. 1, blue bars) were ranked in intermediate positions, with ρ closer to the best models. This result confirmed not only the critical role of remotely sensed data in the GPP empirical upscaling, but also the importance of meteorological data in describing the dynamics both of the driving forces (solar radiation e.g.) and the seasonal limiting factors (e.g., drought and water availability).

The contribution of PFT as an input to a model was minor, and the performance of models that differed only in the use of additional input (RF1000 vs RF1001, RF1000 vs RF1010, RF1001 vs RF1111) was very similar, with slightly better results obtained when PFT was used. This result demonstrated that the PFT information improved the performance of models, though this improvement was limited.

In Fig. 2 were shown the cross correlation between the eight days GPP derived from eddy covariance and predictions by RF (the best predictions and RF with only measured remote sensing variables) and MOD17 (standard and optimised).

From Fig. 2, time series components were predicted well by RF, except for the anomalies. In fact the RMSE of the eight days time step estimated from the best RF predictions was 1.8 g C m·d⁻¹ (Fig. 2) and ρ was approximately 0.84 for the overall time series. High accuracy were also obtained for the among sites evaluation (ρ of 0.76 and RMSE of 0.83 g C m⁻² d⁻¹) and for the seasonality with a RMSE of 1.03 g C m⁻² d⁻¹ and ρ of 0.93. However, the anomalies were not predicted well, with a RMSE of approximately 1.09 g C m⁻² d⁻¹ and a ρ of 0.54. The low ρ confirmed that prediction of anomalies was a difficult task because the prediction was largely dependent on factors (e.g., management, site history, soil carbon pool) that were not directly represented in the drivers. The importance of measured remote sensing variables in the spatial up-scaling of GPP was confirmed by the model based only on that variable (RF1000), which also had high performance, close to that of the best model. For the overall time series, the RMSE of RF1000 was 1.9 g C m⁻² d⁻¹, with a ρ of 0.81, whereas for seasonality, the RMSE was 1.12 g C m⁻² d⁻¹ and the ρ was 0.91. Additionally, similar to the best predictions were the performances for the among-sites (ρ of

<table>
<thead>
<tr>
<th>Combination</th>
<th>Meteorological grid</th>
<th>Land cover grid</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERRA MCD12</td>
<td>MERRA (GMAO)</td>
<td>MCD12</td>
</tr>
<tr>
<td>ERA MCD12</td>
<td>ERA-Interim</td>
<td>MCD12</td>
</tr>
<tr>
<td>MERRA GLC2000</td>
<td>MERRA (GMAO)</td>
<td>GLC2000</td>
</tr>
<tr>
<td>ERA GLC2000</td>
<td>ERA-Interim</td>
<td>GLC2000</td>
</tr>
</tbody>
</table>
0.75, RMSE of 1.01 g C m$^{-2}$ d$^{-1}$) and anomalies ($\rho$ of 0.42, RMSE of 1.23 g C m$^{-2}$ d$^{-1}$). About the anomalies, the performance of RF1000 was very close to the best predictions and was significantly better than models operating primarily with meteorological drivers. These results suggest a minor role of the meteorological drivers at least in the simulation of the average eight days aggregated GPP. In fact, it is possible that the fast or instantaneous cause–effect relationships between photosynthetic activity and meteorological drivers become less important when the data are analysed at eight days time resolution, suggesting that other slow responding drivers gain importance and that those drivers are better represented by remotely sensed data such as vegetation indexes. In addition, it is important to consider that some of the meteorological drivers of GPP are also important drivers for phenology (e.g. incoming solar radiation or temperature) and for this reason in a

![Fig. 1. Performances of RF at the site level. Each bar indicates a $\rho$ value between GPP predicted by RF and estimated by eddy-covariance towers. Among site comparison, seasonality, anomalies and the overall time series were explored. RF1100, RF1110, RF1101 and RF1111 (red bars) used both meteorological data and VI as input, RF0100 and RF0101 (yellow bars) used meteorological data as a driver, RF0110 and RF0111 (green bars) used meteorological data and PAR from a satellite as drivers, and RF1000 and RF1001 (blue bars) used only VI as input. The continuous black lines identify models in which PFT was used (RF1001, RF0101, RF111, RF0110 and RF1111). Vertical continuous black lines indicate the low and high $\rho$ values across models. Vertical dotted black line indicates $\rho$ value for RF1000 (driven by only remote sensing measured data). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)](image1)

![Fig. 2. Cross comparison scatter plots between model output (x-axis, g C m$^{-2}$ d$^{-1}$) and reference GPP derived by eddy covariance (y-axis, g C m$^{-2}$ d$^{-1}$). First column: the results of the best RF in accordance with the ranking in Fig. 1. Second column: the results of the RF driven by only measured remotely sensed data (RF1000). Third column: standard MOD17A2 product extracted at the site location. Fourth column: MOD17+ with parameters optimised using the eddy covariance data.](image2)
machine learning tool the information is already included in the remotely sensed input.

The GPP anomalies were affected both by physiological reactions of green ecosystems to meteorological drivers and by the interannual variability of seasonal greenness. The results suggested that this last component, representing interannual variability of photosynthetically active vegetation density, might be more important than the meteorological drivers at the time resolution of the study. However, the generally poor results in the anomalies simulations confirm the difficulties in the prediction of year-by-year GPP dynamic respect to the multi-year average. This could be due, on the one hand, to some key driver that could be missing (e.g. related to land use history, soil dynamic or management) and, on the other hand, to the uncertainty in the eddy covariance derived GPP that could be larger when analysed year by year respect to when it is averaged to multi-annual means.

The performances of the RF were very close to the performance of MOD17 operating with parameters optimised using the same eddy covariance data (MOD17+); this model showed similar but slightly worse results compared with both RF in Fig. 2 (the best model and the one driven only by VI5). For MOD17+, the \( \rho \) values were 0.79, 0.89 and 0.64 for overall, seasonality and among site comparisons, respectively, whereas the RMSE was approximately 2.04, 1.24 and 1.05 \( \text{g C m}^{-2} \text{d}^{-1} \), respectively. Similar to the RF, the anomalies were not predicted well by MOD17+; the \( \rho \) was 0.4 and the RMSE was 1.27 \( \text{g C m}^{-2} \text{d}^{-1} \).

In particular, the RF1000 predictions very close to the best RF and comparable with the performances of MOD17+ optimised at the site level, should encourage future efforts with the application of machine learning methods such as RF, with only measured remote sensing data to upscale GPP. The MOD17 standard product had the highest RMSE and the lowest correlation. The comparative results of MOD17+ with the best RF suggest the requirement of MOD17 parameter optimization for better GPP prediction using MOD17.

3.1.1. Analysis of accuracy assessment per ecosystem type

A more detailed investigation of the accuracy assessment of GPP by RF was conducted to analyse the uncertainties of output produced at the site level for each model and ecosystem type. Fig. 3 summarises the performance of the RFs evaluated per PFT; \( \rho \) and RMSE were used as metrics. Notably, the primary differences in the accuracy of predictions were among PFTs and not among models, which highlighted that some ecosystems were less predictable than others. However, there were also differences among the models, with generally better results for those using remote sensing inputs.

The correlation coefficients were generally high for all models and PFTs, with the exception of evergreen broadleaf forests (EBF), for which the \( \rho \) was significantly lower and the relative RMSE significantly higher. The models with the lowest performance in this PFT were the ones driven by VI5 (RF1000 and RF1001); this low performance was most likely because of insufficient information from the remote sensing data on the typical summer reduction of GPP caused by drought and warming stress.

In contrast, models driven by only VI5 dataset as input had good performances in PFTs characterised by high variability of green biomass density during the vegetative period (e.g., deciduous broadleaf forests) or by human management (e.g. croplands). For croplands (CRO), all models produced high RMSEs, most likely because of the heterogeneity in management practices, which were different among sites and not described by the meteorological inputs. Nevertheless significant improvements there were for the models operating with VI5 because carrying the information of the changed amount of green vegetation.

These results confirm that at the time resolution used in this study, most of the information about GPP and its dynamic are included in the remotely sensed data that are also directly or indirectly correlated to the meteorological drivers. In fact, in vegetation types where the greenness seasonality is not following any of the meteorological drivers (Evergreen broadleaf forests EBF) errors were significantly higher.

The one-way ANOVA confirmed that the significant variability of RMSEs at the eight days time step was related both to the blocking effect by PFTs \( (p < 0.001) \) and to the dataset used as a driver of GPP \( (p < 0.05) \). However, the multiple comparisons of means (the Tukey test with a significance level of 95%) indicated that RF0100 (a model that used only meteorological information) was significantly less accurate than other models. If the RF0100 was excluded, significant differences among models were not found.

3.2. Analysis of uncertainties at a European spatial scale

To estimate the effects of the uncertainties of the modelled inputs on the European GPP, the RF0111 was applied. This model used as input only modelled data as a grid: meteorological re-analysis, PFT classification and fPAR derived from remote sensing. The uncertainty of RF0111 estimated at the site level with a yearly time step was not significantly different from the best models, thus justifying the use of the model in the analysis.

In the following sections, the uncertainty of the GPP predicted by RF0111 at a European spatial scale was analysed. In particular, in Section 3.2.1, we analysed both the mean annual European GPP predicted by RF0111 and its uncertainty because of the modelled inputs. In Section 3.2.2, we analysed the importance of the two kinds of modelled input (meteorological grids and land cover maps) on the uncertainty of predictions. Finally, in Section 3.2.3, we analysed how the divergence between products has conditioned the uncertainty of predictions.

3.2.1. GPP predicted at a European spatial scale and uncertainty constrained by modelled drivers

The annual European GPP predicted by RF0111 is shown in Fig. 4, which was calculated as the average of the results obtained using as input the four combinations of the available meteorological and land mapping.
Fig. 4. The mean annual GPP predicted by RF0111 at European spatial scale using four different combinations of meteorological and land cover grids (products), then calculating the mean (left). The spatial distribution of the relative uncertainty (%), estimated as the ratio between the standard deviation of the mean annual predictions (each one of them obtained from a specific combination of land cover and meteorological products) and their mean value (right).

Fig. 5. Maps of the intrinsic uncertainties of the models at European spatial scale for RF1000 (uncRF1000, a), RF0111 (uncRF0111, b), and their difference uncRF0111–uncRF1000 (c). Additionally, map for the uncertainties of the RF0111 predictions, introduced by the modelled drivers (uncRF0111*, d), and the boxplot of the intrinsic uncertainties of the models (RF1000 and RF0111) compared with the uncertainties of the predictions (by RF0111) introduced by the use of the modelled drivers (e).
cover products. Primarily, the west of Europe and in particular along the Atlantic coast of the Iberian Peninsula, the west of France and the Italian Peninsula were the areas with highest GPP. High annual GPP was estimated for the southwest of the United Kingdom (UK), the middle of France, in the central Europe (particularly in Germany) and along the Black Sea coast of Turkey. In contrast, low annual GPP was predicted for the North Africa, the southeast of the Iberian Peninsula, particularly in Mediterranean regions, the Anatolian Plateau, and the Scandinavian regions. This spatial pattern was similar to the maps produced by Jung et al. (2008), who applied the process model LPJ and four different diagnostic methods: ANN (Papale & Valentini, 2003), MOD17 + (Jung et al., 2008), FAPAR-based productivity assessment (FPA), and FPA including land cover (FPA + LC) (Jung et al., 2008). The spatial pattern was also similar to the spatial pattern of annual GPP predicted by Beer et al. (2010) and Jung et al. (2011) for the European countries.

The mean European GPP was 1293 (±136–202) g C m⁻² y⁻¹, which was comparable with estimations by other studies. Jung et al. (2008) found that GPP from different data driven approaches ranged from 900 g C m⁻² y⁻¹ to 1100 g C m⁻² y⁻¹ and Jung, Vetter, et al., 2007, with the process models BIOME-BGC, LPJ and ORCHIDEE, reported a mean yearly GPP between 783 and 1042 g C m⁻² y⁻¹. If compared with those studies, RF0111 overestimated European GPP, but the low productivity regions of the north and the east of Europe were not included in the RF0111 simulations. The GPP predicted by RF0111 varied between 440 and 2480 g C m⁻² y⁻¹, results that were comparable with the estimation reported by Beer et al. (2007), who applied a semiempirical watershed derived water use efficiency (WUE) model to estimate a mean GPP that ranged between 1393 and 2071 g C m⁻² y⁻¹.

The uncertainty of annual GPP due to modelled drivers (Figs. 4 and 5d) and estimated pixel-by-pixel as the standard deviation of the mean annual GPP predicted with different gridded products combinations, varied between 0.3 and 437 g C m⁻² y⁻¹, from 4% to 44% of the predicted GPP while the average value was 72 g C m⁻² y⁻¹, 6% of the predicted GPP. The effects of the meteorological and land cover uncertainty were particularly noticeable in the south of Europe, in the Mediterranean region that included the North Africa, the Italian Peninsula, the Hellenic and the Iberian Peninsulas, and the northwest of Turkey, often in correspondence with regions characterised by low predicted GPP (Figs. 4b and 5d). Other regions with relatively high uncertainties were found in the southwest of Scandinavia. In contrast, large areas of France and central Europe, characterised by high estimated productivity, were less sensitive to the drivers used. Because the RMSE of RF0111 at the site level was 165 g C m⁻² y⁻¹ (136–202 g C m⁻² y⁻¹), very close to the best RF model (RF1111) in which the site level RMSE at yearly time step was 152 g C m⁻² y⁻¹ (127–189 g C m⁻² y⁻¹), the importance of the uncertainty caused by the modelled input was clear. This additional uncertainty could be reduced by using as input only directly measured data such as remote sensing reflectances, vegetation indices, or land surface temperatures because the input uncertainties were similar both for training and for upscaling phases. To explore this hypothesis, the uncertainties of the models RF1000 (uncRF1000) and RF0111 (uncRF0111) were upcaled from the site level to the European spatial scale and were compared for the uncertainty of predictions caused by the use of different modelled gridded products (uncRF0111*). The upscaling of models uncertainty was conducted by calculating for each pixel the weighted mean RMSE on the basis of the PFT fractional cover. The results are shown in Fig. 5a and b. The spatial pattern of uncRF0111 and uncRF1000 maps was similar, with values of uncRF0111 slightly lower than for uncRF1000. The spatial pattern of the two maps (Fig. 5a and b) highlighted low uncertainty in Scandinavia and on the Anatolian Peninsula, which were also regions with low estimated GPP. High values of model uncertainties occurred on the Italian and Iberian Peninsulas, in France, in Germany, and in Central and East Europe. The latitudinal trend of models uncertainties was similar to the one for the GPP predicted by RF011, with values higher at southern and middle latitudes than at northern latitudes. By contrast, in the east–west direction, no trend of models uncertainties was discernable, however, the predicted annual GPP was significantly lower at eastern than at western longitudes. The differences between uncertainties of the two models (uncRF0111 and uncRF1000, Fig. 5c) were in a narrow range of values, with a slightly negative mean, and were uniformly distributed across Europe. However, two regions with high differences between the uncertainties of the two models were found. The first, characterised by higher uncertainty of RF0111 than RF1000, was the Ireland and the west of the UK, and the second was the northeast of Europe where RF1000 produced clearly higher uncertainty. The prevalence of slightly negative values in Fig. 5c indicated that RF0111 was somewhat more accurate than the RF driven only by measured remotely sensed data (RF1000). Moreover, the mean value of unRF1000 across Europe was 138 g C m⁻² y⁻¹, with a range of 85–188 g C m⁻² y⁻¹, whereas for uncRF0111, the value was 119 g C m⁻² y⁻¹, with a range of 72–147 g C m⁻² y⁻¹ (Fig. 5e). Nevertheless, when the error caused by uncertainty in the modelled inputs was included (Fig. 5d, uncRF0111* in the predictions, the final uncertainty the output of RF0111 might change. The uncertainty of RF0111 significantly increased, particularly in correspondence with the Atlantic coast, Mediterranean regions, southern Europe, southwestern coast of Scandinavia, and western Ireland and UK, where the uncertainty

---

![Fig. 6. Map of pixels in which the ANOVA test has identified a statistically significant variability of the outputs (a) and the distribution of the uncertainties among the recognised factors (b). NS was used for the pixels not significant at the ANOVA test, LC for the pixels significant only for the use of different land cover products, MT for the pixels significant for the use of different meteorological data, LC + MT, for the pixels significant both for the use of different land cover and meteorological data but with an additive effect, LC × MT for the pixels significant for the interactions of the two factors.](image-url)
caused by different versions of the drivers was high (orange and red regions in Fig. 5d). By contrast, the effects were small in the middle and central Europe, on the Scandinavian Peninsula and in the middle-East.

The distributions of the uncertainties of the two models at a European spatial scale (uncRF1000 and uncRF0111) and the uncertainties of the outputs of RF0111 caused by different modelled drivers

---

**Fig. 7.** Maps of differences between mean annual GPP (g C m$^{-2}$ y$^{-1}$) predicted by each product combination and their mean.

**Fig. 8.** Maps of differences between meteorological products, calculated as the mean of the day-by-day differences ERA-Interim – MERRA. The green areas indicate where the ERA-Interim product was lower than MERRA, whereas red regions indicate the opposite. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Table 4
Mean representativeness of PFTs in MCD12 and GLC2000 across European lands and their differences.

<table>
<thead>
<tr>
<th>PFT</th>
<th>MCD12 (%)</th>
<th>GLC2000 (%)</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBF</td>
<td>2.41</td>
<td>8.64</td>
<td>−6.23</td>
</tr>
<tr>
<td>ENF</td>
<td>9.20</td>
<td>10.85</td>
<td>−1.65</td>
</tr>
<tr>
<td>CRO</td>
<td>21.67</td>
<td>18.45</td>
<td>3.23</td>
</tr>
<tr>
<td>MF</td>
<td>7.77</td>
<td>4.89</td>
<td>2.88</td>
</tr>
<tr>
<td>GRA</td>
<td>3.22</td>
<td>3.43</td>
<td>−0.21</td>
</tr>
<tr>
<td>EBF</td>
<td>0.02</td>
<td>0.05</td>
<td>−0.03</td>
</tr>
<tr>
<td>Others</td>
<td>55.70</td>
<td>53.69</td>
<td>2.01</td>
</tr>
</tbody>
</table>

(uncRF0111∗) are shown in Fig. 5e. The boxes show that although the uncertainties caused by the drivers were generally lower than the ones caused by the models, their contributions were still relevant and in some cases were very high (Fig. 5e). This finding should encourage future efforts for the use of only remotely sensed measured data for upsampling of carbon fluxes. This solution might have an impact on the model capacity to reproduce GPP dynamics because of the missing of key information, but the uncertainty of modelled drivers would be avoided.

3.2.2. Relative importance of meteorological and land cover uncertainties

To disentangle the role of the two drivers in the uncertainty of predicted GPP, the RF0111 simulations were submitted to a pairwise comparison. The mean absolute uncertainty caused by the use of different meteorological products and the same land cover was 100.05 g C m⁻² yr⁻¹ (range 0.32–753.13 g C m⁻² yr⁻¹), whereas the mean absolute uncertainty caused by the different land cover products was 47.15 g C m⁻² yr⁻¹ (range 0–754.41 g C m⁻² yr⁻¹). Although the ranges of uncertainties were approximately the same, if these uncertainties were expressed in relative terms as the ratio between uncertainty and mean annual European GPP, they were 8% (range 0.0003–76%) for the meteorological data and 3.5% (range 0–41.7%) for the land cover maps. Thus, the areas with higher absolute uncertainty caused by the land cover products occurred in pixels with high GPP, whereas the uncertainty of the meteorological drivers was more homogeneously distributed, covering also areas with low GPP.

The important role of the quality of meteorological grids on modelled GPP was also highlighted by Jung, Vetter, et al. (2007), who found a difference of 20% between overall GPP predicted by a reference setup and a change in the meteorological gridded data product. Zhao, Running, and Nemani (2006) found similar results in a study on the effects of the reanalysed meteorological data on the global GPP and NPP estimated by MOD17. In that study, the largest difference occurred between predictions based on the National Centres for Environmental Prediction/National Centre for Atmospheric Research (NCEP/NCAR) dataset and the European Centre for Medium-Range Weather Forecasts (ECMWF ERA-40) dataset and was approximately 23 GtC yr⁻¹, whereas the relative error for the GPP ranged from 16% (ECMWF ERA-40) to 24% (NCEP/NCAR).

Thus, the possible conclusion was that uncertainties in land cover products had a secondary importance compared with the meteorological data. Jung, Vetter, et al. (2007) also found that uncertainty in the GPP predicted by a process model driven with two different land cover classification schemes was less than 10%, approximately.

Regions where the use of different meteorological or land cover products led to a significantly different mean annual GPP were identified by the two-way ANOVA test, conducted pixel by pixel. The results (Fig. 6) indicated that the regions where the differences in GPP among products were statistically significant were concentrated in the south of Europe, in the middle-east and in the north of Europe (Scandinavia and UK).

Additionally, for the spatial extent, the most important factor in the uncertainty of GPP was the meteorological dataset, which was statistically significant for approximately 3.0 × 10⁶ km² (Fig. 6, yellow area), whereas land cover misclassification was significant for an area of 6.9 × 10⁵ km² (Fig. 6, cyan area). The magnitude of the effects of land cover and meteorological data were similar (Fig. 6b), with a mean induced uncertainty of 89 g C m⁻² yr⁻¹ for land cover divergence (Fig. 6b, LC) and 94 g C m⁻² yr⁻¹ for differences in meteorological products (Fig. 6b, MT), and ranges that varied from 13 to 335 g C m⁻² yr⁻¹ and from 12 to 434 g C m⁻² yr⁻¹, respectively. Jointly, meteorological and land cover factors were statistically significant for an area of 1.23 × 10⁶ km² of which 1 × 10⁶ km² was by an additive effect (Fig. 6, orange area) and 2.3 × 10⁵ km² was by their interaction (Fig. 6, red area).

The boxes in Fig. 6b show that where both land cover and meteorological uncertainties were statistically significant there was an increase in the estimated uncertainties. For the pixels in which the combined effect was additive (Fig. 6b, LC + M), the mean value of uncertainties was 122 g C m⁻² yr⁻¹. For the pixels in which the interactions between meteorological and land cover factors was statistically significant (Fig. 6b, LC * M), the mean yearly uncertainty was approximately 193 g C m⁻² yr⁻¹.

We also used alternative nonparametric tests such as Friedman’s and Kruskal–Wallis tests of significance, and the results were similar. (see Supplementary material, Appendix A).

3.2.3. Analysis of roles and relationships between drivers and GPP uncertainties

The effects of the different driver combinations on the upscaled GPP are shown in Fig. 7 and significant differences between predictions provided by each driver combination and the mean annual GPP were found. The spatial pattern of the uncertainties was highlighted in the Mediterranean regions and middle-eastern Europe. The mean annual GPP produced using meteorological data from the ERA-Interim was higher than the one obtained using MERRA in Southern (e.g. Mediterranean regions) and Middle-East Europe. In Central Europe, the west of the UK

Fig. 9. Maps showing for each 0.5 × 0.5 degree aggregated pixel the PFT where the two land cover products MCD12 and GLC2000 show higher differences in the % of coverage (a) and the values of the difference calculated as %_MCD−%_GLC (b).
and Ireland, at northern latitudes and in cold areas at lower latitudes (e.g., Alps Mountains), the output produced with ERA-Interim grids led to lower GPP compared with the average and MERRA.

The only area where the land cover product was likely responsible for the difference was the Iberian Peninsula, where the GPP estimated using GLC2000 was greater than the GPP estimated using the MCD12 product.

This pattern could be explained by the analysis of the primary differences between the drivers.

By Fig. 8, showing the mean differences among meteorological products, it was noticeable that the shortwave global incoming radiation and VPD were systematically lower in the ERA-Interim product compared with MERRA, whereas for precipitation, the values were generally greater in the ERA-Interim product than in MERRA. For air temperature, the distribution of divergences was complex, with greater values for MERRA in North Africa, along the coast of Turkey, on the Iberian Peninsula and in north Scandinavia. By contrast, air temperatures of the ERA-Interim were greater than MERRA in Italy, near the Carpathian Mountains, in the north east of Europe and in the southern half of Scandinavia.

The systematic bias between ERA-Interim and MERRA, particularly for VPD and shortwave incoming solar radiation, could explain the differences in GPP products. The higher GPP in the west UK, cold regions and in northern Europe when MERRA was used, could be related to the higher incoming radiation, whereas the higher values of VPD in the same areas did not have any effect because water availability was not a limiting factor in those environments. The opposite result was the higher productivity estimated by ERA-Interim at lower latitudes (water limited environments such as Italy, North Africa and middle East), which could be explained by the systematically lower value of VPD in the ERA-Interim products than in MERRA.

The differences in the land cover classification that led to the uncertainty when this input was used are summarised in Table 4 and Fig. 9. Table 4 summarises the presence of the different PFTs in Europe as extracted by the applied products and the differences. The most important differences in the mean value of fractional land cover across Europe were found for DBFs, which were more represented in GLC2000 than MCD12 and, in contrast, for CROs and MFs that were more present in MCD12 than GLC2000. Less important was the difference for ENF, which was present more in GLC2000, and though very small could account for the differences with GRA and EBF.

The data reported in Table 4 confirmed the findings of Giri et al. (2005), who described a larger degree of presence of forests for GLC2000 at global scales. This relative abundance was also associated with a different definition of forests in the land cover definition scheme adopted by the two land cover products. Fig. 9a shows that for each 0.5° × 0.5° degree pixel, the PFT in which the two land cover products had the higher difference; in Fig. 9b, its value was calculated as \%PFT_MCD12 − \%PFT_GL2000.

Correspondence between differences in classification and uncertainty in GPP was found on the Iberian Peninsula, where the GPP predicted by GLC2000 was significantly higher than MCD12 (Fig. 7), and the fractional land cover of CRO, ENF and MF estimated by GLC2000 was higher than MCD12 (Fig. 9, green areas). The higher GPP for the predictions by GLC2000 was also affected by the absence of a savannah-like land cover class in GLC2000. Because savannahs (present in the training set and used as a separate PFT) were characterised by relatively low GPP, this property could explain the higher GPP when the GLC2000 was used. Cases in different regions were also found where the effects on the mean annual GPP was low, although the differences in the classification were important. Cases such as these were found at north western France, in northern UK, and in eastern Europe that involved primarily CRO and MF, which was most likely because the two PFTs yielded similar total annual GPP values. Clearly, an analysis of the annual trend or interannual variability would most likely show stronger differences, but such an analysis was beyond the scope of this paper.

4. Conclusions

In this paper, it was presented the application of the Random Forests algorithm to estimate eight days GPP (at the site level) and the mean annual European budget. The results showed that RF methods were promising and comparable with other machine learning approaches published before, including MTE (Beer et al., 2010; Jung et al., 2011) and ANN (Papale & Valentini, 2003, Beer et al., 2010), or semiparametric LUE models such as MOD17 (Beer et al., 2010; Running, Thornton, Nemani, & Glassy, 2000; Sjöström et al., 2011).

The application at a European spatial scale confirmed the potential applicability of the RF method for GPP prediction; the mean annual values of upscaled product was in the range of estimations reported by other authors, both at the site level and at the European spatial scale (Beer et al., 2007, 2010; Jung, Vetter et al., 2007; Jung et al., 2007).

At the site level, using a leave-one-out cross-comparison strategy, the best results were obtained by the combination of meteorological and remotely sensed data as input (RF1111); however the differences with results obtained by the other models (that have used in input a reduced datasets) were not significant, except for the RF trained with only meteorological inputs (RF100). The models driven by only measured remotely sensed data (RF1000) gave very good results, close to the best model, which confirmed the importance of remote sensing data in the spatial upsampling exercises. However, the accuracy of predictions of the different models was variable for different PFTs. Particularly for EBF sites, the models parameterised using only remotely sensed data were unable to capture the effects of limiting factors such as drought and water stress, which were instead better identified when VPD was used as driver. Notably, when PFT was used as a driver, significant improvements in the accuracy of predictions did not occur, while the variables describing the seasonality and vegetation health (IPAR and remotely sensed vegetation indices) were fundamental to predict GPP across PFTs.

These findings were the basis for the second part of the analysis in which the role of differences in the modelled meteorological spatially explicit drivers and misclassification of land cover maps were analysed in relation to the GPP uncertainty at a European scale.

About the uncertainty of the upscaled GPP, a significant role of the uncertainties of the land cover and meteorological drivers was found, with the differences in land cover classification having a less significant role compared with those in the meteorological fields. However, a significant effect of land cover misclassification was observed where the PFTs having different capabilities (and magnitudes) of GPP were involved.

The uncertainty caused by the differences between the modelled input grids was very high, and in some regions it was greater than the intrinsic uncertainty of the models.

The modelled inputs could be avoided by using only measured remotely sensed data as drivers. However, in some PFTs, where the meteorological conditions and induced stress factors did not affect the canopy greenness, the effects on averaged GPP were not well captured by the available remote sensing data. Additionally, the uncertainties caused by specific disturbances that affected the quality of remotely sensed data such as acquisition geometry, atmospheric conditions, snow, clouds, and unmixed pixels that were not addressed in this paper could also play an important role.

With the availability of new sensors in the near future (e.g., the Sentinel constellation and the fluorescence bands currently available at coarse resolution and under evaluation for future missions like EarthExplorer 8 candidate’s mission FLEX) and the expansion of the network of site level measurements in FLUXNET, further exploration of the use empirical models based only on directly measured drivers will continue to capture fast responses of vegetation to stress factors (e.g., with fluorescence or non-photochemical quenching using PRI), to monitor disturbance and management history through multitemporal images and to quantify ecosystem stocks using radar information.
Acknowledgement

The MODIS data products were obtained from the Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC) and the Earth Observation System Data and Information System (EOSDIS). MERRA data have been provided by the Global Modeling and Assimilation Office (GMAO) at NASA Goddard Space Flight Center through the NASA Global Environmental Multiscale (GEMS) online archive. ECMWF ERA–Interim data have been provided by ECMWF data server. GLC2000 data have been provided by EU-JRC. This work used eddy covariance data acquired by the FLUXNET community and in particular by the following networks: CarboEuropeIP, CarbOptical, and CARBOMONT. The authors thank all these data providers for the high quality of measurements/products made available to the scientific community.

The authors thank Martin Jung, Markus Reichstein and Max Planck Institute (MPI) for Biogeochemistry for their support, discussions and hospitality.

Gianluca Tramontana has been supported by the EU project GEOCARBON. Kazuhito Ichii acknowledges the support by the Environment Research and Technology Development Fund (2-1401) from the Ministry of the Environment of Japan. Gustau Camps-Valls acknowledges the support of the Spanish Ministry of Economy and Competitiveness (MINECO) under project LIFE-VISION TIN2012-38102-C03-01. Enrico Tomelleri acknowledges the financial support of the GHG-Europe project. Dario Papale thanks ICOS-INWIRE for the support.

Appendix A. Analysis of the importance of meteorological and land cover uncertainties by nonparametric tests and comparison with ANOVA results

Fig. A1 and Table A1 show the primary findings concerning the identification of regions where the use of different meteorological or land cover products as drivers led to significantly different estimates of mean annual GPP, as determined by Friedman’s and Kruskal–Wallis nonparametric statistical tests.

This was performed as confirmation of the ANOVA results (Section 3.2.2) with the use of statistical tests based on different assumptions.

Similar to Section 2.4.2, the nonparametric tests were conducted pixel-by-pixel with a significance level of 0.05.

The spatial pattern of pixels in which significant effects of meteorological and land cover products were recognised is shown in Fig. A1. The results obtained with nonparametric analysis were similar to the findings by ANOVA, as shown in Fig. 6a.

In Fig. A1, the cyan pixels represent the regions in which the differences in land cover products had a significant role in the uncertainty of the GPP, and the yellow pixels represent the regions in which the differences in meteorological products had a significant role in the uncertainty of the GPP, whereas the red pixels represent the regions where both factors were statistically significant. The patterns of the two maps were similar to the pattern obtained by the ANOVA. Larger effects of divergence among meteorological products were found than for land cover products. In particular, in the Mediterranean region, in middle-eastern Europe, pixels were identified in which the divergence in land cover products caused statistically significant differences in GPP predictions were located on the Iberian and Hellenic Peninsulas, in western Anatolia, and in sparse and fragmented regions of Scandinavia and northern Europe. In these regions, and also on the Italian Peninsula and in middle-eastern Europe, pixels were identified in which both factors were statistically significant.

In Table A1, it was compared the estimated surfaces for the factors recognised by the applied statistical tests. As expected, the estimated surface per category was consistent among the methods, with results of the Kruskal–Wallis tests slightly different from the other methods.

These results also supported our expectation concerning the importance of the meteorological reanalysis uncertainty on the uncertainty of GPP, which could be considered more relevant than the land cover maps.

Table A1

<table>
<thead>
<tr>
<th>Factors</th>
<th>ANOVA (km² × 10⁶)</th>
<th>Friedman’s (km² × 10⁶)</th>
<th>Kruskal–Wallis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land cover</td>
<td>0.69</td>
<td>0.73</td>
<td>0.88</td>
</tr>
<tr>
<td>Meteo</td>
<td>3.04</td>
<td>3.07</td>
<td>2.28</td>
</tr>
<tr>
<td>Both</td>
<td>1.23</td>
<td>1.11</td>
<td>1.11</td>
</tr>
<tr>
<td>Not affected</td>
<td>4</td>
<td>4.06</td>
<td>4.65</td>
</tr>
</tbody>
</table>

These results also supported our expectation concerning the importance of the meteorological reanalysis uncertainty on the uncertainty of GPP, which could be considered more relevant than the land cover maps.

References


Fig. A1. Significant factors (land cover or meteorological products) affecting mean annual GPP variability recognised by nonparametric tests: Friedman’s (left) and Kruskal–Wallis (right). Cyan zones are significantly affected by the uncertainty of land cover products, yellow are zones significantly affected by differences between meteorological products, and red regions are significantly affected both by differences in meteorological and in land cover products. Tests were conducted at the significance level of 0.05.