Please find our point-by-point response to Reviewer 3’s comments below.

I already commented on the original submission, just because of my interest in this type of work. At that time we were working on merging two ET products with a somehow similar approach, therefore my interest in learning from this work and providing some feedback. The first results from our work are now under discussion at https://www.hydrol-earth-syst-sci-discuss.net/hess-2017-573/, in case the authors are interested in our results.

We thank the reviewer for sharing their work, it’s definitely of interest. We’ve read it and will watch its progression in HESS.

I glanced quickly through the new manuscript and the authors reply to the reviewer comments, and the paper is in a better shape now and possible nearly ready for publication. Still, I feel that there are still a few things that may perhaps be better highlighted in the final manuscript.

(1) This is just an opinion, but I see this more as an initial exercise on merging flux estimates that any offering for a merged product that the community may feel ready to adopt. The monthly time scale, period chosen, and products selected for the merging are a bit constraining the utility of that product. In particular, leaving aside broadly adopted products (e.g., PT-JPL), combining obsolete and newer versions of a product (e.g. GLEAM), or just focusing on MODIS era (i.e., 2000 onwards) are not helping here, even if the choices have been well explained.

Yes, there is no question that this approach has strengths and limitations and can always be improved, indeed some of the potential improvements for future versions were suggested in the paper. It is also true that including more component products will likely lead to a better weighted product, but as the reviewer noted we have clearly explained our motivations for producing this first, proof-of-concept version of DOLCE. As noted previously, we did build a version that included CSIRO-global and PT-JPL for the period 2000 to 2006, and while we do not show the results here, the out-of-sample experiments showed that DOLCE was in fact slightly better with these products included. We nevertheless decided that a longer time period was more important than these very minor gains (i.e. 2000-2009).

The use of both ‘obsolete and newer versions’ of a product are presumably perceived as an issue because of the potential for duplicate information to bias the averaging or uncertainty estimation process. We wish to reassure the Reviewer that the weighting approach we used is resilient to the incorporation of redundant information. Including 100 duplicates of a product (even with the addition of separate noise terms) will not result in any qualitative change in the final weighted product or uncertainty estimate – it is one of the strengths of this approach. The two copies of GLEAM were in fact sufficiently different to add predictive capacity to the final weighted product.

Expanding the time period requires incorporating a wide range of products so that different periods of time will be covered by a different combination of products (using different weights). This will eventually lead to temporal discontinuities in DOLCE, and the magnitude of these discontinuities is something that we need to explore in more detail before knowing whether they are acceptable. This is indeed something we plan to explore and hopefully adopt in future versions of DOLCE if the resultant discontinuity is reasonable. More on this issue below.
(2) The authors limit the time period and exclude products so no time discontinuities exist. However, spatial discontinuities do not seem to be a problem, as they select geographical areas with different products. So, for instance, if I am doing a spatial average over an area containing regions with different merged products, is that a valid estimate? The point I am raising is not to also remove spatial discontinuities, but to signal that no merging will possibly result in a perfect product. Perhaps being more flexible in the time compositing to allow a data-merging-richer product with more extended time coverage would be of interest.

This is indeed an interesting idea, and as pointed to above, we would like to explore how big the temporal discontinuities are for the next generation of DOLCE. To make sure we have address the reviewer’s concerns, we have extended the conclusion:

In this work, we used 6 latent heat products that have different spatial coverage and we managed to create a global product by deriving DOLCE tiers that differ in the component products used to derive them. This might have resulted in spatial discontinuities that we haven’t explored in great detail. We might also expand the time coverage of DOLCE by incorporating different component products in different time periods, however doing so could lead to large temporal discontinuities in the derived product. Merging a variety of different ET products is not straightforward and requires subjective decision-making, but we aim to explore extending temporal coverage using different combinations of component products in future iterations of DOLCE.

(3) All bias corrections for ET products that I have seen so far implied some type of climatological component, so I am still a bit puzzled about the annual-value bias correction applied here. I may be missing something, but if you compare all-stations fluxes with the corresponding product X fluxes, derive a bias number, and then bias-correct the product X fluxes, it is quite likely that at a large number of locations the resulting fluxes will be unphysical. For instance, at locations with strong seasonal cycles and very low winter values, depending on the sign of the bias correction you may get negative winter fluxes, or unrealistically high fluxes. Those fluxes are later merged, so perhaps the weights may take care of this and produce realistic merged fluxes. Perhaps the authors want to comment on this.

Yes, this is an excellent point, the potential for this be an issue is indeed real. We of course did not use the bias-corrected products individually, only in the final weighted estimate, where we can confirm that values appeared to be physically plausible. We have added text to the manuscript to make readers aware of this potential issue:

We also note that the bias correction terms were calculated from the temporal mean of the monthly bias at all the flux tower sites across the time period 2000-2009. As a result, the individual bias corrected products may show unphysical ET estimates especially in locations of strong seasonal cycle and at high latitudes especially when ET values are close to zero. However, after applying the weights to the component products, the produced values in the merged product were all realistic.
The removal of MPI from the merged product, with an original weight of close to 0.5, to produce nearly statistically identical results is a surprise. Figures 3-4 and 11-12 are nearly identical. I have to confess that I do not understand this. The removal of one of the apparently more contributing pieces of information for the merging does not seem to affect the skill of the merge product to get closer to the tower fluxes.

Yes indeed, it was a surprise for us too. The participation of MPI in the weighting only slightly improved the out-of-sample tests. Remember though that the weighting approach adopted here is a fitting exercise to match the tower fluxes, so that the product characteristics that a weight of 0.5 for MPI might produce could be effectively replicated by combining other products with different weights once MPI is removed. Recall that the weights are assigned to account for both performance differences and dependence (in the form of error covariance against tower data) between the constituent products.

Points 3 and 4 make me think that a few examples of time series at a couple of tower locations displaying the original products to be merged, the bias corrected products, the resulting merge for different configurations, etc., may have been very illustrative. The boxplots and global maps are great at conveying the large picture, but a few time series may have been very useful to pass a clearer message regarding the merging methodology.

Yes, we have no issue with this suggestion. We’ve added Figure S4 in the supplementary material and we referred to it in the manuscript:

*Fig. S4 shows examples of time series of merged products before and after bias correction with the resultant weighted product at five tower locations.*
Yes, in many cases the improvements of the weighted product over the simple mean are marginal, this is an important point to note. There are at least three very important reasons why we should use this approach, however.

First, the weighted mean is insensitive to the addition of redundant information, unlike the simple mean. This helps us alleviate concerns such as the one the Reviewer raises above – the inclusion of two versions of GLEAM that we suspect contain a great deal of duplicate information. In-sample (i.e. with the sites used to train the weights) the weights represent an analytical solution to the best linear combination fit of products, so are by definition better than any other possible linear combination. The addition of more products can only improve the final result in this case – redundant information in the form of duplicate products has no effect. We test the ability of the weights out-of-sample of course (at sites not used in training), so technically this may no longer be the case.

Second, we use the weighting approach to try to tell whether flux tower data at the point scale offers any information about grid box scale fluxes. To do this, we check whether the (somewhat minor) improvements that the weighting offers in-sample (at the sites used for training) are still evident out-of-sample (at sites not used in training). If they do, as was generally the case, it tells us that the nature of component product biases and error covariances is consistent across sites. This can only be the case if point based tower data has information about the grid box scale.

Finally, using the weights allows us to create uncertainty estimates that are considerably more meaningful than simply using the range or standard deviation of component products. We can use the weights to ensure that we have a time and space varying uncertainty estimate that is consistent with the weighted product’s error variance against the tower data.

We have nevertheless included a figure that explores the Reviewer’s suggestion, and outlined these three motivations in more details in the manuscript:

(6) Perhaps my largest concern comes from glancing through the first column of Figures 3-4, and 11-12 and noticing again how little improvement the merged product seem to offer compared with the simple mean product. MSE, COR, and MSRD mean improvements over the different sample selections are close to the zero line. I would speculate (I may be wrong) that if we plotted differences of the weighted-mean and simple mean products with respect to common products (e.g., MPI and LandFlux, as in the paper), the pattern in the differences would be quite close. In other words, I think it is a valid question to ask if all the trouble of merging the products with the weighted methodology are justified given the close performance of the weighted mean. Based on this work and my own research, I am still not convinced. The authors stated in the conclusion, “the DOLCE product performs better than any of its six constituent members overall”, but based on the boxplot results it could be that the simple mean also do so. Perhaps this could have been explored a bit more.

Figure S4: Time series at five sites showing the products that were used to derive DOLCE tier1 before bias correction (right) and after bias correction with the resulting merge (DOLCE) (left).
The weighting applied here adds only a slight improvement to the simple mean of the merged products. Figure S5 displays the seasonal mean difference ET between (a) MPIBGC and DOLCE tier 1 and (b) MPIGBC and the simple mean of the six component products of DOLCE tier 1. The same spatial mask was applied in the plots and the differences were computed as in Fig. 6. The plots show that the differences exhibit similar patterns with stronger magnitude in the simple mean case. This might raise the question of whether it is worth applying a weighted approach to merge ET. There are however at least three compelling reasons to use this approach: (a) it removes the possibility of similar component products with redundant information biasing the mean (b) it allows us to confirm that point-scale flux tower measurements do indeed contain information about fluxes at the grid box scale, and (c) it allows us to create an uncertainty estimate that better utilises the information in the component products than simply using their range.
Figure S4: Difference between the seasonal global ET (a) from MPIBGC and DOLCE and (b) from MPIBGC and the simple mean of the merged products, calculated over the period 2000–2009.
Abstract. Accurate global gridded estimates of evapotranspiration (ET) are key to understanding water and energy budgets, as well as being required for model evaluation. Several gridded ET products have already been developed which differ in their data requirements, the approaches used to derive them and their estimates, yet it is not clear which provides the most reliable estimates. This paper presents a new global ET dataset and associated uncertainty with monthly temporal resolution for 2000–2009. Six existing gridded ET products are combined using a weighting approach trained by observational datasets from 159 FLUXNET sites. The weighting method is based on a technique that provides an analytically optimal linear combination of ET products compared to site data, and accounts for both the performance differences and error covariance between the participating ET products. We examine the performance of the weighting approach in several in-sample and out-of-sample tests that confirm that point-based estimates of flux towers provide information at the grid scale of these products. We also provide evidence that the weighted product performs better than its six constituent ET product members in four common metrics. Uncertainty in the ET estimate is derived by rescaling the spread of participating ET products so that their spread reflects the ability of the weighted mean estimate to match flux tower data. While issues in observational data and any common biases in participating ET datasets are limitations to the success of this approach, future datasets can easily be incorporated and enhance the derived product.
observations of ET and most of the drivers and outputs of LSMs with soundly quantifiable uncertainty and sufficient accuracy to make diagnostic evaluation possible. This fact has led to a range of model evaluation, comparison and benchmarking studies using FLUXNET data (e.g. Abramowitz, 2005, 2012; Best et al., 2015; Chen et al., 1997; Stöckli et al., 2008; Wang et al., 2007). Haughton et al. (2016) clearly showed that the flux tower data quality is good enough to provide diagnostic constraint on LSMs at the site scale. However, the point scale is not the spatial scale at which these models are typically used. In either climate projections or weather forecast applications, gridded model estimates are required at the regional or global scale, with grid cell surface areas of order 25 km²–40 000 km². The relevance of model diagnostic information at the site scale to broader scale gridded simulations remains unclear.

The advancement of remote sensing technology and satellite image processing techniques has dramatically improved the estimation of many components of LSM forcing (e.g. surface air temperature, precipitation, radiation components and vegetation properties) (Loew et al., 2016). These datasets together with in-situ surface observations have provided constraint on the reanalysis products that provide the basis of global gridded LSM forcing products. On the other hand, very few LSM outputs (such as evapotranspiration or sensible heat flux) can be directly observed by remote sensing, and the only currently available way to derive them is by the use of modelling schemes or empirical formulations that use satellite based datasets (Jung et al., 2011; Fisher et al., 2008; Miralles et al., 2011; Martens et al., 2016; Mu et al., 2007; 2011; Su, 2002; Zhang et al., 2016; 2010). This approach leads to large and usually unquantifiable uncertainties. As a result, gridded LSM evaluation products show considerable differences (Ershadi et al., 2014; Jiménez et al., 2011; McCabe et al., 2016; Michel et al., 2016; Vinukollu et al., 2011). If an accurate description of these uncertainties was available, then it might be possible to evaluate gridded LSM simulations in certain circumstances, but until now very few studies quantify LSM uncertainties at regional or global scales (Badgley et al., 2015; Loew et al., 2016; Zhang et al., 2016).

Several gridded ET products have already been developed (see Table 1). These ET products differ in their data requirements, the approaches used to derive them and their estimates (Wang and Dickinson, 2012). So far, it is not at all clear which product provides the most reliable estimates. Recently, inter-comparison studies have aimed to evaluate and compare the available gridded ET products. For instance, Vinukollu et al. (2011), ran simulations of the Surface Energy Balance System SEBS, Penman–Monteith scheme by Mu (PM-Mu) and modified Priestley–Taylor (PT–JPL) forced by a compiled gridded dataset. The monthly model ET estimates were then compared with ET observations from 12 eddy-covariance towers. The root mean square deviation (RMSD) and biases for all the models fell within a small range with the highest correlation with tower data for PT–JPL, and the lowest for SEBS. These results somewhat disagree with the work of Ershadi et al. (2014) who inspected the performance of SEBS, the single-source Penman–Monteith, advection-aridity (AA) complementary method (Brutsaert and Stricker, 1979) and PT–JPL using a high quality forcing dataset from 20 FLUXNET stations and found that the models can be ranked from the best to the worst model as: PT–JPL followed closely by SEBS then PM and finally AA. In the same study, a more detailed analysis revealed that no single model was consistently the best across all landscapes. More recently, in the Energy and Water cycle Experiment (GEWEX) LandFlux project, McCabe et al. (2016)
ran simulations of the algorithms in SEBS, PT–JPL, PM-Mu and Global Land Evaporation Amsterdam Model (GLEAM) from common global-scale gridded forcing data, as well as site-based forcing data, and assessed their response relative to ET measurements from forty-five globally distributed FLUXNET towers. The results indicated that PT–JPL achieved the highest statistical performance, followed closely by GLEAM, whereas PM-MU and SEBS tended to under- and over-estimate fluxes respectively. Furthermore, in the Water Cycle Multi-mission Observation Strategy for Evapotranspiration (WACMOS-ET) project-part 1, Michel et al. (2016) carried out validation experiments for PT–JPL, PM-MU, SEBS and GLEAM against in situ measurements from 85 FLUXNET towers and found that there was no single best performing model. In each of the evaluation studies described above, tower data from FLUXNET provide ground truth for gridded ET datasets by comparing grid cell values to those measured at the site scale. Most gridded ET products have a 0.5-degree resolution, so that each grid cell can represent an area of around 2500 km². The fetch of flux tower measurements varies depending on terrain, vegetation and weather, but is typically under 1 km² (Burba and Anderson, 2010). None of these studies directly address this obvious scale mismatch, and the degree to which surface heterogeneity might nullify any information that flux towers provide about fluxes at these larger scales. Indeed, all the evaluation studies that compared gridded estimates against flux tower observations highlighted this point as a limitation in the evaluation approach. Zhang et al. (2016) pointed out that merging an ensemble of gridded ET products using a sophisticated data fusion method is likely to generate a better ET product with reduced uncertainty. Ershadi et al. (2014) and McCabe et al. (2016) noted in their studies that the multi-product mean produces improved estimates relative to individual ET products. Similarly, the analysis result of the LandFlux-EVAL project (Mueller et al., 2013) suggested that deriving ET values using the mean of multiple datasets outperforms the ET values from individual datasets. These results do suggest that flux towers likely provide some information at the gridded scale, since there are solid theoretical reasons to expect the mean to outperform individual estimates (Annan and Hargreaves, 2010; Bishop and Abramowitz, 2013), although this fact was not noted in these studies. That is, we would expect that the mean of a range of relatively independent approaches – whose errors are somewhat uncorrelated – would provide a better estimate to observations if those observations were of the same system being simulated. In this study, we provide an even stronger vindication of this relationship in section 2.3 below.

This paper introduces a new method for deriving globally gridded ET as well as its spatio-temporal uncertainty by combining existing gridded ET estimates. The method is based on the ensemble weighting and rescaling technique suggested by Bishop and Abramowitz (2013). The technique provides an analytically optimal linear combination of ensemble members that minimizes mean square error when compared to an observational dataset, and so accounts for both the performance differences and error covariance between the participating ET products (for example, caused by the fact that different ET datasets might share a forcing product). In this way, at least in-sample, the optimal linear combination is insensitive to the addition of redundant information. We use a broad collection of FLUXNET sites as the observational constraint. We examine the performance of the weighting approach in several in-sample and out-of-sample tests that confirm flux towers do indeed provide information at the grid scale of these products. Having confirmed this, we note that this “Derived
Optimal Linear Combination Evapotranspiration” (DOLCE) is more observationally constrained than the individual participating gridded ET estimates and therefore provides a valuable addition to currently available gridded ET estimates. Sections 2.1 and 2.2 present the gridded ET datasets and the flux tower ET used to derive the weighted ET product. Sect. 2.3 describes the weighting approach used to calculate the linear combination ET and its uncertainty. In Sect. 2.4 we explain our methods for testing the weighting approach and the representativeness of the point scale to the grid scale. We then present and discuss our results in Sect. 3 and 4 before concluding.

2 Methods, data and experimental setup

To build the DOLCE product, six global gridded ET products, all classified as diagnostic datasets based on Mueller et al. (2011), are weighted based on their ability to match site-level data from 159 globally distributed flux tower sites. DOLCE is derived at 0.5° spatial resolution and monthly temporal resolution for 2000–2009.

2.1 ET datasets

This study employs monthly values from ten existing gridded ET datasets (Table 1). Only six of these datasets – referred to in this study as the Diagnostic Ensemble – are used for building DOLCE, while all of the ten are used to evaluate its performance (we’ll refer to these as the Reference Ensemble). The reasons for restricting the Diagnostic Ensemble are (a) to maximize the time period covered by DOLCE (see Table 1), (b) to avoid temporal discontinuities in the derived product that can result from using different component products in different time periods , (c) to maximize the number of flux tower sites that can inform the weights (noting that datasets have different spatial coverage), and (d) to avoid LSM-based estimates in the final DOLCE product, so that its validity for LSM evaluation is clearer. For these reasons, CSIRO-global, LandFlux-EVAL-ALL, LandFlux-EVAL-DIAG and PT–JPL were removed from the Diagnostic Ensemble.

The weighting approach becomes increasingly effective as more products are included in the weighted mean, but since different products have different spatial coverage, there is a need to strike a balance between maximizing the number of products included in DOLCE and maximizing spatial coverage of DOLCE. To resolve this issue, three different subsets of the Diagnostic Ensemble were chosen to derive DOLCE “tiers”. The three DOLCE tiers differ in their spatial coverage of ET and uncertainty, the number of ET products included, and the number of flux towers employed to compute them. DOLCE tier 1 was derived by employing all the products in the Diagnostic Ensemble and 138 sites, and has a spatial coverage equal to the intersection of the global land covered by these products, as shown in Red in Fig. 1. In DOLCE tier 2, GLEAM v2B was excluded and the remaining products were used to compute ET and uncertainty, increasing the spatial coverage and allowing 151 observational flux tower sites. The newly added regions in DOLCE tier 2 are shown in orange in Fig. 1. Finally, ET and uncertainties in the regions shown in yellow were computed using only two ET products, GLEAM V2A and GLEAM V3A, to create DOLCE tier 3, using 159 flux sites. With this approach, we manage to create a global product without sacrificing data quality in regions where more information is available.
2.2 Flux tower data

The flux tower data used in this work is a composite of daily values from the FLUXNET2015-Tier 1 (FN) and LaThuile2007 (LT) Free Fair Use databases (Baldocchi, 2008; Baldocchi et al., 2001; Papale et al., 2012; ORNL DAAC, 2015). We applied quality control and filtering to the site data as follows:

1. Omit the observations that don’t fall within the temporal coverage of DOLCE.
2. Omit the daily ET observations if less than 50% of half-hourly ET was observed on that day (as opposed to gap-filled).
3. Omit the monthly ET aggregates that have been calculated from less than 15 daily mean ET values.
4. Correct the remaining site-months of ET observations from the LT dataset for energy balance non-closure on a site-by-site basis (energy closure corrected FN daily data was used)
5. Keep only the sites that are located within the spatial coverage of products in any of the three tiers of the Diagnostic Ensemble (Fig. 2).
6. Exclude irrigated sites or those that have only one monthly record.

Applying (1), (2) and (3) reduced the number of tower sites that were available at the beginning of the analysis from 246 to 172 sites, and produced a total of 7891 site-months (out of a possible 18468). Applying (5) and (6) reduced the number of sites against which the weighting is tested to 138 sites in DOLCE tier 1, increasing to 151 sites in DOLCE tier 2 and 159 in DOLCE tier 3. In (4), two different correction techniques were applied for energy balance non-closure at LT sites. Both involve ensuring that \( R_n - G = H + LE \) where \( R_n \) is the net radiation, \( G \), \( H \) and \( LE \) are the ground heat flux, sensible heat flux and latent heat flux respectively, at either the monthly timescale or over the length of the entire site record. Applying a correction technique for energy imbalance at LT sites required applying (2) and (3) for the other components of energy imbalance (i.e. \( R_n \), \( G \) and \( H \)), which means that the sites that had to undergo a correction for the energy imbalance, should have monthly estimates for all the fluxes of the energy budgets, where each monthly value has been calculated from at least 15 daily mean flux values. Because of this constraint, many sites were disregarded from the analysis. One of the two correction method for energy imbalance at LT sites distributes the residual errors in heat fluxes according to the Bowen ratio (BR) (Bowen, 1926; Sumner and Jacobs, 2005; Twine et al., 2000) such that:

\[
BR = \frac{H_{uncorrected}}{LE_{uncorrected}} \tag{1}
\]

is applied using

\[
F_{corrected} = \frac{R_n - G}{H_{uncorrected} + LE_{uncorrected}} \times F_{uncorrected} \tag{2}
\]

where \( F \) is either \( H \) or \( LE \), so that BR is unchanged between corrected and uncorrected fluxes. At first, the Bowen Ratio technique was applied on monthly basis, but this occasionally seemed to give erratic results in the corrected flux values. However, when a single correction per site was applied across all the monthly records, the results were more consistent.
The second approach calculates ET as the Residual term in the Energy Balance Equation \( \text{LE}_{\text{corrected}} = R_n - H - G \). The two correction methods were tested separately but no qualitative differences were noticed. All results below use the BR technique. In (6), we expect that some of the weighting models will largely underestimate the flux at irrigated sites, a result of a missing irrigation module in their scheme (Jung et al., 2011; Miralles et al., 2011). Because of this, the error bias of these models at the irrigated sites will modify the mean error bias (i.e. mean bias across all the sites) significantly, which will affect the weighting in favour of the products that can represent better irrigation. We excluded these sites as we do not want the products to be weighted for their inclusion/non-inclusion of physical processes.

We used daily averages of latent heat flux represented by “LE_CORR” in FN dataset. In LT dataset, we employed the components of energy imbalance and their associated flags (in brackets), represented by \( G_f \) (\( G_{fqcOK} \)), for soil heat flux, \( H_f \) (\( H_{fqcOK} \)) for sensible heat flux, \( Rn_f \) (\( Rn_{fqcOK} \)) for surface net radiation and \( LE_f \) (\( LE_{fqcOK} \)) for latent heat flux.

### 2.3 Weighting approach

The weighting approach presented here was suggested by (Bishop and Abramowitz, 2013). Given an ensemble of ET estimates and a corresponding observational dataset across time and space, the weighting builds a linear combination of the ensemble members that minimizes the mean square difference (MSD) with respect to the observational dataset such that: If \( x_{kj} \) is the \( j \)th time-space step of the \( k \)th bias-corrected ET product (i.e. after subtracting the mean error from the product), and a linear combination of the ET estimates is expressed as

\[
\mu^j_e = w^T x^j = \sum_{k=1}^{K} w_k x_{kj} , \tag{3}
\]

then the weights \( w^T \) provide an analytical solution to the minimization of

\[
\sum_{j=1}^{J} (\mu^j_e - y^j)^2 \tag{4}
\]

subject to the constraint that \( \sum_{k=1}^{K} w_k = 1 \), where \( j \in [1, J] \) are the time-space steps, \( k \in [1, K] \) represent the ET products, \( y^j \) is the \( j \)th observed time-space step.

The solution is expressed as

\[
w = \frac{A^{-1} 1}{1^T A^{-1} 1} , \tag{5}
\]

where \( 1^T = [1,1,\ldots,1] \) and \( A \) is the \( k \times k \) error covariance matrix of the gridded products. Further details are in (Abramowitz et al., 2015; Bishop and Abramowitz, 2013; Zeller and Hehn, 1996).

As noted above, this weighting approach has two key advantages. 1) It provides an optimal solution to minimizing mean square error differences between the weighted ET estimates and the observational tower data in-sample. 2) It accounts for the error covariance between the participating datasets (e.g. caused by the fact that single datasets may share ET schemes or forcing), that is, they may not provide independent estimates. The analytical solution guarantees that the weighted mean will perform as well or better than the unweighted mean or any individual estimate included in the linear combination, at least on the data used to train the weights (that is, in-sample). Moreover, the addition of a new gridded ET product to the ensemble...
will not degrade the performance of the weighted product, even if it consists of a duplicate product or a poor performing product. In order to confirm this in-sample weighting improvement, we performed an in-sample test presented in the supplementary material (Fig. S1). However, there is no guarantee that the weighted product will necessarily perform well out-of-sample, at a collection of sites not used for training. We explore this more below.

We use the ensemble dependence transformation process presented in Bishop and Abramowitz (2013) to calculate the spatiotemporal uncertainty of DOLCE. This involves first quantifying the discrepancy between our weighted ET estimate, \( \mu_e \), and the flux tower data, expressed as an error variance, \( s_e^2 \), over time and space. We then transform our Diagnostic Ensemble so that its variance about \( \mu_e \) at a given time-space step, \( \sigma_{e,j}^2 \), averaged over all time and space steps where we have flux tower data, is equal to \( s_e^2 \). This process ensures that the spread of the transformed Diagnostic Ensemble provides a better uncertainty estimate than simply using the spread of all products in the original Diagnostic Ensemble, since the spread of the transformed ensemble now accurately reflects uncertainty in those grid cells where flux tower data is available.

We calculate the discrepancy between observations and our weighted mean as

\[
s_e^2 = \frac{\sum_{j=1}^{J} (\mu_e^j - y^j)^2}{J-1} \tag{6}
\]

Next, we wish to ensure that

\[
\frac{1}{J} \sum_{j=1}^{J} \sigma_{e,j}^2 = s_e^2 \tag{7}
\]

but the variance of our existing Diagnostic Ensemble will not, in general, satisfy this equation. To transform it so that it does, we first modify the coefficients from Eq. (3), so that they’re guaranteed to all be positive:

\[
\tilde{w} = \frac{w^T + (\alpha-1) 1^T}{\alpha} \tag{8}
\]

where \( \alpha = 1 - K \min(w_k) \) and \( \min(w_k) \) is the smallest negative weight (and \( \alpha \) is set 1 if all \( w_k \) are non-negative). We then transform the ensemble using

\[
\tilde{x}_k^j = \mu_e^j + \beta(x_k^j - \bar{x}^j) - \mu_e^j \tag{9}
\]

where

\[
\beta = \sqrt{\frac{s_e^2}{\frac{1}{J} \sum_{j=1}^{J} \sum_{k=1}^{K} \tilde{w}_k (x_k^j - \bar{x}^j) - \mu_e^j)^2}} \tag{10}
\]

If we then define the weighted variance estimate

\[
\sigma_{e,j}^2 = \sum_{k=1}^{K} \tilde{w}_k (\tilde{x}_k^j - \mu_e^j)^2 \tag{11}
\]

we ensure that both Eq. (6) above holds, and also that \( \mu_e^j = \sum_{k=1}^{K} w_k x_k^j = \sum_{k=1}^{K} \tilde{w}_k \tilde{x}_k^j \) (see Bishop and Abramowitz, 2013, for proofs). We then use \( \sqrt{\sigma_{e,j}^2} \) as the spatially and temporally varying estimate of uncertainty standard deviation, which we will refer to below as uncertainty.
2.4 Experimental setup

We employed four statistical metrics that reflect how well a gridded ET product represents the quality controlled flux tower observations: Mean Square Error \( \text{MSE} = \text{mean} (\text{dataset} - \text{observation})^2 \); Mean Bias = \( \text{mean} |\text{dataset} - \text{observation}| \); Correlation \( \text{COR} = \text{corr} (\text{observation}, \text{dataset}) \) and Modified Relative Standard Deviation \( \text{MRSD} = \frac{\sigma_{\text{dataset or observation}}}{\max (\text{mean} (\text{observation}), q)} \).

We use a modified relative standard deviation metric MRSD that measures the variability of latent heat flux relative to the mean of the flux measured at each site. This ensures that a comparison between MRSD for a product and observations can tell us whether a product’s variability is too large or too small (unlike relative standard deviation). The term ‘q’ is a threshold representing the 2\(^{nd}\) percentile of the distribution of observed mean flux (i.e. temporal mean ET) across all sites (\( \approx 13 \text{ W/m}^2 \)), which guarantees that MRSD calculated across many sites is not dominated by sites where the mean flux (denominator in MRSD Equation above) approaches zero. We looked at the bias in MRSD for each product considered- i.e. \( |\text{MRSD}_{\text{dataset}} - \text{MRSD}_{\text{observation}}| \), and showed the performance improvement of the weighted mean.

In every test, we split the available sites into in-sample and out-of-sample sites (that is, calibration and validation sites). First, we applied the weighting process at the in-sample sites to weight the individual datasets of the Diagnostic Ensemble and derive a weighted product. Then, we calculated the four metrics above using the out-of-sample sites and displayed results by showing:

a. The percentage performance improvement of the derived weighted product compared to the equally weighted mean (\( \text{Dmean} \)) of the Diagnostic Ensemble;

b. The percentage performance improvement of the derived weighted product compared to each individual product in the Reference Ensemble, yielding 10 different values of performance improvement.

c. The aggregate (\( \text{Ragg} \)) of the values of performance improvement calculated across all 10 products in the Reference Ensemble.

We display the results of performance improvement datasets calculated in (a-c) above in 12 box and whisker plots. In each boxplot, the lower and upper hinges represent the first (\( Q_1 \)) and third (\( Q_3 \)) quartiles respectively of the performance improvement datasets and the line located inside the boxplot represents the median value. The extreme of the lower whisker represents the minimum of 1) \( \max (\text{dataset}) \) and 2) \( (Q_3 + \text{IQR}) \), while the extreme of the upper whisker is the maximum of 1) \( \min (\text{dataset}) \) and 2) \( (Q_3 + \text{IQR}) \)), where IQR is the interquartile range of the performance improvement dataset. If the median performance improvement is positive, this indicates that the weighting offers an improvement in more than half of the data presented in the boxplot.

We first divide sites between the in-sample and out-of-sample groups by randomly selecting 25 % of the sites as out-of-sample. The remaining sites form the in-sample training set, and are used to calculate a scalar bias correction term and a weight for each participating gridded ET product. These bias correction terms and the weights are then applied to the products at the out-of-sample sites. The test was repeated 5000 times with different random selection of sites being out of sample.
Next, we repeat the process with just one site in the out-of-sample testing group. The bias corrections and weights are therefore derived on all sites except one, and tested on the single out-of-sample site. The same test was repeated for all the participating sites.

3. Results

The results for the 25 % out-of-sample test are displayed in the box and whisker plots presented in Fig. 3 (a), (b), (c) and (d). The MSE plot in Fig. 3 (a) indicates that across most of the random selections of the 25 % out-of-sample sites (57 % of random selections), the MSE of the weighted product is slightly better (with maximum 12 % improvement) than the equally weighted mean of the Diagnostic Ensemble. The weighting also improved results for almost all random selections of sites when compared to each individual dataset in the Reference Ensemble. The MRSD plot in Fig. 3 (b) shows that the weighting succeeded in improving MRSD relative to Dmean and PT-JPL by no more than 4 % and 36 % respectively across the majority of the combinations of sites, but didn’t improve MRSD relative to the other reference products. This is perhaps unsurprising, given that both the weighted and unweighted mean should have decreased variability when the variations in individual products are not temporally coincident. The COR plot in Fig. 3 (c) shows that despite the expected drop in variability in the weighted product, the correlation with site data has in fact improved across most of the different random selections of the 25 % out-of-sample sites relative to all of the Reference Ensemble datasets. The improvement was minimal relative to Dmean (maximum 2 %) and MPI (maximum 3 %). Similarly, the Mean Bias plot (Fig. 3 (d)) shows a clear improvement relative to Dmean and the Reference Ensemble datasets across most combinations of sites. It is important to reinforce that these results are for sites that were not used to train the weights. As detailed in section 2.3, performance improvement at training sites is expected, but the fact that the weighting delivers improvements at sites that were not included in training data indicates that there is indeed information content about the larger scales in site data.

The results of the one site out-of-sample tests are displayed in Fig. 4 (a), (b), (c) and (d). These boxplots indicate that across three metrics (MSE, MRSD and Mean Bias) the weighted mean clearly outperformed all the products in the Reference Ensemble. However, when the performance improvement of the weighted product is compared against Dmean the improvement was marginal. In MSE, an improvement was achieved at more than half (53 %) of the sites (Fig. 4 (a)). In MRSD, 51 % of the sites showed a small improvement relative to Dmean (Fig. 4 (b)). In COR, the performance decreased at 57 % of the sites, and a minimal improvement over the Reference Ensemble was shown at more than half of sites (Fig. 4 (c)). Finally, in Mean Bias, 58 % of the sites showed performance improvement over Dmean (Fig. 4 (d)). Part of the success of the weighting approach relative to the multi-product mean is due to the bias correction applied before the weighting. Figure S3 in supplementary material separates the effect of each step, and Fig. S4 shows examples of time series of merged products before and after bias correction with the resultant weighted product at five tower locations. We also note that the bias correction terms were calculated from the temporal mean of the monthly bias at all the flux tower sites across the time period 2000-2009. As a result, the individual bias corrected products may show unphysical ET estimates especially in...
locations of strong seasonal cycle and at high latitudes especially when ET values are close to zero. However, after applying
the weights to the component products, the produced values in the merged product were all realistic.

The results of the one site out-of-sample test suggest that spatial heterogeneity - the discrepancy between site and grid scale estimates - is significant, as the spread of these boxplots is large. These sites are more likely to be where the fetch of the flux site observations cannot represent the data of the entire grid (estimates). A further analysis (not shown here) has indicated that sites that show poor improvement in Fig. 4 (a), (b), (c) and (d) have a consistently high bias against all of the gridded products.

To investigate the effect this spatial heterogeneity is having on the out-of-sample tests, we also repeat the one site out-of-sample experiments using a subset of the sites within grid cells that are deemed to be relatively homogeneous (Fig. 4 (e), (f), (g) and (h)). Homogeneity is defined in this case by using only those sites that have the same IGBP vegetation type as the grid cell that contains them. Almost all the gridded ET products use the dominant land cover type for computing ET, yet this is not always the same as the land cover type surrounding the flux site. Since this mismatch occurs in heterogeneous terrains, we will denote the sites that have this property as “HET-case” sites, whereas the sites that show agreement with the underlying grid cell vegetation type are denoted as “HOM-case” sites. We retrieve the IGBP vegetation cover data of the grid cells from MODIS Land Cover for year 2009 at 0.5° spatial resolution (http://glof.umd.edu/data/lc/; Friedl et al., 2010).

The vegetation type of individual sites is taken from metadata on the FLUXNET website. We show the distribution of HET-case and HOM-case sites by biome type in Table S2 in the supplements. There is some expectation that the weighting will show better performance if it is trained with HOM-case sites only, although HOM-case sites consist of about one-thirds of the total number of sites used in this study. To further investigate this idea, we carried out the one site out-of-sample test training and testing on HOM-case sites data only. Figures 4 (e), (f), (g) and (h) show that the HOM-case subset of tower sites does indeed improve the result marginally at least in terms of MSE, COR and Mean Bias when compared to the original dataset (HOM & HET-case; Fig. 4 (a), (b), (c) and (d)). We did not investigate the 25% out-of-sample tests with HOM-case sites, since the number of in-sample sites would become too small, and likely lead to overfitting. In both cases it is important to note that many individual sites agree poorly with the weighted product compared to some other products. The distinction between the results shown in Fig. 3 versus Fig. 4 serves to highlight that DOLCE, and indeed any other large scale gridded ET product, is not suitable for estimation of an individual site’s fluxes, even if prediction over many sites shows notable improvement.

On the basis of the aggregate out-of-sample improvement that this approach offers over existing gridded ET products, in terms of MSE, MRSD, COR and Mean Bias against site data, we now present details of the DOLCE product, which is trained using all site data and derived from the combination of the six diagnostic products. The 3 GLEAM products were resampled to 0.5° using bilinear interpolation to match the spatial resolution of DOLCE.

We calculated four statistics - Mean Bias, RMSE, Standard Deviation (SD) difference (i.e. $\sigma_{DOLCE} - \sigma_{observation}$) and Correlation - to see how well DOLCE performs at each site. The results are displayed in Fig. 5 (a), (b), (c) and (d). At about half of sites, DOLCE had values between −6 and 6 W m$^{-2}$, −5 and 5 W m$^{-2}$ in the Mean bias and SD difference metrics.
respectively, and values greater than 0.93 W m\(^{-2}\) and less than 14 W m\(^{-2}\) in correlation and RMSE respectively. DOLCE metrics vary in the range \([-43.4–30.1\) W m\(^{-2}\)] for mean bias, \([-25.8–23.5\) W m\(^{-2}\)] for SD difference, \([0.09–0.99\) W m\(^{-2}\)] for correlation and \([4.2–60.9\) W m\(^{-2}\)] for RMSE.

We also calculated the on-site metrics (i.e., RMSE, Mean Bias, SD difference, and Correlation) for DOLCE separately at each subset of sites (i.e. HOM-case and HET-case) and we displayed the results in two boxplots (Fig. 5 (left column)). These boxplots show that most of the sites at which DOLCE showed low performance are HET sites, since the end of whiskers is larger for the HET-case sites.

We tested the performance of DOLCE at three irrigated sites that were excluded from the weighting for reasons explained earlier (section 2.2), by computing the four statistics. A description of these sites and the results are shown in Table 2. The results show that the performance of DOLCE is reasonable at US-Ne1 and US-Ne2 and low at US-Twt. These results are discussed further below.

We now look at the differences between DOLCE and two widely used ET products. Figures 6 and 7 show the seasonal mean difference ET between MPIBGC and DOLCE and between LandFlux-EVAL-DIAG and DOLCE respectively. The differences were computed from seasonal means using the same spatial mask over the period 2000–2009 for MPIBGC and 2000–2005 for LandFlux-EVAL-DIAG. The seasonal plots in Fig. 6 show that, overall, DOLCE has lower ET than MPIBGC. DOLCE tends to have higher ET values in the Asian and Australian tropics during the austral summer and autumn. DOLCE has higher ET values in Sahel during Sep–Nov and in the high plateau of Madagascar during Dec–May. In the Amazon, DOLCE exhibits higher ET values between June and November. Higher values of DOLCE are seen in Guiana highlands throughout the year. The Brazilian highlands show higher values for MPIBGC from June to November. The Mid latitudes of the Northern hemisphere show higher values for MPIBGC between March and August but no significant differences are seen in this area between September and February.

In Fig. 7, there are large differences in ET showing higher values in LandFlux-EVAL-DIAG ET over the Amazon, the rainforests of Southeast Asia and in the high plateau of Madagascar throughout the year. DOLCE tends to exhibit higher ET values in the Ethiopian highlands and Myanmar between June and November.

The spatial distribution of DOLCE mean ET and its seasonal variability (standard deviation) over the austral Summer (Dec–Feb) and Winter (Jun–Aug) from 2000 to 2009 is shown in Fig. 8 (a) and (b) respectively. The seasonal variability of ET is larger in the warm season but is always small over Antarctica, Greenland and the deserts in North Africa (Sahara), the middle east (Arabian Peninsula desert) and Asia (i.e. Gobi, Takla Makan and Thar). The average uncertainty shown in if Fig. 8 (c) is bigger in the warm season, this is in agreement with the relatively large size of the flux in the warm season, and its seasonal variability shown in Fig. 8 (d) is also in agreement with the seasonal variability of the flux. Figure 8 (e) is intended to give some indication of the reliability of DOLCE. Regions in blue show grid cells that satisfy \(\frac{\text{Uncertainty SD}}{\text{mean ET}} \leq 1\). Those shown in green have \(|\text{mean ET}| < 5\) and Uncertainty SD < 10 but don’t satisfy \(\frac{\text{Uncertainty SD}}{\text{mean ET}} \leq 1\), so that the uncertainty estimates are higher than their associated ET, but both the uncertainty and the ET estimates are very small. All the remaining
grid cells, perhaps those least reliable, are shown in red. The global maps in Fig. 8 (e) shows that most of the non-reliable uncertainty values are located in the land added in tier 3 (Fig. 2 in yellow). This is not surprising, because uncertainty values have been derived using only two gridded ET datasets that are also not observationally constrained due to the lack of the sites representing the terrestrial ecosystems and environmental conditions in these areas. Some regions show unreliable uncertainty estimates during the cold season only, located in the plateau of Tibet, over the Andes, the Australian deserts, the South-African deserts, American Great basin and Rocky Mountains.

4. Discussion

Many studies have analysed the systematic and random errors of latent heat flux in FLUXNET measurements (Dirmeyer et al., 2016; Göckede et al., 2008; Richardson et al., 2006). These studies have detected errors of magnitudes that cannot be neglected. A recent study (Cheng et al., 2017) showed that the computed eddy-covariance fluxes have errors in the applied turbulence theory that lead to the underestimation of fluxes, and that this is likely to be one of the causes of the lack of surface energy closure. In this study, we 1) used the flag assigned to the observed flux, to filter out the low-quality data and 2) used energy-balance-corrected FLUXNET data which has higher per-site mean values than the raw data at most of the sites (85% of them). We expect that filtering together with the use of corrected data will reduce the magnitude of the uncertainty in the observational data used here and compensate to a certain extent for the underestimation due to the systematic errors. However, we have not formally explored a range of approaches to addressing this. The possibility of systematic biases in FLUXNET data remains, and this could clearly lead to systematic biases in DOLCE.

We have also assumed that error across sites is uncorrelated, which, given the distribution of sites, is unlikely to be true, meaning that the effective number of sites is probably somewhat smaller than those shown in Fig. 2. Given this dependence is likely to vary depending on a range of time varying factors, we have left the job of attempting to disentangle this issue for future work.

Nevertheless, the results of the one site out-of-sample and 25 % out-of-sample tests above suggest that the weighting of gridded ET products in this way produces a more capable ET product overall. Critically, the fact that the weighting improves out of sample performance suggests that while the representativeness of point-scale measurement for the grid scale may not exist at every single site, it does exist across all these sites as a whole. We also investigated using bilinear interpolation instead of direct grid cell to tower comparison (not shown), but found no qualitative differences.

It might be reasonable to assume that eliminating sites that poorly represent grid cell properties might further improve these results, yet the results of categorizing sites into HOM-case and HET-case subsets suggested that this is not necessarily a simple process. Our characterisation was based on a relatively simple vegetation characterisation, yet there are a range of soil-vegetation interactions that likely affect ET in each location, that are not captured in this classification. Among HET-case sites, for example, the worst performance occurred at AU-Fog. This poor performance is a reflection of the large discrepancy between the land cover seen by the flux tower which is a wetland landscape in a monsoonal climate, and thus
characterised by high evaporation rates throughout the year and the Savannah land cover seen by the satellite products where the vegetation dries or dies in the dry season. Also, we note that none of the six diagnostic products was able to reproduce the flux measured at this site. Perhaps unsurprisingly, DOLCE scored the worst values for correlation, RMSE and Mean Bias in AU-Fog.

Aside from AU-Fog, three sites showed low correlation (less than 0.5) between DOLCE and observations. These include ID-Pag, ES-LgS and IL-Yat (see Supplementary Table S1). IL-Yat site is an Evergreen Needleleaf Forest site located over a desert (Bare) grid cell and therefore belongs to the HET-case which might be the reason for the discrepancy between the satellite derived ET and the observed ET. ID-Pag belongs to the HOM-case, but a burning event occurred in a nearby forest between mid-August and late October 2002, and caused the site to be covered with a dense smoke-haze. Although these three months were excluded from the analysis, it is very likely that the burned region produced extremely different ET compared to the surrounding unburned landscape, making it a HET-case site until the regeneration of the vegetation and might be the reason why the weighting products couldn’t reproduce ET in the following couple of months. ES-LgS is an Open Shrubland site that belongs to the HOM-case. This site operated from 2007 to 2009 and is currently inactive. The reason of the low correlation of DOLCE with the observed fluxes is not clear, however previous studies (Ershadi et al., 2014; McCabe et al., 2016) stated that the complexity of Shrubland landscape is the cause of the poor performance of many models (PM, GLEAM and PT–JPL). In the case of DOLCE, the poor performance was revealed in one metric only (i.e. correlation). These results nevertheless lead us to expect that if we construct DOLCE by incorporating HOM-case sites only, we might get a better product, but the small number of sites satisfying this property, the fact that the separation of sites into HOM-case and HET-case can lead to a separation of land covers, and the difficulty in defining a meaningful definition for expected flux homogeneity are limiting factors. Determining whether DOLCE performs better at HOM-case sites or HET-case sites is inconclusive. Even that the worst performance of DOLCE was achieved in HET-case sites, the boxplots in Fig. 5 (a), (d) show that the value of the median, lower and upper quartiles are better in HET-case for two metrics (i.e. RMSE and COR). While we expect that calibrating the weighting with HOM-case could lead to a better product, we don’t expect to see DOLCE overall performing better in any of the two groups.

Figure 9 shows that DOLCE performs differently for different vegetation types, at least as sampled by the flux towers we use here. Previous studies have shown that different ET estimation approaches perform differently over different biomes. For example, Ershadi et al. (2014) found that the PM models outperform PT–JPL across Grasslands, Croplands and Deciduous Broadleaf Forests, but PT–JPL has higher performance than PM models over Shrublands and Evergreen Needleleaf Forest biomes. McCabe et al. (2016) observed poor performance of GLEAM, PM models and PT–JPL over Shrublands and low performance over the forest biomes and higher performance over short canopies. The number of sites for each vegetation type in our case varies, with four for Open Shrublands (OSH), 39 for Evergreen Needleleaf Forest (ENF), nine for Evergreen Broadleaf Forest (EBF), 18 for Deciduous Broadleaf Forest (DBF), 10 for Mixed Forests (MF), four for Woody Savannas (WSA), four for Savannas (SAV), 42 for grassland (GRA), eight for Wetland (WET), one for Cropland/Natural Vegetation Mosaic (VEG) and 19 for Croplands (CRO). Of the 159 sites that are used to weight the tier 3 products (GLEAM v2A and
GLEAM v3A), only 142 sites are within the spatial coverage of DOLCE. This decrease is caused by the coarser resolution of DOLCE compared to GLEAM products. The remaining 17 sites were located on grids that were identified as land grid cells in GLEAM’s spatial resolution (0.25°) but water body in DOLCE’s spatial resolution (0.5°). The EBF box and whisker plot in Fig. 9 (d) shows the correlation of DOLCE at eight EBF sites, out of which two sites are located in the tropics. The lowest correlation seen in this biome type is at the tropical sites ID-Pag (0.26) and BR-Sa3 (0.62). This suggests that DOLCE tends to represent ET at the extratropical sites better than the tropics, and this is not surprising since most of the sites that were used to calibrate DOLCE were extratropical sites.

DOLCE has also shown a weak performance at US-Twt, which is an irrigated rice paddy. This site gets flooded in spring and drains in early fall, then the rice is harvested. Only nine months were available for this site, which coincide with the flood and drain period between spring and fall. DOLCE could not depict the flooding and draining event, probably because none of the weighting products can represent such phenomena, so it is expected that the effects of seasonal flooding are not represented in DOLCE.

In this study, we sought a single weight for each product to apply globally. But we have reason to believe that different products are likely to perform better in different environments, so that different weights in different climatic circumstances might well improve the result of weighting overall. A similar suggestion was made in the studies of Ershadi et al. (2014) and Michel et al. (2016) who highlighted the need to develop a composite model, where individual models are assigned weights based on their performance across particular biome types and climate zones. We therefore tried to cluster flux tower sites into groups (such as vegetation type) so that each group maintains enough members to allow the in- and out-of-sample testing approach used above. We tried clustering by vegetation type, climate zone and aridity index, and implemented the same one site out-of-sample testing approach as above, but this time, in each cluster different sets of weights will be assigned to the weighting products. The motivation for this approach was to try to reduce the number of sites that did not show improvement as a result of the weighting, and ideally improve performance of DOLCE overall, yet none of these clustering approaches delivered any improvement, possibly due to the considerably reduced sample size in each cluster (increasing the likelihood of overfitting). It might also be a result of grid cell heterogeneity within the small sample that constitutes each vegetation type. Finally, as noted above, we have no guarantee that vegetation type is necessarily a good proxy for surface flux behaviour. A summary of this investigation and a plot showing the results of the clustered weighting by vegetation type (Fig. S2) are included in the supplementary material.

There are relatively few towers located in the Southern Hemisphere and the tropics (14 out of 159 sites) and none located in the dry climates over South West Asia and North Africa. The weighting was therefore mostly driven by the ability of products to match sites located in the temperate and cold zones of the Northern Hemisphere, so that performance in climate zones with low FLUXNET site density was under-represented when deriving DOLCE. This might raise questions about the performance of DOLCE in the tropics and the Southern Hemisphere. To evaluate DOLCE in these areas we calculated the four site metrics (RMSE, Mean bias, SD difference and Correlation) separately for two groups of sites, 1) those located in the Northern Hemisphere excluding Tropics and 2) sites located in the Tropics and/or Southern Hemisphere. We excluded
the two sites ID-Pag and AU-Fog in this exercise since both are wetland sites, and so would complicate a determination of whether these two groups had notable behavioural differences. If systematic behavioural differences did exist between these two groups, we would expect relatively poorer performance of DOLCE at group2 sites compared to group1 sites. The results, shown in Fig. 10, appear inconclusive. DOLCE performed marginally worse at group2 sites overall, however with the limited number of sites in group2, the validation of the performance of DOLCE in the tropics and Southern Hemisphere remains somewhat uncertain. The uneven distribution of eddy-covariance sites between the Northern and Southern Hemisphere and across the different climates might also explain why much of the largest seasonal differences DOLCE-MPI and DOLCE-LANDFLUX-EVAL shown in Fig. 6 and Fig. 7 reside in the low latitudes (tropics) and the Southern Hemisphere and the persistent differences between DOLCE and LANDFLUX-EVAL in the tropics throughout the year. The expansion of the FLUXNET network into these areas that are lacking observations is clearly something that would improve DOLCE and LSM evaluation more broadly.

In this study, DOLCE was derived by combining remote sensing products that use an empirical approach (i.e. MPIBGC) and a more physical approach (i.e. MOD16, GLEAM v2A, GLEAM v2B and GLEAM v3A and PML). Constructing an ET product by combining different approaches can take advantage of their desirable features and reduce their limitations (Zhang et al., 2016). This is indeed reflected by the enhanced performance of DOLCE over all the biome types against the reference products. Table 3 shows that the weights were attributed almost equally to each approach (i.e. ≈ 0.5 for both MPIBGC and the physical ensemble in tier 1 and tier 2), which indicates that the two approaches contributed equally in deriving DOLCE, it is not surprising that MPIBGC was the most weighted product since it is highly calibrated with flux tower data. In a further analysis, we left MPIBGC out and we performed the out of sample tests using the five remaining products. We wanted here to test how the weighting will perform without MPIBGC. The plots in Fig. 11 and 12 show the results of 25 % out-of-sample test and one site out-of-sample test respectively. Overall, without including MPIBGC, the weighting offers a smaller performance improvement than that offered when MPIBGC is a member of weighting ensemble (Fig. 3 and Fig. 4 (a-d)). The distribution of the weights when MPIBGC is absent from the weighting is 0.3 for both PML and GLEAM_v3A, 0.2 for GLEAM_v2B, 0.13 for MOD16 and 0.07 for GLEAM_v2A.

It is important to mention that the weighting approach offers the possibility of enhancing DOLCE in the future by incorporating any ET dataset that becomes available, for example, WECANN, (Alemohammad et al., 2016), and HOLAPS v1.0, (Loew et al., 2016).

The limitations of the weighting approach presented here arise from issues in the observational data and limitations in the weighted datasets that are employed. FLUXNET data inevitably has some instrumentation-related problems (Göckede et al., 2008) that affect the data quality of the measured flux in some terrains and under some environmental conditions. Moreover, FLUXNET sites are not globally-representative of all terrestrial ecosystems, and not evenly distributed across the biome types, which might lead to biases when computing the weights of the products in favour of the product that outperforms over the most frequent cover types. Also, in some areas where tower density is relatively high, the information from different
towers is not necessarily independent. Finally, a common imperfection in all the weighted products due to for example anthropogenic water management, will of course lead to the same imperfection in the derived product. The uncertainty estimate presented here is firmly grounded in the spread of existing gridded ET products, but is better than this spread alone, since this spread has been recalibrated so that the uncertainty of DOLCE where flux tower data exists is precisely the spread of the recalibrated products.

5. Conclusion

In this study, we presented a new global ET product with monthly temporal resolution for 2000–2009 at 0.5° spatial resolution and a calibrated estimate of its uncertainty. The approach used to weight existing gridded ET products accounted for both the performance differences and error covariance between the participating ET products. The DOLCE product performs better than any of its six constituent members overall, as well outperforming other well-known gridded ET products (CSIRO-global, LandFlux-EVAL-ALL, LandFlux-EVAL-DIAG and PT–JPL) across a range of metrics. The weighting applied here adds only a slight improvement to the simple mean of the merged products. Figure S5 displays the seasonal mean difference ET between (a) MPIBGC and DOLCE tier 1 and (b) MPIGBC and the simple mean of the six component products of DOLCE tier 1. The same spatial mask was applied in the plots and the differences were computed as in Fig. 6. The plots show that the differences exhibit similar patterns with stronger magnitude in the simple mean case. This might raise the question of whether it is worth applying a weighted approach to merge ET. There are however at least three compelling reasons to use this approach: (a) it removes the possibility of similar component products with redundant information biasing the mean (b) it allows us to confirm that point-scale flux tower measurements do indeed contain information about fluxes at the grid box scale, and (c) it allows us to create an uncertainty estimate that better utilises the information in the component products than simply using their range. In this work, we used 6 latent heat products that have different spatial coverage and we managed to create a global product by deriving DOLCE tiers that differ in the component products used to derive them. This might have resulted in spatial discontinuities that we haven’t explored in great detail. We might also expand the time coverage of DOLCE by incorporating different component products in different time periods, however, doing so could lead to large temporal discontinuities in the derived product. Merging a variety of different ET products is not straightforward and requires subjective decision-making, but we aim to explore extending temporal coverage using different combinations of component products in future iterations of DOLCE. It was shown that despite the scale mismatch between the flux tower and the grid cell, the ensemble of flux towers as a whole does provide information about the grid cells that contain them, since the improvements delivered by the weighting approach were evident in sites not used to derive the weights. While the representativeness of the point scale for the grid scale is enhanced by only considering sites that lie within homogeneous grid cells we suggest that an optimal definition of homogeneity for flux behaviour be the subject of future investigation. Nevertheless, DOLCE appears to outperform existing gridded ET products overall, and offers the opportunity for improvement as more flux tower data and new gridded ET products become available. Expanding
DOLCE over longer time periods and incorporating more diagnostic ET datasets (such as PT-JPL, CSIRO-global, GLEAM-V3B, SEBS, WECANN and HOLAPS) will be carried out in future versions.

6. Competing interests

The authors declare that they have no conflict of interest.

7. Data availability

DOLCE v1.0 can be downloaded from geonetwork.nci.org.au and its DOI is http://dx.doi.org/10.4225/41/58980b55b0495

8. Acknowledgements

Sanaa Hobeichi acknowledges the support of the Australian Research Council Centre of Excellence for Climate System Science (CE110001028). Gab Abramowitz and Jason Evans acknowledge the support of the Australian Research Council Centre of Excellence for Climate Extremes (CE170100023). This work used eddy covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices. This work also used eddy covariance data acquired by the FLUXNET community and in particular by the following networks: AmeriFlux (U.S. Department of Energy, Biological and Environmental Research, Terrestrial Carbon Program (DE - FG02 - 04ER63917 and DE - FG02 - 04ER63911)), AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet - Canada (supported by CFCAS, NSERC, BIOCAP, Environment Canada, and NRCan), GreenGrass, KoFlux, LBA, NECC, OzFlux, TCOS - Siberia, USCCC. We acknowledge the financial support to the eddy covariance data harmonization provided by CarboEuropeIP, FAO - GT OS - TCO, iLEAPS, Max Planck Institute for Biogeochemistry, National Science Foundation, University of Tuscia, Université Laval and Environment Canada and US Department of Energy and 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, University of Virginia.

9. References


Tables

Table 1: Gridded ET products used in this paper.

<table>
<thead>
<tr>
<th>ET product and Reference</th>
<th>Abbreviation</th>
<th>Time period &amp; Spatial Resolution</th>
<th>Forcing data source</th>
<th>Calculation Method(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSIRO-global (Zhang et al., 2010a)</td>
<td>CS</td>
<td>1983–2006 0.5° Also available at flux tower distributed across all global biome types</td>
<td>Meteorological observations from flux tower distributed across all global biome types</td>
<td>An extended ET product of CSIRO (Zhang et al., 2010b) that covers a global domain NDVI-based PM model</td>
</tr>
<tr>
<td>Model</td>
<td>Version</td>
<td>Time Period</td>
<td>Input Data</td>
<td>Equation and Model Components</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------</td>
<td>-------------------</td>
<td>------------------------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>GLEAM-V2A</td>
<td>G2A</td>
<td>1980–2011</td>
<td>Remote sensing based observations</td>
<td>PT equation</td>
</tr>
<tr>
<td>(Miralles et al., 2011)</td>
<td></td>
<td>0.25°</td>
<td>Gauged based precipitation</td>
<td>Canopy Interception Model, Soil water module and Stress module</td>
</tr>
<tr>
<td>GLEAM-V2B</td>
<td>G2B</td>
<td>2000–2011</td>
<td>Remote sensing based observations</td>
<td>PT equation</td>
</tr>
<tr>
<td>(Miralles et al., 2011)</td>
<td></td>
<td>0.25°</td>
<td></td>
<td>Canopy Interception Module, Soil water module and Stress module</td>
</tr>
<tr>
<td>GLEAM-V3A</td>
<td>G3A</td>
<td>1980–2014</td>
<td>Satellite based inputs</td>
<td>A revised version of GLEAM V2A in which new satellite-observed geophysical variables have been incorporated and the representation of the surface soil moisture and evaporation has been improved</td>
</tr>
<tr>
<td>(Martens et al., 2016)</td>
<td></td>
<td>0.25°</td>
<td>Multi-source precipitation</td>
<td></td>
</tr>
<tr>
<td>LandFlux-Eval-Diag</td>
<td>LFD</td>
<td>1989–2005</td>
<td>Simple mean of 5 diagnostic ET datasets</td>
<td></td>
</tr>
<tr>
<td>(Mueller et al., 2011, 2013)</td>
<td></td>
<td>1°</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LandFlux-Eval-All</td>
<td>LFA</td>
<td></td>
<td>Simple mean of 14 Diagnostic, LSM and Reanalysis datasets.</td>
<td></td>
</tr>
<tr>
<td>(Mueller et al., 2011, 2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>products</td>
<td></td>
<td>0.5° also available at</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Four metrics (RMSE, Mean bias, SD difference and Correlation) of DOLCE at three irrigated sites, and the number of available monthly records for each site.

<table>
<thead>
<tr>
<th>Site-Code</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Description</th>
<th>RMSE</th>
<th>Mean bias</th>
<th>SD difference</th>
<th>Correlation</th>
<th>Number of months</th>
</tr>
</thead>
<tbody>
<tr>
<td>US-Ne1</td>
<td>-96.4766</td>
<td>41.1651</td>
<td>Rice paddy</td>
<td>16.6</td>
<td>-7.41</td>
<td>-9.25</td>
<td>0.96</td>
<td>103</td>
</tr>
<tr>
<td>US-Ne2</td>
<td>-96.4701</td>
<td>41.1649</td>
<td>Mead irrigated continuous maize site</td>
<td>15.8</td>
<td>-5.05</td>
<td>-7.44</td>
<td>0.95</td>
<td>103</td>
</tr>
</tbody>
</table>

(Mu et al., 2011) reanalysis data
Remote sensing inputs from MODIS 8-day retrievals

MPIBGC (Jung et al., 2011) MPI 1982–2011
0.5° FLUXNET data from 253 sites
Remote sensing datasets from (SeaWiFS)
Empirical methods: a Model Tree Ensemble (MTE) Machine learning techniques

PML PM-Leuning model (Zhang et al., 2015) PML 1981–2012
0.5° GMAO Reanalysis products
PM Leuning method

PT–JPL (Fisher et al., 2008) PT 1984–2006
1° Meteorological reanalysis data from ISLSCP –II
Remote sensing based observations from monthly AVHRR data
PT equation
<table>
<thead>
<tr>
<th>Product</th>
<th>DOLCE tier 1</th>
<th>DOLCE tier 2</th>
<th>DOLCE tier 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>weight</td>
<td>Bias W m⁻²</td>
<td>weight</td>
</tr>
<tr>
<td>MOD16</td>
<td>0.041</td>
<td>3.756</td>
<td>0.05</td>
</tr>
<tr>
<td>MIBGC</td>
<td>0.495</td>
<td>3.837</td>
<td>0.537</td>
</tr>
<tr>
<td>GLEAM-ν2A</td>
<td>-0.026</td>
<td>6.180</td>
<td>0.123</td>
</tr>
<tr>
<td>GLEAM-ν2B</td>
<td>0.192</td>
<td>-3.571</td>
<td></td>
</tr>
<tr>
<td>GLEAM-ν3A</td>
<td>0.171</td>
<td>-5.483</td>
<td>0.151</td>
</tr>
<tr>
<td>PML</td>
<td>0.127</td>
<td>4.982</td>
<td>0.139</td>
</tr>
</tbody>
</table>

Table 3: (1) Bias of weighting products and (2) weights assigned to the bias corrected products in the case of each of the three DOLCE tiers, and the number of flux tower sites used to feed the weighting.
Figures

Figure 1: Spatial coverage of DOLCE tier 1 (red), regions added in tier 2 (orange) and tier 3 (yellow). Spatial coverage varies temporally; May 2000 is shown here as a representative example.

Figure 2: Location of the 159 flux tower sites involved in building DOLCE, classified by vegetation type given by International Geosphere-Biosphere Programme (IGBP): Croplands (CRO), Closed Shrublands (CSH), Deciduous Broadleaf Forest (DBF), Evergreen Broadleaf Forest (EBF), Evergreen Needleleaf Forest (ENF), Grasslands (GRA), Mixed Forests (MF), Open Shrublands (OSH), Savannas (SAV), Cropland/Natural Vegetation mosaic (VEG), Wetlands (WET), and Woody Savannas (WSA).
Figure 3: Box and whisker plots displaying the percentage improvement that the weighted product provides in the 25% out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean bias (d), when compared to the equally weighted mean of the Diagnostic Ensemble (Dmean), aggregated Reference Ensemble (Ragg) and each member of the reference ensemble. Box and whisker plots represent 5000 entries, each entry is generated through randomly selecting 25% of sites to be out sample.
Figure 4: In (a), (b), (c) and (d), as for Fig. 3 but showing the one site out-of-sample test. Box and whisker plots are generated through selecting one site to be out sample and are repeated for all 138 sites. Products marked with * have limited spatiotemporal availability relative to the diagnostic ensemble, and testing against the LFA, LFD, CS and PT products was limited to 110, 108, 108 and 72 sites respectively. In (e), (f), (g) and (h), the one site out-of-sample test is trained by HOM-case sites data only.
Figure 5: The global maps in the right column show the performance of DOLCE against in-situ measurements at each of the 142 sites by calculating four statistics: (a) RMSE, (b) Mean bias, (c) SD difference and (d) Correlation. The boxplots on the left display the results of these four metrics calculated separately for the sites that are satisfying the HOM-case and the HET-case.
Figure 6: Difference between the seasonal global ET from MPIBGC and DOLCE calculated over the period 2000–2009.
Figure 7: Difference between the seasonal average ET from LandFlux-EVAL-DIAG and DOLCE calculated over the period 2000–2005.
Figure 8: Seasonal (a) global ET and (b) its variability (standard deviation), (c) time average of uncertainty (the standard deviation uncertainty $\sqrt{\sigma^2}$ shown in Equation 11) (d) standard deviation of uncertainty over time (e) reliability, defined as high ($\left| \frac{\text{Uncertainty SD}}{\text{mean ET}} \right| \leq 1$ in blue), medium($\left| \text{mean ET} \right| \leq 5$, $\text{Uncertainty SD} < 10$ and $\left| \frac{\text{Uncertainty SD}}{\text{mean ET}} \right| \geq 1$ in green) and low (in red). DJF is shown in the left column and JJA in the right column. The global mean values in (a-d) are area weighted.
Figure 9: Four statistics, (a) RMSE, (b) Mean bias, (c) SD difference and (d) Correlation, calculated for DOLCE at 142 flux tower sites and displayed by biome types. See Fig. 2 for biome abbreviations.
Figure 10: Four statistics, (a) RMSE, (b) Mean bias, (c) SD difference and (d) Correlation, calculated for DOLCE separately at 129 flux towers located at the Northern Hemisphere excluding the tropics (NH) and at 11 towers in the Southern Hemisphere and the tropics (SH & TROPICS).
Figure 11: Box and whisker plots displaying the percentage improvement that the weighted product excluding MPIBGC provides in the 25% out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias (d), when compared to the equally weighted mean (Dmean) of the Diagnostic Ensemble, aggregated Reference Ensemble (Ragg) and each member of the reference ensemble. Box and whisker plots represent 5000 entries, each entry is generated through randomly selecting 25% of sites to be out sample.
Figure 12: Box and whisker plots displaying the percentage improvement that the weighted product excluding MPIBGC provides in the one site out-of-sample test for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias (d), when compared to the equally weighted mean (Dmean) of the Diagnostic Ensemble, aggregated Reference Ensemble (Ragg) and each member of the reference ensemble. Products marked with * have limited spatiotemporal availability relative to the diagnostic ensemble, and testing against the LFA, LFD, CS and PT products was limited to 110, 108, 108 and 72 sites respectively.
Table S1: Fluxnet sites used to calculate the weighting. 159 Fluxnet sites used to derive DOLCE. The IGBP Land Cover Type Classification at the site level is taken from the site investigator’s descriptions found in fluxnet.ornl.gov when available or otherwise from literature citation. Number of site-months indicates the number of site months available in this study for each site. The Dataset is the source of the Data, that is Fluxnet dataset published in Nov, 2016 (FN) or the 2007 LaThuile free fair use dataset (LT).

<table>
<thead>
<tr>
<th>Site code</th>
<th>Site name</th>
<th>Country</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Dataset</th>
<th>IGBP at the site level</th>
<th>Number of site-months</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT-Neu</td>
<td>Neustift/Stubai Valley</td>
<td>Austria</td>
<td>11.3175</td>
<td>47.11667</td>
<td>FN</td>
<td>GRA</td>
<td>96</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>AU-AdE</td>
<td>Adelaide River</td>
<td>Australia</td>
<td>131.11780</td>
<td>-13.07690</td>
<td>FN</td>
<td>WSA</td>
<td>19</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>AU-DaP</td>
<td>Daly River Pasture</td>
<td>Australia</td>
<td>131.31810</td>
<td>-14.063300</td>
<td>FN</td>
<td>GRA</td>
<td>25</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>AU-DaS</td>
<td>Daly River SAV</td>
<td>Australia</td>
<td>131.38800</td>
<td>-14.159283</td>
<td>FN</td>
<td>WSA</td>
<td>24</td>
<td>(Jamali et al., 2011)</td>
</tr>
<tr>
<td>AU-Dry</td>
<td>Dry River</td>
<td>Australia</td>
<td>132.37060</td>
<td>-15.258800</td>
<td>FN</td>
<td>WSA</td>
<td>12</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>AU-Fog</td>
<td>Fogg Dam</td>
<td>Australia</td>
<td>131.30720</td>
<td>-12.545200</td>
<td>FN</td>
<td>WET</td>
<td>32</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>AU-How</td>
<td>Howard Springs</td>
<td>Australia</td>
<td>131.15</td>
<td>-12.4952</td>
<td>LT</td>
<td>SAV</td>
<td>67</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>AU-Stp</td>
<td>Sturt Plains</td>
<td>Australia</td>
<td>133.35030</td>
<td>-17.150800</td>
<td>FN</td>
<td>GRA</td>
<td>13</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>AU-Tum</td>
<td>Tumbarumba</td>
<td>Australia</td>
<td>148.15160</td>
<td>-35.656600</td>
<td>FN</td>
<td>EBF</td>
<td>102</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>AU-Wac</td>
<td>Wallaby Creek</td>
<td>Australia</td>
<td>145.1873</td>
<td>-37.429</td>
<td>LT</td>
<td>EBF</td>
<td>32</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>BE-Bra</td>
<td>Brasschaat (De Inslag Forest)</td>
<td>Belgium</td>
<td>51.30917</td>
<td>4.52056</td>
<td>FN</td>
<td>MF</td>
<td>102</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>BE-Jal</td>
<td>Jalhay</td>
<td>Belgium</td>
<td>6.07333</td>
<td>50.56389</td>
<td>LT</td>
<td>MF</td>
<td>10</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>BE-Lon</td>
<td>Lonzee</td>
<td>Belgium</td>
<td>4.74613</td>
<td>50.55159</td>
<td>FN</td>
<td>CRO</td>
<td>69</td>
<td>(Moureaux, et al., 2006)</td>
</tr>
<tr>
<td>BE-Vie</td>
<td>VieIsalm</td>
<td>Belgium</td>
<td>5.99805</td>
<td>50.30507</td>
<td>FN</td>
<td>MF</td>
<td>116</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>BR-Sa3</td>
<td>Santarem-Km83-Logged Forest</td>
<td>Brazil</td>
<td>-54.9714</td>
<td>-3.01803</td>
<td>FN</td>
<td>EBF</td>
<td>44</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>Site code</td>
<td>Site name</td>
<td>Country</td>
<td>Longitude</td>
<td>Latitude</td>
<td>Dataset</td>
<td>IGBP at the site level</td>
<td>Number of site-months</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td>------------------------</td>
<td>------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>BW-Ma1</td>
<td>Maun-Mopane Woodland</td>
<td>Botswana</td>
<td>23.56033</td>
<td>-19.9165</td>
<td>LT</td>
<td>SAV</td>
<td>19</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>CA-Mer</td>
<td>Eastern Peatland-Mer Bleue</td>
<td>Canada</td>
<td>-75.5186</td>
<td>45.4094</td>
<td>LT</td>
<td>WET</td>
<td>70</td>
<td>(Lafleur, et al., 2003)</td>
</tr>
<tr>
<td>CA-Qcu</td>
<td>Quebec - Eastern Boreal, Black Spruce/Jack Pine Cutover</td>
<td>Canada</td>
<td>-74.0365</td>
<td>49.2671</td>
<td>LT</td>
<td>ENF</td>
<td>64</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>CA-Qfo</td>
<td>QC-Eastern Old Black Spruce (EOBS)</td>
<td>Canada</td>
<td>-74.3421</td>
<td>49.6925</td>
<td>FN</td>
<td>ENF</td>
<td>74</td>
<td>(Bergeron et al., 2007)</td>
</tr>
<tr>
<td>CA-SF1</td>
<td>Saskatchewan - Western Boreal, forest burned in 1977</td>
<td>Canada</td>
<td>-105.8176</td>
<td>54.4850</td>
<td>FN</td>
<td>ENF</td>
<td>40</td>
<td>(Mkhabela et al., 2009)</td>
</tr>
<tr>
<td>CA-SF2</td>
<td>Saskatchewan - Western Boreal, forest burned in 1989</td>
<td>Canada</td>
<td>-105.8775</td>
<td>54.2539</td>
<td>FN</td>
<td>ENF</td>
<td>48</td>
<td>(Mkhabela et al., 2009)</td>
</tr>
<tr>
<td>CH-Cha</td>
<td>Chamau</td>
<td>Switzerland</td>
<td>8.41044</td>
<td>47.210222</td>
<td>FN</td>
<td>GRA</td>
<td>44</td>
<td>(Merbold et al., 2014)</td>
</tr>
<tr>
<td>CH-Dav</td>
<td>Davos-Seehorn forest</td>
<td>Switzerland</td>
<td>9.8559</td>
<td>46.8153</td>
<td>FN</td>
<td>ENF</td>
<td>113</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>CH-Fru</td>
<td>Früebüel</td>
<td>Switzerland</td>
<td>8.53778</td>
<td>47.115833</td>
<td>FN</td>
<td>GRA</td>
<td>45</td>
<td>(Imer et al., 2013)</td>
</tr>
<tr>
<td>CN-Cng</td>
<td>Changling</td>
<td>China</td>
<td>123.5092</td>
<td>44.5934</td>
<td>FN</td>
<td>GRA</td>
<td>31</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>CN-Du2</td>
<td>DulongGRA</td>
<td>China</td>
<td>116.28360</td>
<td>42.046700</td>
<td>FN</td>
<td>GRA</td>
<td>25</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>CN-HaM</td>
<td>Haibei Alpine Tibet site</td>
<td>China</td>
<td>101.18</td>
<td>37.37</td>
<td>FN</td>
<td>GRA</td>
<td>36</td>
<td>(Li et al., 2013)</td>
</tr>
<tr>
<td>CZ-wet</td>
<td>CZECHWET</td>
<td>Czech Republic</td>
<td>14.77035</td>
<td>49.024650</td>
<td>LT</td>
<td>WET</td>
<td>9</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>Site code</td>
<td>Site name</td>
<td>Country</td>
<td>Longitude</td>
<td>Latitude</td>
<td>Dataset</td>
<td>IGBP at the site level</td>
<td>Number of site-months</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------------</td>
<td>--------------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td>------------------------</td>
<td>------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>DE-Geb</td>
<td>Gebesee</td>
<td>Germany</td>
<td>10.9143</td>
<td>51.1001</td>
<td>LT</td>
<td>CRO</td>
<td>107</td>
<td>(Revill et al., 2013)</td>
</tr>
<tr>
<td>DE-Gri</td>
<td>Grillenburg-grass station</td>
<td>Germany</td>
<td>13.51253</td>
<td>50.94947</td>
<td>FN</td>
<td>GRA</td>
<td>71</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>DE-Hai</td>
<td>Hainich</td>
<td>Germany</td>
<td>10.453</td>
<td>51.07917</td>
<td>FN</td>
<td>DBF</td>
<td>120</td>
<td>(Ershadi et al., 2014)</td>
</tr>
<tr>
<td>DE-Kli</td>
<td>Klingenberg - CRO</td>
<td>Germany</td>
<td>13.52251</td>
<td>50.89288</td>
<td>FN</td>
<td>CRO</td>
<td>60</td>
<td>(Revill et al., 2013)</td>
</tr>
<tr>
<td>DE-Lkb</td>
<td>Lackenberg</td>
<td>Germany</td>
<td>13.3047</td>
<td>49.0996</td>
<td>FN</td>
<td>ENF</td>
<td>7</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>DE-Meh</td>
<td>Mehrstedt 1</td>
<td>Germany</td>
<td>10.65547</td>
<td>51.27531</td>
<td>LT</td>
<td>MF</td>
<td>40</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>DE-Obe</td>
<td>Oberbärenburg</td>
<td>Germany</td>
<td>13.7196</td>
<td>50.7836</td>
<td>FN</td>
<td>ENF</td>
<td>23</td>
<td>(Jung et al., 2009)</td>
</tr>
<tr>
<td>DE-Seh</td>
<td>Selhausen</td>
<td>Germany</td>
<td>6.4496531</td>
<td>50.870623</td>
<td>FN</td>
<td>CRO</td>
<td>31</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>DE-Tha</td>
<td>Anchor Station Tharandt - old spruce</td>
<td>Germany</td>
<td>13.56694</td>
<td>50.96361</td>
<td>FN</td>
<td>ENF</td>
<td>120</td>
<td>(Wei et al., 2014)</td>
</tr>
<tr>
<td>DE-Wet</td>
<td>Wetzstein</td>
<td>Germany</td>
<td>11.45753</td>
<td>50.4535</td>
<td>LT</td>
<td>ENF</td>
<td>60</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>DK-Fou</td>
<td>Foulum</td>
<td>Denmark</td>
<td>9.58722</td>
<td>56.4842</td>
<td>LT</td>
<td>CRO</td>
<td>8</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>DK-Lva</td>
<td>Lille Valby (Rimi)</td>
<td>Denmark</td>
<td>12.08330</td>
<td>55.683300</td>
<td>LT</td>
<td>GRA</td>
<td>23</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>DK-Ris</td>
<td>Risbyholm</td>
<td>Denmark</td>
<td>12.09722</td>
<td>55.530278</td>
<td>LT</td>
<td>CRO</td>
<td>24</td>
<td>(Gilmanov et al., 2010)</td>
</tr>
<tr>
<td>DK-Sor</td>
<td>Soroe</td>
<td>Denmark</td>
<td>11.64464</td>
<td>55.485869</td>
<td>FN</td>
<td>DBF</td>
<td>102</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>DK-ZaF</td>
<td>Zackenberg Fen</td>
<td>Denmark</td>
<td>-20.5545</td>
<td>74.4814</td>
<td>FN</td>
<td>WET</td>
<td>3</td>
<td>(Soegaard and Nordstroem, 1999)</td>
</tr>
<tr>
<td>ES-ES1</td>
<td>El Saler</td>
<td>Spain</td>
<td>-0.31881</td>
<td>39.34597</td>
<td>LT</td>
<td>ENF</td>
<td>78</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>Site code</td>
<td>Site name</td>
<td>Country</td>
<td>Longitude</td>
<td>Latitude</td>
<td>Dataset</td>
<td>IGBP at the site level</td>
<td>Number of site-months</td>
<td>Reference</td>
</tr>
<tr>
<td>----------</td>
<td>-------------------</td>
<td>------------</td>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
<td>------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>ES-ES2</td>
<td>El Saler-Sueca</td>
<td>Spain</td>
<td>-0.31528</td>
<td>39.27556</td>
<td>LT</td>
<td>CRO</td>
<td>32</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>ES-LgS</td>
<td>Laguna Seca</td>
<td>Spain</td>
<td>-2.9658</td>
<td>37.0979</td>
<td>FN</td>
<td>OSH</td>
<td>30</td>
<td>fluent.ornl.gov</td>
</tr>
<tr>
<td>ES-LMa</td>
<td>Las Majadas del Tietar</td>
<td>Spain</td>
<td>-5.77336</td>
<td>39.9415</td>
<td>LT</td>
<td>SAV</td>
<td>36</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>ES-VDA</td>
<td>Vall d'Alinya</td>
<td>Spain</td>
<td>1.4485</td>
<td>42.15218</td>
<td>LT</td>
<td>GRA</td>
<td>28</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>FI-Hyy</td>
<td>Hyytiala</td>
<td>Finland</td>
<td>24.2950</td>
<td>61.8475</td>
<td>FN</td>
<td>ENF</td>
<td>80</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>FI-jok</td>
<td>Jokioinen</td>
<td>Finland</td>
<td>23.5135</td>
<td>60.8986</td>
<td>FN</td>
<td>CRO</td>
<td>39</td>
<td>(Reichstein et al., 2005)</td>
</tr>
<tr>
<td>FI-Kaat</td>
<td>Kaamanen wetland</td>
<td>Finland</td>
<td>27.2950</td>
<td>69.1407</td>
<td>FN</td>
<td>WET</td>
<td>82</td>
<td>(Aurela et al., 2001)</td>
</tr>
<tr>
<td>FI-Lom</td>
<td>Lompolojänkkä</td>
<td>Finland</td>
<td>24.2092</td>
<td>67.9972</td>
<td>FN</td>
<td>WET</td>
<td>36</td>
<td>fluent.ornl.gov</td>
</tr>
<tr>
<td>FI-Sod</td>
<td>Sodankyla</td>
<td>Finland</td>
<td>26.6378</td>
<td>67.3619</td>
<td>FN</td>
<td>ENF</td>
<td>110</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>FR-Gri</td>
<td>Grignon</td>
<td>France</td>
<td>1.95191</td>
<td>48.84422</td>
<td>FN</td>
<td>CRO</td>
<td>68</td>
<td>(Loubet et al., 2011)</td>
</tr>
<tr>
<td>FR-Hes</td>
<td>Hesse Forest-Sarrebourg</td>
<td>France</td>
<td>7.06556</td>
<td>48.67416</td>
<td>LT</td>
<td>DBF</td>
<td>80</td>
<td>(Reichstein et al., 2005)</td>
</tr>
<tr>
<td>FR-LBr</td>
<td>Le Bray (after 6/28/1998)</td>
<td>France</td>
<td>-0.7693</td>
<td>44.71711</td>
<td>LT</td>
<td>ENF</td>
<td>102</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>FR-Lq1</td>
<td>Laqueuille</td>
<td>France</td>
<td>2.73583</td>
<td>45.64306</td>
<td>LT</td>
<td>GRA</td>
<td>36</td>
<td>(Gilmanov et al., 2007)</td>
</tr>
<tr>
<td>FR-Lq2</td>
<td>Laqueuille extensive</td>
<td>France</td>
<td>2.73703</td>
<td>45.63919</td>
<td>LT</td>
<td>GRA</td>
<td>36</td>
<td>(Gilmanov et al., 2007)</td>
</tr>
<tr>
<td>FR-Pue</td>
<td>Puechabon</td>
<td>France</td>
<td>3.59583</td>
<td>43.741390</td>
<td>FN</td>
<td>EBF</td>
<td>83</td>
<td>(Wei et al., 2014)</td>
</tr>
<tr>
<td>HU-Bug</td>
<td>Bugacpuszta</td>
<td>Hungary</td>
<td>19.6013</td>
<td>46.6911</td>
<td>LT</td>
<td>GRA</td>
<td>50</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>HU-Mat</td>
<td>Matra</td>
<td>Hungary</td>
<td>19.726</td>
<td>47.8469</td>
<td>LT</td>
<td>GRA</td>
<td>32</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>Site code</td>
<td>Site name</td>
<td>Country</td>
<td>Longitude</td>
<td>Latitude</td>
<td>Dataset</td>
<td>IGBP at the site level</td>
<td>Number of site-months</td>
<td>Reference</td>
</tr>
<tr>
<td>----------</td>
<td>---------------------</td>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td>------------------------</td>
<td>-----------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>ID-Pag</td>
<td>Palangkaraya (PDF)</td>
<td>Indonesia</td>
<td>114.036</td>
<td>-2.345</td>
<td>LT</td>
<td>EBF</td>
<td>18</td>
<td>(Hirano et al., 2007)</td>
</tr>
<tr>
<td>IE-Ca1</td>
<td>Carlow1</td>
<td>Ireland</td>
<td>-6.9181</td>
<td>52.8588</td>
<td>LT</td>
<td>CRO</td>
<td>33</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>IE-Dri1</td>
<td>Dripsey</td>
<td>Ireland</td>
<td>-8.75181</td>
<td>51.986694</td>
<td>LT</td>
<td>GRA</td>
<td>31</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>IL-Yat1</td>
<td>Yatir</td>
<td>Israel</td>
<td>35.0515</td>
<td>31.3450</td>
<td>FN</td>
<td>ENF</td>
<td>67</td>
<td>(Reichstein et al., 2003)</td>
</tr>
<tr>
<td>IT-Amp</td>
<td>Amplero</td>
<td>Italy</td>
<td>13.60516</td>
<td>41.9041</td>
<td>LT</td>
<td>GRA</td>
<td>54</td>
<td>(Gilmanov et al., 2007)</td>
</tr>
<tr>
<td>IT-Cas</td>
<td>Castellarano</td>
<td>Italy</td>
<td>8.7175222</td>
<td>45.0700472</td>
<td>LT</td>
<td>CRO</td>
<td>6</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>IT-Col</td>
<td>Collelongo-Piana</td>
<td>Italy</td>
<td>13.58814</td>
<td>41.84936</td>
<td>LT</td>
<td>DBF</td>
<td>72</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>IT-Cpz</td>
<td>Castelporziano</td>
<td>Italy</td>
<td>12.3761</td>
<td>41.7052</td>
<td>FN</td>
<td>EBF</td>
<td>94</td>
<td>(Wei et al., 2014)</td>
</tr>
<tr>
<td>IT-Lec</td>
<td>Lecceto</td>
<td>Italy</td>
<td>11.26975</td>
<td>43.30359</td>
<td>LT</td>
<td>EBF</td>
<td>17</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>IT-LMa</td>
<td>La Mandria</td>
<td>Italy</td>
<td>7.58259</td>
<td>45.15258</td>
<td>LT</td>
<td>DBF</td>
<td>30</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>IT-Mal</td>
<td>Malga Arpaco</td>
<td>Italy</td>
<td>11.70334</td>
<td>46.11402</td>
<td>LT</td>
<td>GRA</td>
<td>32</td>
<td>(Gilmanov et al., 2007)</td>
</tr>
<tr>
<td>IT-MBo</td>
<td>Monte Bondone</td>
<td>Italy</td>
<td>11.04583</td>
<td>46.01468</td>
<td>LT</td>
<td>GRA</td>
<td>84</td>
<td>(Gilmanov et al., 2007)</td>
</tr>
<tr>
<td>IT-Non</td>
<td>Nonantola</td>
<td>Italy</td>
<td>11.09109</td>
<td>44.69019</td>
<td>LT</td>
<td>MF</td>
<td>40</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>IT-PT1</td>
<td>Zerbolo-Parco Ticino-Canarrazzo</td>
<td>Italy</td>
<td>9.06104</td>
<td>45.20087</td>
<td>FN</td>
<td>DBF</td>
<td>31</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>IT-Ren</td>
<td>Renon/Ritten (Bolzano)</td>
<td>Italy</td>
<td>11.43369</td>
<td>46.58686</td>
<td>FN</td>
<td>ENF</td>
<td>80</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>IT-Ro2</td>
<td>Roccarespampani 2</td>
<td>Italy</td>
<td>11.92093</td>
<td>42.390260</td>
<td>FN</td>
<td>DBF</td>
<td>77</td>
<td>(Tedeschi et al., 2014)</td>
</tr>
<tr>
<td>Site code</td>
<td>Site name</td>
<td>Country</td>
<td>Longitude</td>
<td>Latitude</td>
<td>Dataset</td>
<td>IGBP at the site level</td>
<td>Number of site-months</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------</td>
<td>---------</td>
<td>-----------</td>
<td>------------</td>
<td>---------</td>
<td>------------------------</td>
<td>-----------------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>IT-SRo</td>
<td>San Rossore</td>
<td>Italy</td>
<td>10.2844</td>
<td>43.727861</td>
<td>FN</td>
<td>ENF</td>
<td>109</td>
<td>(Reichstein et al., 2005)</td>
</tr>
<tr>
<td>IT-Tor</td>
<td>Torgnon</td>
<td>Italy</td>
<td>7.5781</td>
<td>45.8444</td>
<td>FN</td>
<td>GRA</td>
<td>16</td>
<td>(Galvagno et al., 2013)</td>
</tr>
<tr>
<td>NL-Ca1</td>
<td>Cabauw</td>
<td>Netherland</td>
<td>4.927</td>
<td>51.971</td>
<td>LT</td>
<td>GRA</td>
<td>46</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>NL-Haa</td>
<td>Haastrecht</td>
<td>Netherland</td>
<td>4.80556</td>
<td>52.003611</td>
<td>LT</td>
<td>GRA</td>
<td>10</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>NL-Hor</td>
<td>Horstermeerv</td>
<td>Netherland</td>
<td>5.07130</td>
<td>52.240350</td>
<td>FN</td>
<td>GRA</td>
<td>50</td>
<td>(Sulkava et al., 2011)</td>
</tr>
<tr>
<td>NL-Lan</td>
<td>Langerak</td>
<td>Netherland</td>
<td>4.9029</td>
<td>51.9536</td>
<td>LT</td>
<td>CRO</td>
<td>11</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>NL-Loo</td>
<td>Loobos</td>
<td>Netherland</td>
<td>5.74356</td>
<td>52.166581</td>
<td>FN</td>
<td>ENF</td>
<td>119</td>
<td>(Gash and Dolman, 2003)</td>
</tr>
<tr>
<td>NL-Lut</td>
<td>Lutjewad</td>
<td>Netherland</td>
<td>6.35603</td>
<td>53.398922</td>
<td>LT</td>
<td>CRO</td>
<td>9</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>NL-Mol</td>
<td>Molenweg</td>
<td>Netherland</td>
<td>4.63908</td>
<td>51.65</td>
<td>LT</td>
<td>CRO</td>
<td>7</td>
<td>(Gilmanov et al., 2007)</td>
</tr>
<tr>
<td>PL-Wet</td>
<td>Polwet</td>
<td>Poland</td>
<td>16.3094</td>
<td>52.7622</td>
<td>LT</td>
<td>WET</td>
<td>24</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>PT-Esp</td>
<td>Espirra</td>
<td>Portugal</td>
<td>-8.6018</td>
<td>38.6394</td>
<td>LT</td>
<td>EBF</td>
<td>43</td>
<td>(Wei et al., 2014)</td>
</tr>
<tr>
<td>PT-Mi1</td>
<td>Mitra (Evora)</td>
<td>Portugal</td>
<td>-8.00006</td>
<td>38.54064</td>
<td>LT</td>
<td>EBF</td>
<td>20</td>
<td>(Reichstein et al., 2003)</td>
</tr>
<tr>
<td>PT-Mi2</td>
<td>Mitra IV Tojal</td>
<td>Portugal</td>
<td>-8.02455</td>
<td>38.4765</td>
<td>LT</td>
<td>GRA</td>
<td>29</td>
<td>(Gilmanov et al., 2007)</td>
</tr>
<tr>
<td>RU-Cok</td>
<td>Chokurdakh</td>
<td>Russia</td>
<td>147.4943</td>
<td>70.8291</td>
<td>FN</td>
<td>OSH</td>
<td>29</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>RU-Fyo</td>
<td>Fyodorovskoye wet spruce stand</td>
<td>Russia</td>
<td>32.92208</td>
<td>56.46153</td>
<td>FN</td>
<td>ENF</td>
<td>118</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>RU-Ha1</td>
<td>Hakasia steppe</td>
<td>Russia</td>
<td>90.00215</td>
<td>54.72517</td>
<td>FN</td>
<td>GRA</td>
<td>19</td>
<td>(Belelli Marchesini)</td>
</tr>
<tr>
<td>Site code</td>
<td>Site name</td>
<td>Country</td>
<td>Longitude</td>
<td>Latitude</td>
<td>Dataset</td>
<td>IGBP at the site level</td>
<td>Number of site-months</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td>-------------------------</td>
<td>------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>RU-Ha2</td>
<td>Ubs Nur- Hakasija-recovering grassland</td>
<td>Russia</td>
<td>89.95664</td>
<td>54.77301</td>
<td>LT</td>
<td>GRA</td>
<td>8</td>
<td>(Belelli Marchesini et al., 2007)</td>
</tr>
<tr>
<td>RU-Ha3</td>
<td>Ubs Nur-Hakasija-Site 3</td>
<td>Russia</td>
<td>89.07785</td>
<td>54.70455</td>
<td>LT</td>
<td>GRA</td>
<td>7</td>
<td>(Belelli Marchesini et al., 2007)</td>
</tr>
<tr>
<td>RU-Zot</td>
<td>Zotino</td>
<td>Russia</td>
<td>89.3508</td>
<td>60.8008</td>
<td>FN</td>
<td>ENF</td>
<td>19</td>
<td>(Eugster et al., 2000)</td>
</tr>
<tr>
<td>SE-Deg</td>
<td>Degero</td>
<td>Sweden</td>
<td>19.5500</td>
<td>64.1833</td>
<td>FN</td>
<td>GRA</td>
<td>56</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>SE-Faj</td>
<td>Fajemyr</td>
<td>Sweden</td>
<td>13.55351</td>
<td>56.26547</td>
<td>LT</td>
<td>WET</td>
<td>12</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>SE-Fla</td>
<td>Flakaliden</td>
<td>Sweden</td>
<td>19.4569</td>
<td>64.1128</td>
<td>FN</td>
<td>ENF</td>
<td>35</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>SE-Nor</td>
<td>Norunda</td>
<td>Sweden</td>
<td>17.4795</td>
<td>60.0865</td>
<td>FN</td>
<td>ENF</td>
<td>24</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>SE-Sk1</td>
<td>Skyttorp1 young</td>
<td>Sweden</td>
<td>17.9181</td>
<td>60.1250</td>
<td>FN</td>
<td>ENF</td>
<td>11</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>SE-Sk2</td>
<td>Skyttorp</td>
<td>Sweden</td>
<td>17.8401</td>
<td>60.1297</td>
<td>FN</td>
<td>ENF</td>
<td>15</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>SK-Tat</td>
<td>Tatra</td>
<td>Slovak Republic</td>
<td>20.1635</td>
<td>49.12078</td>
<td>LT</td>
<td>ENF</td>
<td>4</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>UK-AMo</td>
<td>Auchencorth Moss-Scotland</td>
<td>UK</td>
<td>-3.23889</td>
<td>55.79167</td>
<td>LT</td>
<td>GRA</td>
<td>8</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>UK-EBu</td>
<td>Easter Bush-Scotland</td>
<td>UK</td>
<td>-3.20578</td>
<td>55.866</td>
<td>LT</td>
<td>GRA</td>
<td>16</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>UK-ESa</td>
<td>East Saltoun</td>
<td>UK</td>
<td>-2.85861</td>
<td>55.90694</td>
<td>LT</td>
<td>CRO</td>
<td>22</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>UK-Gri</td>
<td>Griffin- Aberfeldy-Scotland</td>
<td>UK</td>
<td>-3.7981</td>
<td>56.6072</td>
<td>FN</td>
<td>ENF</td>
<td>35</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>UK-Ham</td>
<td>Hampshire</td>
<td>UK</td>
<td>-0.8583</td>
<td>51.5353</td>
<td>LT</td>
<td>DBF</td>
<td>22</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>UK-Her</td>
<td>Hertfordshire</td>
<td>UK</td>
<td>-0.476080</td>
<td>51.7837982</td>
<td>LT</td>
<td>CRO</td>
<td>6</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>Site code</td>
<td>Site name</td>
<td>Country</td>
<td>Longitude</td>
<td>Latitude</td>
<td>Dataset</td>
<td>IGBP at the site level</td>
<td>Number of site-months</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------------</td>
<td>---------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td>------------------------</td>
<td>------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>UK-PL3</td>
<td>Pang/ Lambourne (forest)</td>
<td>UK</td>
<td>-1.26667</td>
<td>51.45</td>
<td>LT</td>
<td>DBF</td>
<td>24</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>UK-Tad</td>
<td>Tadham Moor</td>
<td>UK</td>
<td>-2.82864</td>
<td>51.2071</td>
<td>LT</td>
<td>GRA</td>
<td>9</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>US-AR1</td>
<td>ARM USDA UNL OSU Woodward Switchgrass 1</td>
<td>USA</td>
<td>-99.4200</td>
<td>36.4267</td>
<td>FN</td>
<td>GRA</td>
<td>8</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>US-ARb</td>
<td>ARM Southern Great Plains burn site- Lamont</td>
<td>USA</td>
<td>-98.0402</td>
<td>35.5497</td>
<td>FN</td>
<td>GRA</td>
<td>20</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>US-ARc</td>
<td>ARM Southern Great Plains control site- Lamont</td>
<td>USA</td>
<td>-98.0400</td>
<td>35.5465</td>
<td>FN</td>
<td>GRA</td>
<td>20</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>US-ARM</td>
<td>OK - ARM Southern Great Plains site- Lamont</td>
<td>USA</td>
<td>-97.4888</td>
<td>36.6058</td>
<td>FN</td>
<td>CRO</td>
<td>84</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>US-Aud</td>
<td>AZ - Audubon Research Ranch</td>
<td>USA</td>
<td>-110.51</td>
<td>31.5907</td>
<td>LT</td>
<td>GRA</td>
<td>49</td>
<td>(Hilton et al., 2014)</td>
</tr>
<tr>
<td>US-Bar</td>
<td>NH - Bartlett Experimental Forest</td>
<td>USA</td>
<td>-71.2881</td>
<td>44.0646</td>
<td>LT</td>
<td>MF</td>
<td>23</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>US-Bkg</td>
<td>Brookings</td>
<td>USA</td>
<td>-96.8362</td>
<td>44.3453</td>
<td>FN</td>
<td>GRA</td>
<td>31</td>
<td>(Gilmanov et al., 2005)</td>
</tr>
<tr>
<td>US-Blo</td>
<td>CA - Blodgett Forest</td>
<td>USA</td>
<td>-120.633</td>
<td>38.8952</td>
<td>FN</td>
<td>ENF</td>
<td>91</td>
<td>(Reichstein et al., 2003)</td>
</tr>
<tr>
<td>US-Bo1</td>
<td>IL - Bondville</td>
<td>USA</td>
<td>-88.2904</td>
<td>40.0062</td>
<td>LT</td>
<td>CRO</td>
<td>83</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>US-Cop</td>
<td>Corral Pocket</td>
<td>USA</td>
<td>-109.3900</td>
<td>38.0900</td>
<td>FN</td>
<td>GRA</td>
<td>48</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>US-FPe</td>
<td>MT - Fort Peck</td>
<td>USA</td>
<td>-105.102</td>
<td>48.3077</td>
<td>LT</td>
<td>GRA</td>
<td>75</td>
<td>(Ershadi et al., 2014)</td>
</tr>
<tr>
<td>Site code</td>
<td>Site name</td>
<td>Country</td>
<td>Longitude</td>
<td>Latitude</td>
<td>Dataset</td>
<td>IGBP at the site level</td>
<td>Number of site-months</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td>-----------------------</td>
<td>----------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>US-GBT</td>
<td>GLEES</td>
<td>USA</td>
<td>-106.2397</td>
<td>41.3658</td>
<td>FN</td>
<td>ENF</td>
<td>63</td>
<td>(Zeller and Henn, 1996)</td>
</tr>
<tr>
<td>US-GLE</td>
<td>GLEES</td>
<td>USA</td>
<td>-106.2399</td>
<td>41.3665</td>
<td>FN</td>
<td>ENF</td>
<td>57</td>
<td>(Frank et al., 2014)</td>
</tr>
<tr>
<td>US-Goo</td>
<td>MS - Goodwin Creek</td>
<td>USA</td>
<td>-89.8735</td>
<td>34.2547</td>
<td>LT</td>
<td>VEG</td>
<td>48</td>
<td>fluxnet.ornl.gov</td>
</tr>
<tr>
<td>US-Ho1</td>
<td>ME - Howland Forest (main tower)</td>
<td>USA</td>
<td>-68.7402</td>
<td>45.2041</td>
<td>LT</td>
<td>MF</td>
<td>60</td>
<td>(Wei et al., 2014)</td>
</tr>
<tr>
<td>US-Ho2</td>
<td>ME - Howland Forest (west tower)</td>
<td>USA</td>
<td>-68.747</td>
<td>45.2091</td>
<td>LT</td>
<td>MF</td>
<td>60</td>
<td>(Wei et al., 2014)</td>
</tr>
<tr>
<td>US-Los</td>
<td>WI - Lost Creek</td>
<td>USA</td>
<td>-89.9792</td>
<td>46.0827</td>
<td>FN</td>
<td>CSH</td>
<td>98</td>
<td>(Sulman et al., 2010)</td>
</tr>
<tr>
<td>US-Me1</td>
<td>Metolius - Eyerly burn</td>
<td>USA</td>
<td>-121.5000</td>
<td>44.5794</td>
<td>FN</td>
<td>ENF</td>
<td>10</td>
<td>(Irvine &amp; Hibbard, 2007)</td>
</tr>
<tr>
<td>US-Me2</td>
<td>Metolius- intermediate aged ponderosa pine</td>
<td>USA</td>
<td>-121.5574</td>
<td>44.4523</td>
<td>FN</td>
<td>ENF</td>
<td>89</td>
<td>(Irvine &amp; Hibbard, 2007)</td>
</tr>
<tr>
<td>US-Me4</td>
<td>Metolius-old aged ponderosa pine</td>
<td>USA</td>
<td>-121.6224</td>
<td>44.4992</td>
<td>FN</td>
<td>ENF</td>
<td>12</td>
<td>(Irvine &amp; Hibbard, 2007)</td>
</tr>
<tr>
<td>US-MMS</td>
<td>IN - Morgan Monroe State Forest</td>
<td>USA</td>
<td>-86.4131</td>
<td>39.3231</td>
<td>FN</td>
<td>DBF</td>
<td>120</td>
<td>(Baldocchi et al., 2001)</td>
</tr>
<tr>
<td>US-MOz</td>
<td>MO - Missouri Ozark Site</td>
<td>USA</td>
<td>-92.2</td>
<td>38.7441</td>
<td>LT</td>
<td>DBF</td>
<td>31</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>US-Ne3</td>
<td>NE - Mead - rainfed maize-soybean rotation site</td>
<td>USA</td>
<td>-96.4397</td>
<td>41.1797</td>
<td>FN</td>
<td>CRO</td>
<td>103</td>
<td>(Verma et al., 2005)</td>
</tr>
<tr>
<td>US-NR1</td>
<td>Niwot Ridge Forest (LTER NWT1)</td>
<td>USA</td>
<td>-105.54600</td>
<td>40.032902</td>
<td>FN</td>
<td>ENF</td>
<td>120</td>
<td>(Monson et al., 2002)</td>
</tr>
<tr>
<td>US-Ohio</td>
<td>OH - Oak Openings</td>
<td>USA</td>
<td>-83.8438</td>
<td>41.5545</td>
<td>LT</td>
<td>DBF</td>
<td>24</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>US-SP1</td>
<td>FL - Slashpine- Austin Cary- 65yrs</td>
<td>USA</td>
<td>-82.2188</td>
<td>29.7381</td>
<td>LT</td>
<td>ENF</td>
<td>16</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>Site code</td>
<td>Site name</td>
<td>Country</td>
<td>Longitude</td>
<td>Latitude</td>
<td>Dataset</td>
<td>IGBP at the site level</td>
<td>Number of site-months</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------------------------------------------------------</td>
<td>---------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td>-----------------------</td>
<td>----------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>US-SP2</td>
<td>FL - Slashpine-Mize-clearcut-3yr.regen</td>
<td>USA</td>
<td>-82.2448</td>
<td>29.7648</td>
<td>LT</td>
<td>ENF</td>
<td>60</td>
<td>(Hilton et al., 2014)</td>
</tr>
<tr>
<td>US-SP3</td>
<td>FL - Slashpine-Donaldson-mid-rot-12yrs</td>
<td>USA</td>
<td>-82.1633</td>
<td>29.7548</td>
<td>LT</td>
<td>ENF</td>
<td>54</td>
<td>(Hilton et al., 2014)</td>
</tr>
<tr>
<td>US-SRG</td>
<td>Santa Rita GRA</td>
<td>USA</td>
<td>-110.8277</td>
<td>31.7894</td>
<td>FN</td>
<td>GRA</td>
<td>21</td>
<td>(Scott et al., 2010)</td>
</tr>
<tr>
<td>US-SRM</td>
<td>AZ - Santa Rita Mesquite</td>
<td>USA</td>
<td>-110.866</td>
<td>31.8214</td>
<td>FN</td>
<td>WSA</td>
<td>72</td>
<td>(Scott et al., 2010)</td>
</tr>
<tr>
<td>US-Syv</td>
<td>MI - Sylvania Wilderness Area</td>
<td>USA</td>
<td>-89.3477</td>
<td>46.242</td>
<td>FN</td>
<td>MF</td>
<td>65</td>
<td>(Desai et al., 2005)</td>
</tr>
<tr>
<td>US-Ton</td>
<td>CA - Tonzi Ranch</td>
<td>USA</td>
<td>-120.966</td>
<td>38.4316</td>
<td>FN</td>
<td>DBF</td>
<td>102</td>
<td>(Vargas et al., 2011)</td>
</tr>
<tr>
<td>US-Var</td>
<td>CA - Vaira Ranch-Ione</td>
<td>USA</td>
<td>-120.951</td>
<td>38.4133</td>
<td>FN</td>
<td>GRA</td>
<td>109</td>
<td>(Ma et al., 2007)</td>
</tr>
<tr>
<td>US-WCr</td>
<td>WI - Willow Creek</td>
<td>USA</td>
<td>-90.0799</td>
<td>45.8059</td>
<td>FN</td>
<td>DBF</td>
<td>79</td>
<td>(Cook et al., 2004)</td>
</tr>
<tr>
<td>US-Whs</td>
<td>Walnut Gulch Lucky Hills Shrub</td>
<td>USA</td>
<td>-110.0522</td>
<td>31.7438</td>
<td>FN</td>
<td>OSH</td>
<td>30</td>
<td>(Cavanaugh et al., 2011)</td>
</tr>
<tr>
<td>US-Wi0</td>
<td>WI - Young red pine (YRP)</td>
<td>USA</td>
<td>-91.0814</td>
<td>46.6188</td>
<td>LT</td>
<td>ENF</td>
<td>6</td>
<td>(Noormets et al., 2007)</td>
</tr>
<tr>
<td>US-Wi1</td>
<td>WI - Intermediate hardwood</td>
<td>USA</td>
<td>-91.2329</td>
<td>46.7305</td>
<td>LT</td>
<td>DBF</td>
<td>5</td>
<td>(Noormets et al., 2007)</td>
</tr>
<tr>
<td>US-Wi3</td>
<td>Mature hardwood</td>
<td>USA</td>
<td>-91.0987</td>
<td>46.6347</td>
<td>FN</td>
<td>DBF</td>
<td>16</td>
<td>(Noormets et al., 2007)</td>
</tr>
<tr>
<td>US-Wi4</td>
<td>WI - Mature red pine</td>
<td>USA</td>
<td>-91.1663</td>
<td>46.7393</td>
<td>LT</td>
<td>ENF</td>
<td>23</td>
<td>(Hilton et al., 2014)</td>
</tr>
<tr>
<td>US-Wi5</td>
<td>WI - Mixed young jack pine</td>
<td>USA</td>
<td>-91.0858</td>
<td>46.6531</td>
<td>LT</td>
<td>ENF</td>
<td>7</td>
<td>(Stoy et al., 2013)</td>
</tr>
<tr>
<td>US-Wi6</td>
<td>WI - Pine barrens #1</td>
<td>USA</td>
<td>-91.2982</td>
<td>46.6249</td>
<td>LT</td>
<td>OSH</td>
<td>11</td>
<td>(Noormets et al., 2007)</td>
</tr>
<tr>
<td>US-Wi7</td>
<td>WI - Red pine</td>
<td>USA</td>
<td>-91.0693</td>
<td>46.6491</td>
<td>LT</td>
<td>MF</td>
<td>5</td>
<td>fluxnet.ornl.</td>
</tr>
</tbody>
</table>
### In-sample test

The in-sample test shows that the weighting approach is monotonically not increasing (MSE) and that it converges to the same results no matter where we start. This confirms that the weighting technique always builds a better performing ET estimates from an ensemble of ET products relative to the in-sample observational dataset (Fig. S1).

### Clustering test structure and results

We assigned one of the five following biome groups to each flux tower site. Each biome group is a composite of two or more land cover types. We performed the combination based on the similarity of the combined components, by doing that we increase the available number of sites in each cluster:

- MF DBF SH: sites covered by mixed forest (MF), deciduous broadleaf forest (DBF) or Shrublands (Open (OSH) or closed (CSH))
- ENF EBF: includes Evergreen Needleleaf Forest (ENF) and Evergreen Broadleaf Forest (EBF) sites
- SAV WSA GRA: includes sites covered by Savannas (SAV), Woody Savannas (WSA) or Grasslands
- CRO VEG: includes Croplands (CRO) and Grasslands (GRA)
- Remaining: Sites that don’t belong to any of the four groups defined above were put together in a “remaining group”.

In each particular cluster, one site was taken out and the remaining sites were used to train the weighting and derive a set of bias correction terms and weights. The mean of the weighting ensemble was calculated and the derived cluster-dependent weighted product was evaluated against the out-sample site in a particular cluster. This has resulted in calculating the statistical metrics MSE, MRSD, COR and Mean Bias – described in Sec. 2.4 separately at every cluster. In a parallel analysis, a non-clustered weighting was performed as explained in Sec. 2.4. The weights generated in the non-clustered

<table>
<thead>
<tr>
<th>Site code</th>
<th>Site name</th>
<th>Country</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Dataset</th>
<th>IGBP at the site level</th>
<th>Number of site-months</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>US-Wi8</td>
<td>WI - Young hardwood clearcut</td>
<td>USA</td>
<td>-91.2524</td>
<td>46.7223</td>
<td>LT</td>
<td>DBF</td>
<td>5</td>
<td>(Noormets et al., 2007)</td>
</tr>
<tr>
<td>US-Wkg</td>
<td>AZ - Walnut Gulch Kendall grasslands</td>
<td>USA</td>
<td>-109.942</td>
<td>31.7365</td>
<td>FN</td>
<td>GRA</td>
<td>68</td>
<td>(Scott et al., 2010)</td>
</tr>
<tr>
<td>ZA-Kru</td>
<td>Skukuza Kruger National Park</td>
<td>South Africa</td>
<td>31.4969</td>
<td>-25.0197</td>
<td>FN</td>
<td>SAV</td>
<td>69</td>
<td>(King et al., 2003)</td>
</tr>
<tr>
<td>ZM-Mon</td>
<td>Mongu</td>
<td>Zambia</td>
<td>23.2528</td>
<td>-15.4378</td>
<td>FN</td>
<td>DBF</td>
<td>22</td>
<td>(King et al., 2003)</td>
</tr>
</tbody>
</table>
weighting were assigned to the participating products at the sites belonging to the remaining group in the clustered weighting. The weighting was repeated for all the sites being taken out of sample.

In the clustering test, we looked at:

1. The performance improvement (%) of the cluster-independent weighted product against the mean of the diagnostic ensemble (No CL).
   
2. The performance improvement (%) of the cluster-independent weighted product against each individual dataset in the Reference Ensemble, this yielded 10 different values of performance improvement. Then, these values were aggregated in one boxplot (No CL Agg).
   
3. The performance improvement (%) of the cluster-dependent weighted product against the mean of the diagnostic ensemble (CL).
   
4. The performance improvement (%) of the cluster-dependent weighted product against each individual dataset in the Reference Ensemble, this yielded 10 different values of performance improvement. Then, these values were aggregated in one boxplot (CL Agg).
   
5. The performance improvement (%) of the cluster-dependent weighted product against the mean of the diagnostic ensemble at a particular cluster. This was calculated for all the clusters.
   
6. The performance improvement (%) of the cluster-dependent weighted product against each individual dataset in Ref at a particular cluster, this yielded 10 different values of performance improvement for each cluster, which were then aggregated and displayed as a single boxplot for every cluster.

The results of the clustered test in Fig. S2 show 1) a significant difference in the performance across the individual biome groups and that 2) the median value of the cluster-dependent weighting has slightly decreased in CL-5 and CL-5 Agg compared to No CL and No CL Agg respectively. (1) and (2) indicate that while the clustered weighting works better over some biomes and for some metrics, overall the biome types classification doesn’t help to better align to the flux behaviour.
Figure S1: The results of the in-sample test showing how the Mean Square Error (MSE) of the weighting changes when we increase the number of ET products involved in the weighting from 1 to 6. The test was repeated for 25 random permutations of the weighting products.
Figure S2: Box and whisker plots displaying the results of the One out-of-sample site test for the cluster independent weighting (No CL and No CL Agg boxplots), the cluster dependent weighting (Cl-5 and CL-5 Agg) and over individual biome types for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias (d).

Table S2: Distribution by land cover of HOM-case sites and HET-case sites at both the site scale and grid cell scale

<table>
<thead>
<tr>
<th>Land Cover</th>
<th>HOM-case (site)</th>
<th>HET-case (site)</th>
<th>HET-case (grid cell)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRO</td>
<td>10</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>CSH</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DBF</td>
<td>1</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>EBF</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>ENF</td>
<td>6</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>GRA</td>
<td>13</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>MF</td>
<td>7</td>
<td>3</td>
<td>32</td>
</tr>
<tr>
<td>OSH</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>SAV</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>VEG</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>WET</td>
<td></td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>WSA</td>
<td></td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Wa (Water)</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>URB (Urban)</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Figure S3: Box and whisker plots displaying the percentage improvement that the weighted product (in black) and the bias corrected mean of the diagnostic ensemble (in red) provide in the one out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean bias (d), when compared to equally weighted mean of the Diagnostic Ensemble (Dmean), aggregated Reference Ensemble (Ragg) and each member of the reference ensemble.
Figure S4: Difference between the seasonal global ET (a) from MPIBGC and DOLCE and (b) from MPIBGC and the simple mean of the merged products, calculated over the period 2000–2009.
**Figure S4:** Time series at five sites showing the products that were used to derive DOLCE tier1 before bias correction (right) and after bias correction with the resulting merge (DOLCE) (left).

References


Noormets, A., Chen, J. and Crow, T. R.: Age-Dependent Changes in Ecosystem Carbon Fluxes in Managed Forests in


