Manuscript hess-2017-147 entitled “Derived Optimal Linear Combination Evapotranspiration (DOLCE): a global gridded synthesis ET estimate”

We would like to thank the reviewers for their constructive comments on our manuscript. This document outlines our point-by-point responses to the reviewer #2 comments and the improvements made to the manuscript. We’ve also added all the modified plots and tables at the end of this document.

Response to Reviewer #2

1. **I have questions about why do you choose to produce DOLCE at 0.5 deg. resolution? Not 0.25 deg or 0.05 deg (the possibilities provided by MOD16, GLEAM ET), even as high as 1km (the possibility will discussed later in my report). I need a strong rebuttal from the authors here.**

   It would be fantastic to produce a meaningful 0.05 degree global ET product, but among all the weighting products, only MOD16 has information at this scale. It could be possible to produce an ET product at 0.25 degree, but both PML and MPI have no information at this scale. One of the benefits of DOLCE is that it combines the strengths of many estimates, one of the limitations is that it requires many estimates in order to provide better information. So, given the very small number of datasets available at high-resolution DOLCE would likely not provide better information.

2. **In addition, before producing DOLCE, are the weighting ETs resampled to 0.5 degree? Or how did you combine ET at different resolution into 0.5 degree DOLCE?**

We thank the reviewer for their comment, this was indeed not mentioned in the manuscript. Yes, we resampled the 3 GLEAM products to 0.5 degree. We've included this information in the manuscript:

The 3 GLEAM products were resampled to 0.5° using bilinear interpolation to match the spatial resolution of DOLCE

3. **My understanding is that ET product at either 0.5, or 1 deg., most of the flux site grid is in-homogeneity, thinking about the land surface covers can varies a lot in 50km*50km resolution. The more higher resolution of ET, the more higher chance could the site measurement represent information for a gird. Please check the reference: (Anderson et al., 2012).**

The reviewer is of course correct, there is a huge mismatch between the fetch of the flux tower (i.e. 1km^2 or less) and the area of the grid cell containing the site, and a grid cell with an area close to the flux site fetch is more likely to be homogeneous than a 50km*50km grid cell. Unfortunately, as we responded to point 1 above, creating DOLCE at a finer resolution was not possible. We also dedicated a considerable portion of the manuscript to exploring the representativeness question (as noted by Reviewer 1), by testing whether the weighted DOLCE product performed well at sites that were not used to derive the weights (the results shown in Figures 3 and 4, and in particular the two columns of plots in Figure 4). We have also added text to make it clear that DOLCE will *not* perform better at every site, but only perform better across a wide range of sites considered collectively:

In both cases it is important to note that many individual sites agree poorly with DOLCE compared to some other products. The distinction between the results shown in Fig. 3 versus Fig. 4 serve 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.

4. DOLCE in this work does not make an extension for spatial resolution, or temporal resolution, even time coverage (10 years, 2000-2009). It is not encouraged for publication, which is similar to already published by Mueller et al. 2015. However, the method used to merge ET products is very useful. But the way of using it at 0.5 deg is not an optimal one.

As discussed in point 1 above, with currently available ET products, it is not possible to produce DOLCE at a higher resolution, however it is always possible to expand the time coverage of DOLCE as its component products evolve, indeed this is the aim for future versions of DOLCE. We added a statement to this effect in the conclusion.

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.

5. I would say whether the site is reported homogeneous in the evaluations is related to the scale. To asses site homogeneity by ET producer is not fare for the operation of these flux sites. I understand site PIs would seek to ensure the site represent one typical land surface. All the flux tower can be taken as homogeneous sites when the evaluated ET is at 10 meter resolution. Move to 0.5 deg grid, all the flux is located in an in-homogeneous grid. Thus you cannot say 'homogeneous sites', but 'homogeneous site grid'. Or most likely homogeneous due to good matching between 0.5 deg grid and flux site.

This is a good point that we do need to clarify, here we use HOM and HET to describe whether the land cover attributed to the grid box (based on the IGBP map at 0.5 grid) matches the land cover at the flux tower. We clarified this in the manuscript and replaced HOM site and HET site with "HOM-case" and "HET-case" respectively.

6. **This also intrigue my interests, if you use the method to combine and calibrate existing ET to be a ET at 1km or higher resolution, which makes all available flux sites homogenous at this scale and can be used to calibrate weighting ET, more land covers can also be used not only DOLCE Tier 1, 2 and 3, say as IGBP 18 land covers. Why don‘t you use IGBP land covers to replace tier 1, 2 and 3 to derive weight for the 18 land covers? More classification or tiers can also help you produce DOLCE at higher resolution. You may say weighting ET cannot provide information at lower resolution.**

   We thank the reviewer for their suggestion. The difference between Tier1, Tier2 and Tier3 is the size of the ensemble of ET products: Tier1 is derived from 6 ET products, Tier2 from 5 and Tier3 from 2 products. We produced Tier2 and Tier3 which have less products than Tier1 to overcome the limitation of the spatial coverage in the excluded products and ensure a global coverage of DOLCE (i.e. composite of Tier1, Tier2 and Tier3), as explained in Section 2.1.

   The idea of producing a weight for each land cover is a great idea, but as explained in the manuscript we didn’t have flux tower covering all biome types, and for some we had very few sites, not enough to calibrate the weights without over-fitting becoming an issue. We did try to group similar biome types so that each group maintained enough members to allow the in- and out-of-sample testing approach – this was outlined in the discussion with results presented in supplementary material. The results showed that clustering by biome type doesn’t improve the weighting (figure S2).

7. **But the weighting ET can be calibrated with flux site at higher spatial resolution. Each weighting ET can be resampled and calibrated to e.g. 1km resolution. Then you can further combine them into a DOLCE high resolution ET.**

   This point is clearly very similar to the first point raised above. Calibrating low resolution products with site observations to produce high resolution product might seem to be a great idea, however the vast majority of the products weighted here do not have information at fine scales, so while there would be data at high resolution, there would be no more information than the low-resolution product. This might lead a user of the product to believe it provided more information than it actually does.

8. **This work will make the DOLCE more useful. The current 0.5 deg. ET does not differentiate from other fused/merged ET at coarse resolution, e.g. Landflux.**

   We thank the reviewer for their suggestion. The out-of-sample tests show that DOLCE @0.5degree) is performing better than Landflux (@1degree) in all the diagnostics we xamined.

9. **In addition, I also think about if your collection of flux tower data is not enough, which may lead to an biased weight for DOLCE calculation. One reviewer also pointed out the limitation of flux tower for the tropical region. Especially, the flux tower play an important role in your method.**

   We have addressed this point in detail in our response to point 1 raised by Reviewer 1.

10. **You mentioned that ‘irrigated sites’ was excluded. What‘s the purpose of doing this?**

    We added the text below in the manuscript to clarify this point:

    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.

11. **is it believed that all the weighting ET product cannot estimate ET for irrigated crops? If yes, this means DOLCE can also has a big errors for irrigated regions or human influenced regions. Then this need pointed out or at lease add discussion about shortage of this dataset. If no, I would be interested to look at the performance of weighting ET products and**

DOLCE at irrigated flux sites, since irrigated crops may also influence global water balance. We cannot blind to this issues when producing a global ET. Can‘t we? I agree if you remove irrigated sites or HT sites, this will makes your ET or paper looks better, but in reality it also expose the shortage of the method and products.

Testing how DOLCE performs at irrigated sites is a good idea. We expanded our analysis to evaluate how DOLCE performs at the three excluded irrigation sites. We displayed the results in Table 2 and added the text below in the Result section:

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 also discussed the results by including this text in the Discussion section:

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 9 months were available for this site, which coincide with the flood and drain period between spring and fall. DOLCE couldn’t depict the flooding and draining event, probably because none of the weighting products can represent such phenomena, so it is expected that representing seasonal flooding is a shortage in DOLCE.

More specific comments list below.

12. There are a lot of errors in the Table 1. This has been pointed out by Carlos Jimenez.

We thank the reviewer for spotting this, we have corrected these.

13. In addition, I also found some ET not included. This is also a 0.05 deg. monthly ET. You may find here: http://en.tpedatabase.cn/portal/MetaDataInfo.jsp?MetaDataId=249454

A great suggestion. Yes, SEBS is a remote sensing product and looks a good addition to the weighting ensemble. In the current version of DOLCE, we considered global ET products that cover the whole period 2000-2009, whereas SEBS misses a few months in 2000. In the future versions of DOLCE, this will not be an issue and we will consider adding SEBS to the weighting ensemble.

14. Table3. Weighting ET products have negative or positive mean bias, if you give 0.5 weight to MPIBGC (3.837 mean bias) and GLEAM-v2B (-3.571) respectively, then DOLCE tier 1 will have a mean bias of 0.266 W/mˆ2. I also calculate mean bias with the weight and mean bias in table 3 for tier 1, then I get a mean bias of 0.9021 (0.041*3.756+0.495*3.837+(-0.026*6.180+....=0.9021), why do you think the weight provided in table 3 is better than my suggested weight of the two ETs?

The bias represents the mean error of each product with respect to the observation. In the weighting technique, we bias correct the product first, then apply the weighting. This was perhaps not clear enough in the manuscript, so we've changed the table caption to read:

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.

So, to answer this question (a) we are not optimising for mean bias, and (b) we are testing out of sample to make sure the weighting approach works well.

15. Page 11 line28-30, please see my comments above

Thanks, this point has been discussed above.

16. Page 5, at what time resolution did you do the energy balance correction in the two equations or methods? Monthly or half-hourly? I would not expect 'no qualitative differ-ences'

between bowen and residual term correction at half-hourly flux data. Secondly, most of the flux sites have no ground heat flux but soil heat flux at 5cm.

We added the text below in the manuscript to clarify this point:

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 (HFqcOK) for sensible heat flux, Rn_f (Rn_fqcOK) for surface net radiation and LE_f (LE_fqcOK) for latent heat flux.

17. What‘s your consideration of G used in the correction.

We applied quality control and filtering for G as highlighted in steps (2) and (3), section 2.2. We added the text below to clarify this point:

Applying a correction technique for energy imbalance at LaThuile 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.

18. Fig.4 LFD or LDF*,LFA or LFA*?

The reviewer is right, we now changed LDF label to LFD and we modified the caption to clarify why we added * to the four products. The modified caption of Fig.4 reads:

*Figure 4: In (a), (b) and (c), as for Fig. 3 but showing the one site out-of-sample tests. 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 (d), (e) and (f), the one out-of-sample test is trained by HOM-case sites data only.*

19. Figure 7, there is no values for the Sahel desert. This is due to non-values from DOLCE or LandFlux? Then I found DOLCE has ET estimate for Sahel desert in Fig. 8. Please explain it.

We created a mask from the spatial intersection of DOLCE, MPI and Landflux so that when we perform the comparison of DOLCE with MPI (fig. 6) and LandFlux (fig. 7) we look at the same areas. This eliminated the Sahel desert from the comparison, as it is not covered by MPI.

20. Fig. 8. How did you say reliability of uncertainty is low? DOLCE has uncertainty with monthly temporal resolution, am I right? I have difficulty understanding 'seasonal variability of global mean ET and its associated uncertainty'. It's better say 'spatial distribution of a) global ET and (b) its associated uncertainty (standard deviation)in Winter and Summer,'.

We've added two extra plots that show the seasonal variability of 1) ET estimates and 2) uncertainty estimates, we changed the plot titles and rewrote the caption to read:

Figure 8: Seasonal (a) global mean ET and (b) its variability (standard deviation), (c) time average of uncertainty (the standard deviation uncertainty shown in Equation 7) (d) standard deviation of uncertainty over time (e) reliability, defined as high ($\frac{\boldsymbol{Uncertainty\ SD}}{\boldsymbol{mean\ ET}} \leq \mathbf{1}$ *in blue), medium(*$|\boldsymbol{mean\ ET}| \leq \mathbf{5}$, $\boldsymbol{Uncertainty\ SD} < \mathbf{10}$ *and* $\frac{\boldsymbol{Uncertainty\ SD}}{\boldsymbol{mean\ ET}} \geq \mathbf{1}$ *in green) and low (in red). DJF is shown in the left column and JJA in the right column.*

21. Line 10. 'Together with the reasonable density .. are reasonably well constrained.' need rephrase.

We have rewritten this to read:

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.

22. Page 1, Line 16, ‘point-based estimates of flux towers provide information at the grid scale of these products.’, are you sure point flux tower can provide information at 50*50 km pixel? Please check my comments above.

As noted above in our response to point 3, we also think this is an interesting question that we have gone to some lengths to address – please see our response above for more detail. The fact that overall, the weighting is succeeding out-of-sample shows that the point-based measurements used to weight DOLCE do indeed provide useful information at the 0.5 degree grid cell scale. So yes, within the caveats noted in the revised manuscript, we are reasonably certain that this is the case.

23. Page 1, Line 21, These ET products differ in their data requirements, the approaches used to derive them and their estimates (Wang and Dickinson, 2012). This is well known. No need citation here.

We acknowledge that many in the community do know this, but are reasonably certain that others do not, and so have left the reference as is – there is no cost to this.

24. Line 24, we provide an even stronger vindication of this relationship. Which part show this? Please give some explanation.

In section 2.3 we showed that combining products will always derive a better performing product at the in-sample sites. We now refer the reader to Section 2.3.

25. Please use either ‘Time-space step’ or ‘Space-time step’;

We thank the reviewer for spotting this. We now made the necessary changes and chose to use “Time-space”.

26. Acknowledgements Where did you use ERA-inerim by saying ‘The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE’.

The reviewer is right, we haven't used ERA-Interim reanalysis data, but part of the term and conditions of the use of FLUXNET data is to use the exact statement included in http://fluxnet.fluxdata.org/data/data-policy/. We assume ERA-Interim is included in the statement because it was used for gap filling.

27. **References: Annan,. . . ..n/a-n/ Fisher, JB has been listed two times. Miralles, D.G, also listed two times. Please check the standard format for references used on HESS website**

Thanks for picking this up, we have removed the duplicates and made the appropriate corrections.

28. **Table s1, please also add RMSE, correlation and mean bias values for each site. This information is also important for DOLCE dataset users.**

We obviously have all these values and they are all included in the plots (Figure 5 and Figure 9). The table is already dense but we are very happy to include them if the Editor thinks that fitting these into the table will not make the table unpublishable.

Tables

Table1: Gridded ET products used in this paper.

ET product and Reference	*Abbreviation*	*Time period & Spatial Resolution*	*Forcing data source*	*Calculation Method(s)*
CSIRO-global (Zhang et al., 2010a)	*CS*	*1983–2006 0.5° Also available at 8km and 1°*	*Meteorological observations from flux tower distributed across all global biome types Remote sensing inputs*	*An extended ET product of CSIRO (Zhang et al., 2010b) that covers a global domain NDVI-based PM model PT equation for open water evaporation*
GLEAM-V2A (Miralles et al., 2011)	*G2A*	*1980–2011 0.25°*	*Remote sensing based observations Gauged based precipitation*	*PT equation Canopy Interception Model, Soil water module and Stress module*

GLEAM-V2B (Miralles et al., 2011)	*G2B*	*2000–2011 0.25°*	*Remote sensing based observations*	*PT equation Canopy Interception Module, Soil water module and Stress module*
GLEAM-V3A (Martens et al., 2016)	*G3A*	*1980–2014 0.25°*	*Satellite based inputs Multi-source precipitation*	*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*
LandFlux-Eval-Diag (Mueller et al., 2011, 2013)	*LFD*	*1989–2005 1°*	*Simple mean of 5 diagnostic ET datasets*	
LandFlux-Eval-All (Mueller et al., 2011, 2013)	*LFA*		*Simple mean of 14 Diagnostic, LSM and Reanalysis datasets.*	
MOD16 MODIS global ET products (Mu et al., 2011)	*MOD*	*2000–2014 0.5° also available at 0.0 5°*	*Global Modeling and Assimilation Office (GMAO) meteorological reanalysis data Remote sensing inputs from MODIS 8-day retrievals*	*PM formula (Monteith J. L., 1965)*
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*	*PT equation*

			Remote sensing based observations from monthly AVHRR data	

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.

Site-Code	*Longitude*	*Latitude*	*Description*	*RMSE*	*Mean bias*	*SD difference*	*Correlation*	*Number of months*
US-Ne1	*-96.4766*	*41.1651*	*Rice paddy*	*16.6*	*-7.41*	*-9.25*	*0.96*	*103*
US-Ne2	*-96.4701*	*41.1649*	*Mead irrigated continuous maize site*	*15.8*	*-5.05*	*-7.44*	*0.95*	*103*
US-Twt	*-121.6521*	*38.1055*	*Mead irrigated maize-soybean rotation site*	*91.9*	*-67.39*	*-55.23*	*0.49*	*9*

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

Land Cover	*HOM-case*	*HET-case (site)*	*HET-case (grid cell)*
CRO	*10*	*7*	*20*
CSH	*0*	*1*	*0*
DBF	*1*	*16*	*0*
EBF	*3*	*5*	*0*
ENF	*6*	*22*	*2*
GRA	*13*	*27*	*5*
MF	*7*	*3*	*32*
OSH	*2*	*1*	*4*
SAV	*3*	*1*	*6*
VEG	*1*		*0*
WET		*5*	*0*

WSA		*4*	*9*
Wa (Water)			*1*
URB (Urban)			*1*

Figures

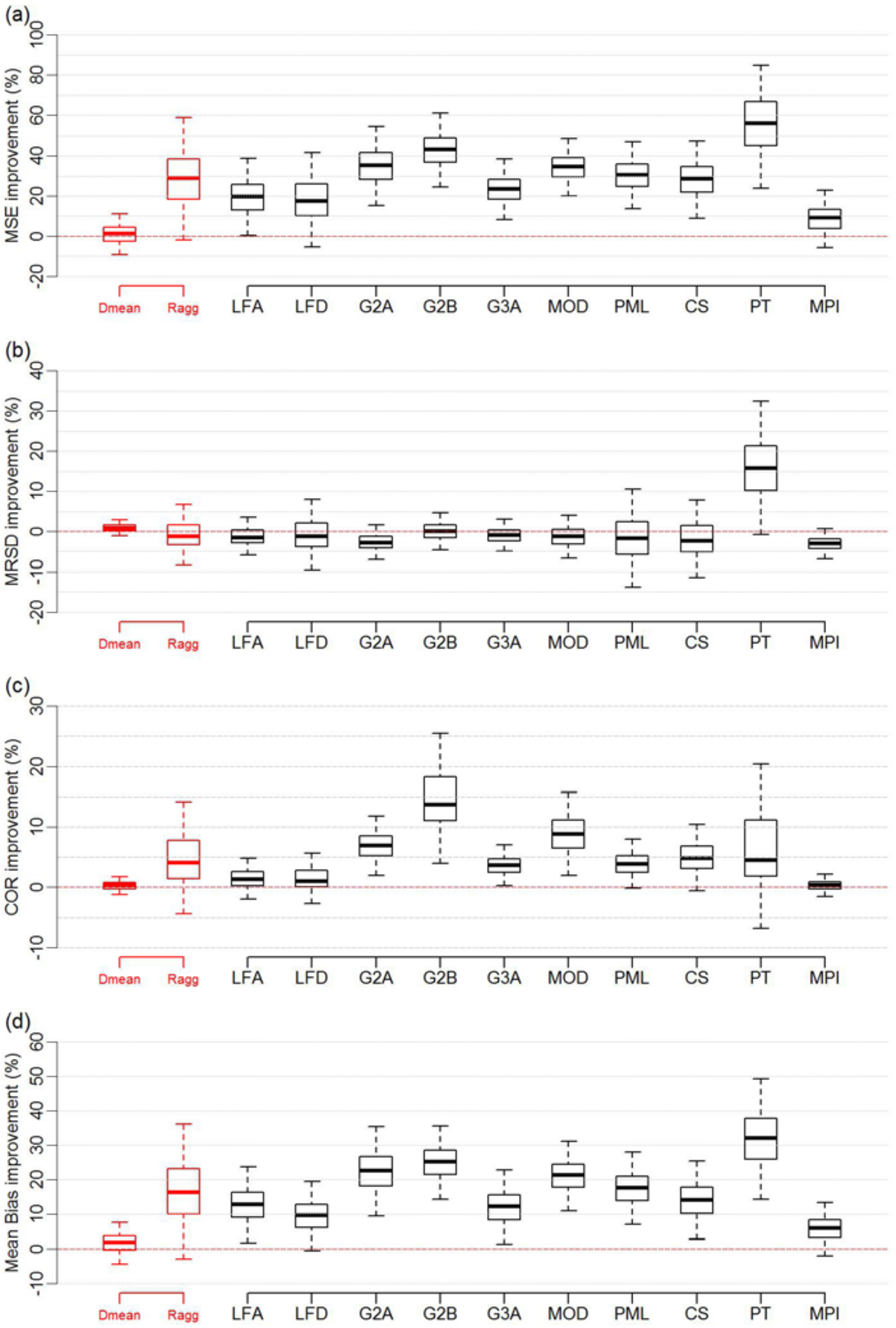

Figure 3: Box and whisker plots displaying the percentage improvement that the weighted product provides in the 25% out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean bias (d), when compared to equally weighted mean of the Diagnostic Ensemble (Dmean), aggregated Reference Ensemble (Ragg) and each member of the reference ensemble. Box and whisker plots represents 5000 entries, each entry is generated through randomly selecting 25% of sites to be out sample.

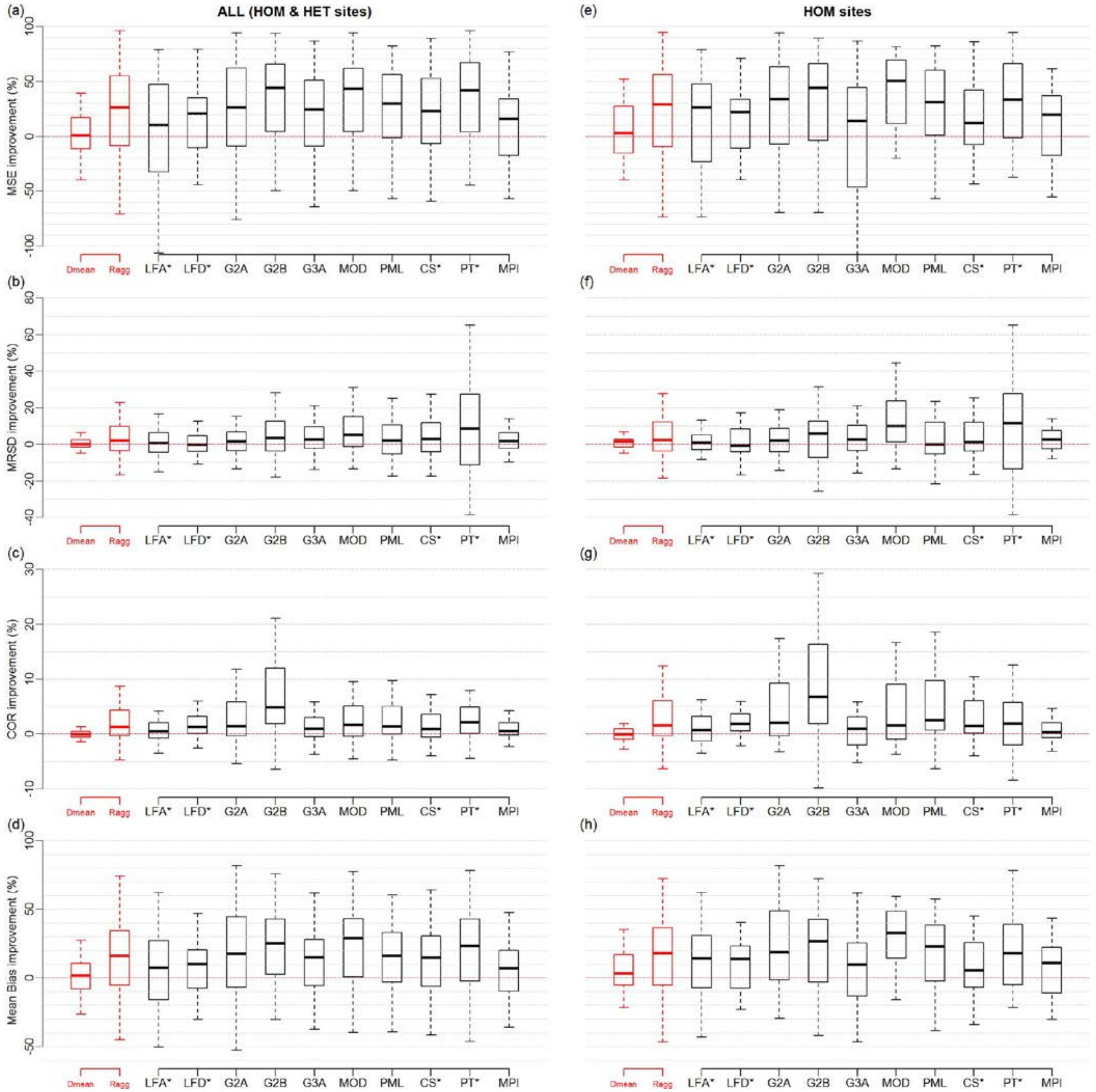

Figure 4: In (a), (b), (c) and (d), as for Fig. 3 but showing the one site out-of-sample tests. Box and whisker plots are generated through selecting one site to be out sample and are repeated for all 138 sites. Products marked with * have limited spatiotemporal availability relative to the diagnostic 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 out-of-sample test is trained by HOM-case sites data only.

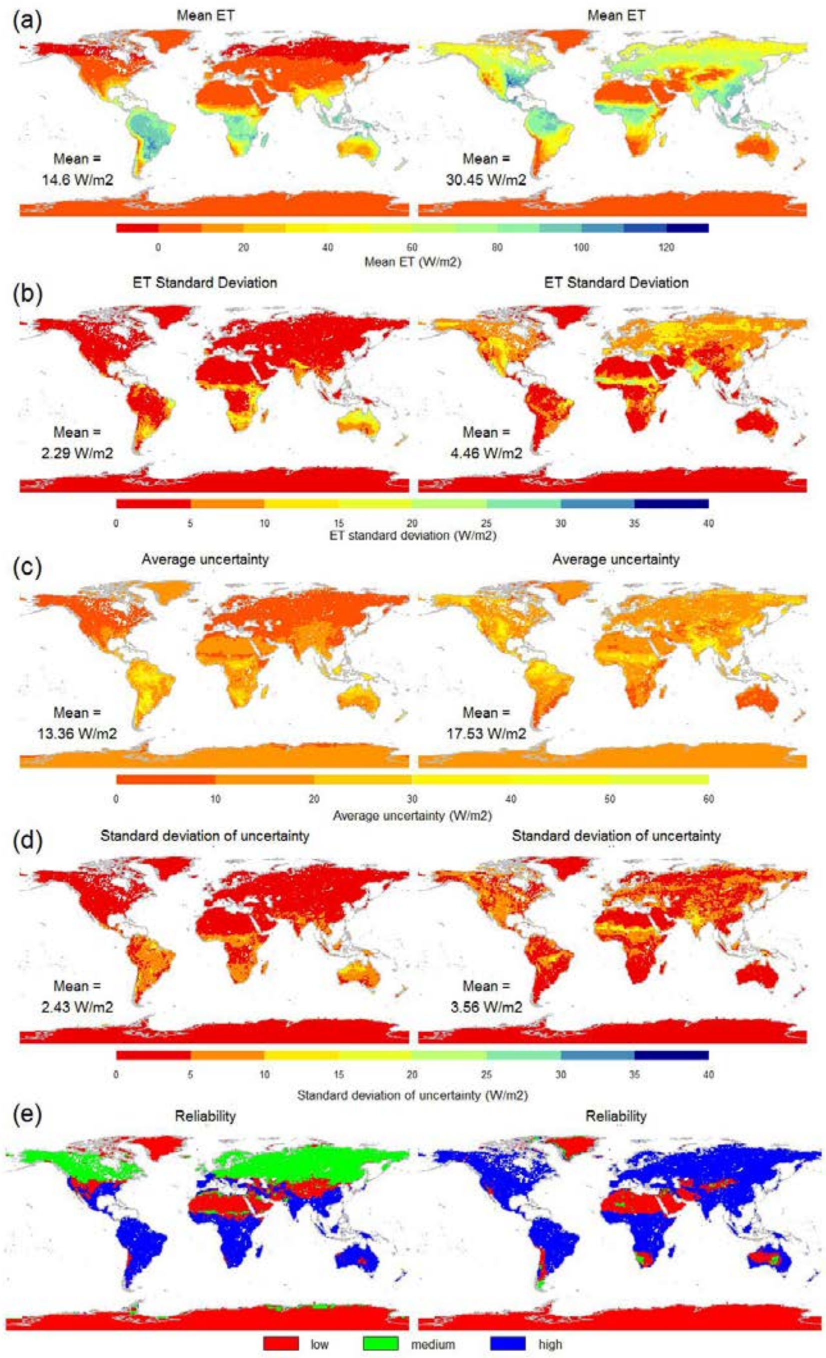

Figure 8: Seasonal (a) global mean ET and (b) its variability (standard deviation), (c) time average of uncertainty (the standard deviation uncertainty shown in Equation 7) (d) standard deviation of uncertainty over time (e) reliability, defined as high ($\frac{Uncertainty\ SD}{mean\ ET} \leq 1$ in blue), medium($|mean\ ET| \leq 5$, $Uncertainty\ SD < 10$ and $\frac{Uncertainty\ SD}{mean\ ET} \geq 1$ in green) and low (in red). DJF is shown in the left column and JJA in the right column.

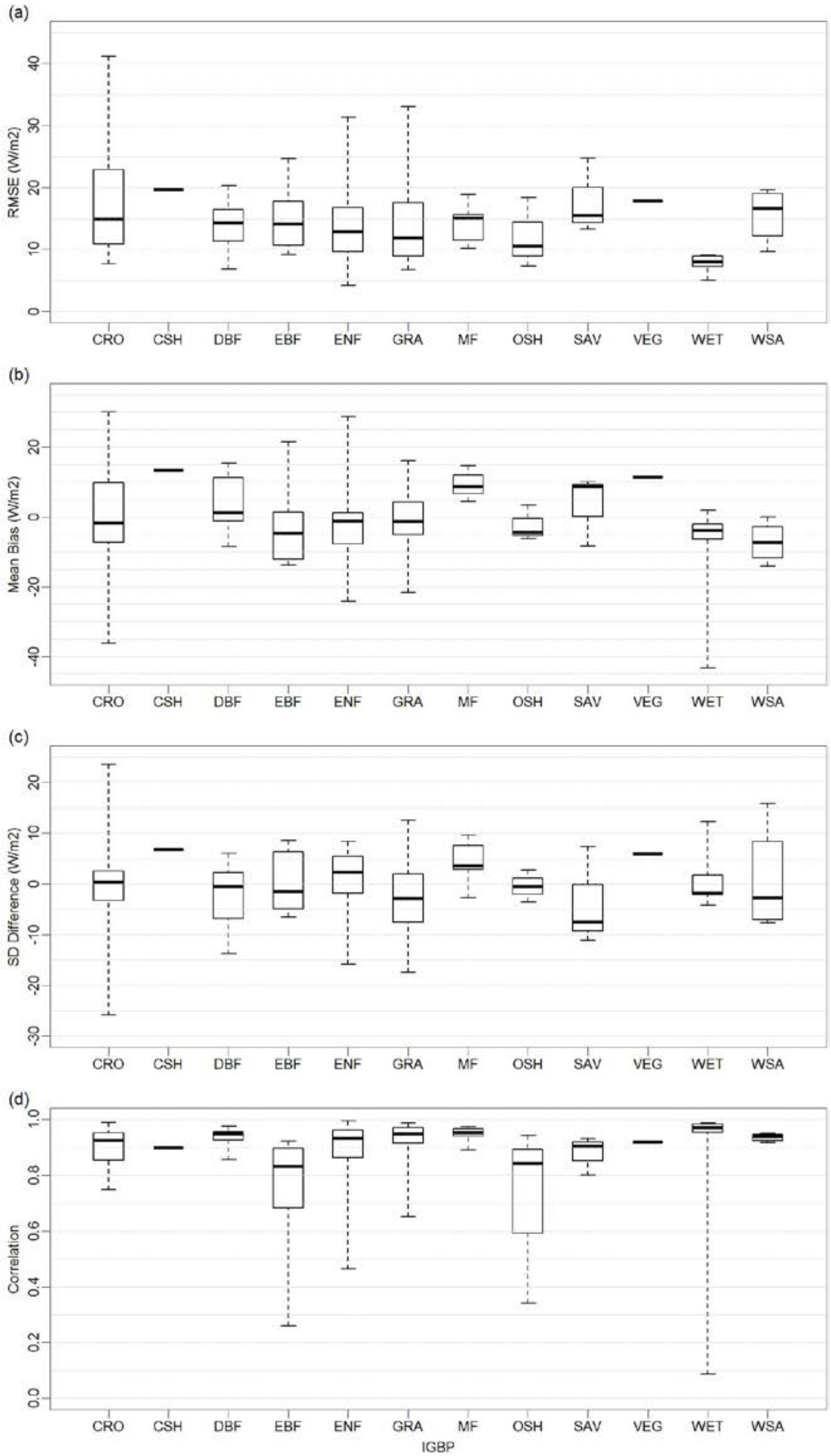

***Figure 9: Four statistics, (a) RMSE, (b) Mean bias, (c) SD difference and (d) Correlation, calculated for DOLCE at 142 flux tower sites and displayed by biome types. See Fig. 2 for biome abbreviations*.**

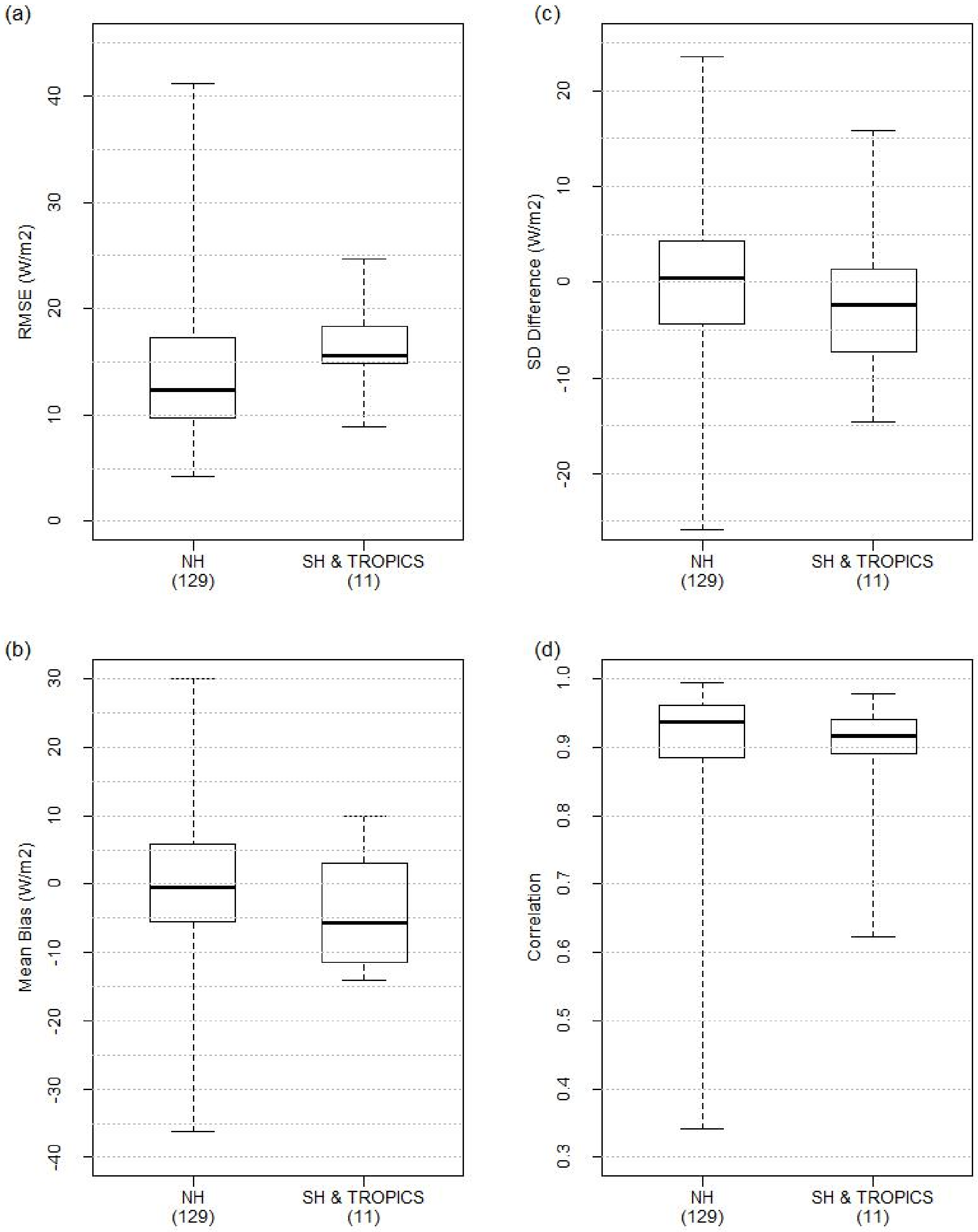

Figure 10: Four statistics, (a) RMSE, (b) Mean bias, (c) SD difference and (d) Correlation, calculated for DOLCE separately at 129 flux towers located at the Northern Hemisphere excluding the tropics (NH) and at 11 towers in the Southern Hemisphere and the tropics (SH & TROPICS)

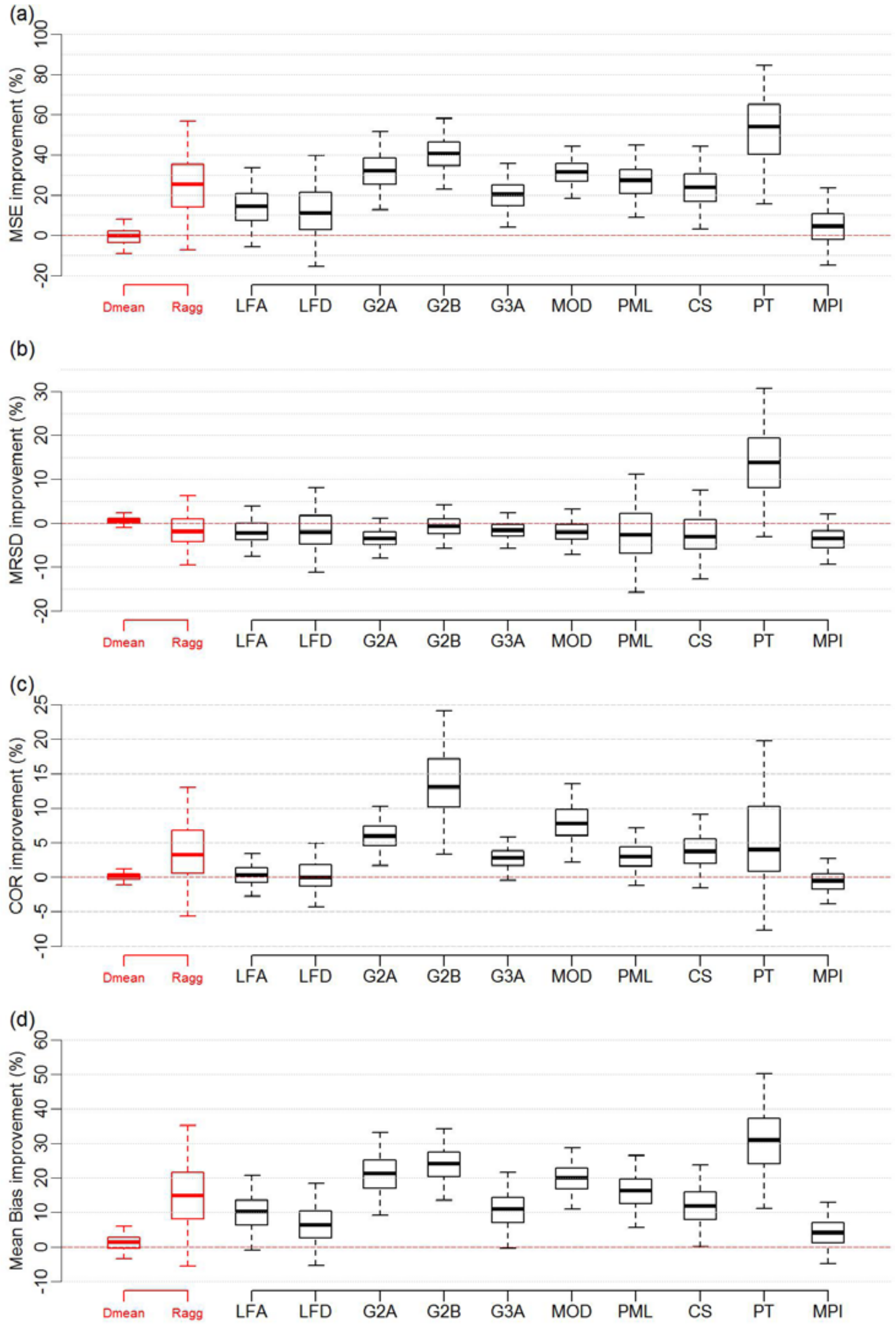

Figure 11: Box and whisker plots displaying the percentage improvement that the weighted product excluding MPIBGC provides in the 25% out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias (d), when compared to equally weighted mean (Dmean) of the Diagnostic Ensemble, aggregated Reference Ensemble (Ragg) and each member of the reference ensemble. Box and whisker plots represents 5000 entries, each entry is generated through randomly selecting 25% of sites to be out sample

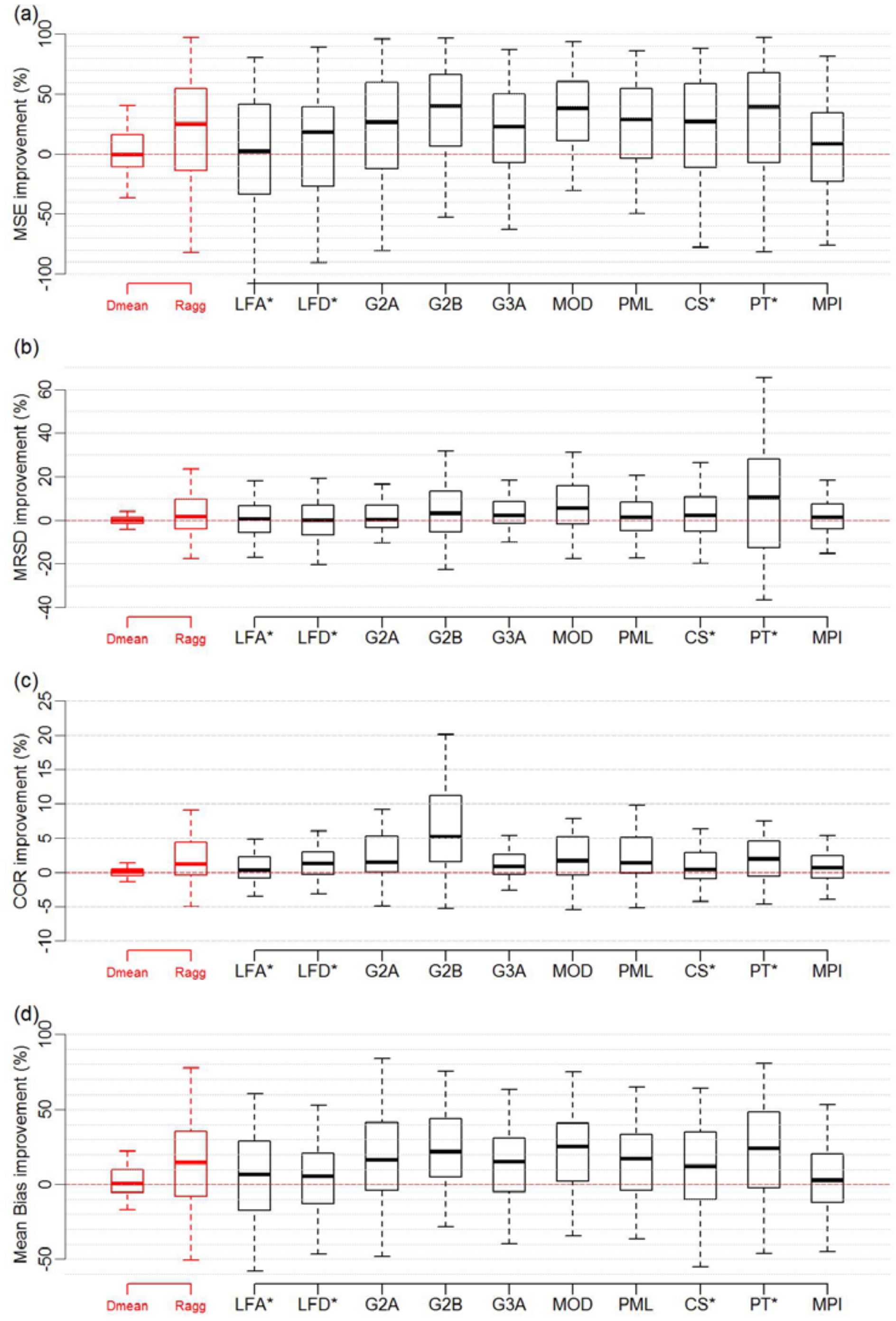

Figure 12: Box and whisker plots displaying the percentage improvement that the weighted product excluding MPIBGC provides in the one out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias (d), when compared to equally weighted mean (Dmean) of the Diagnostic Ensemble, aggregated 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.

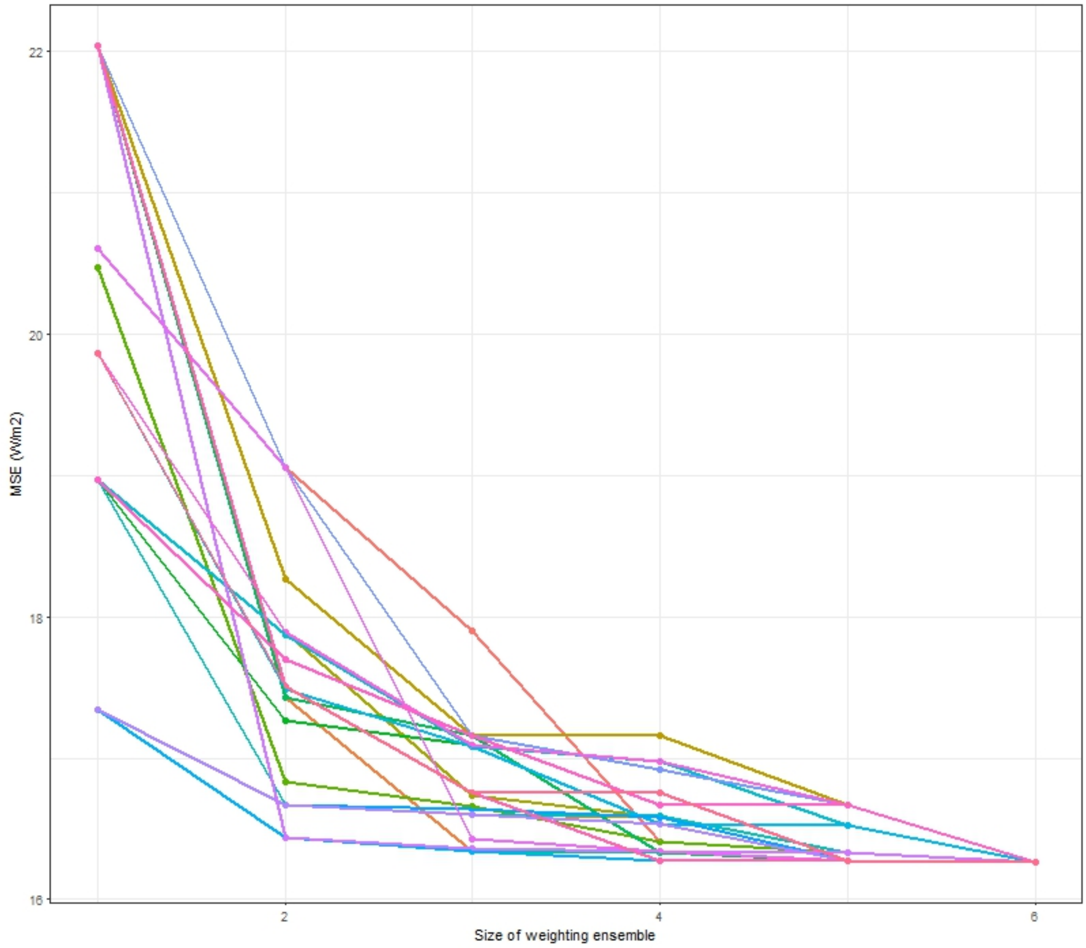

Figure S1: The results of the in-sample test showing how the Mean Square Error (MSE) of the weighting changes when we increase the number of ET products involved in the weighting from 1 to 6. The test was repeated 25 times of a random selection of n products (2 ≤ n ≤ 6).

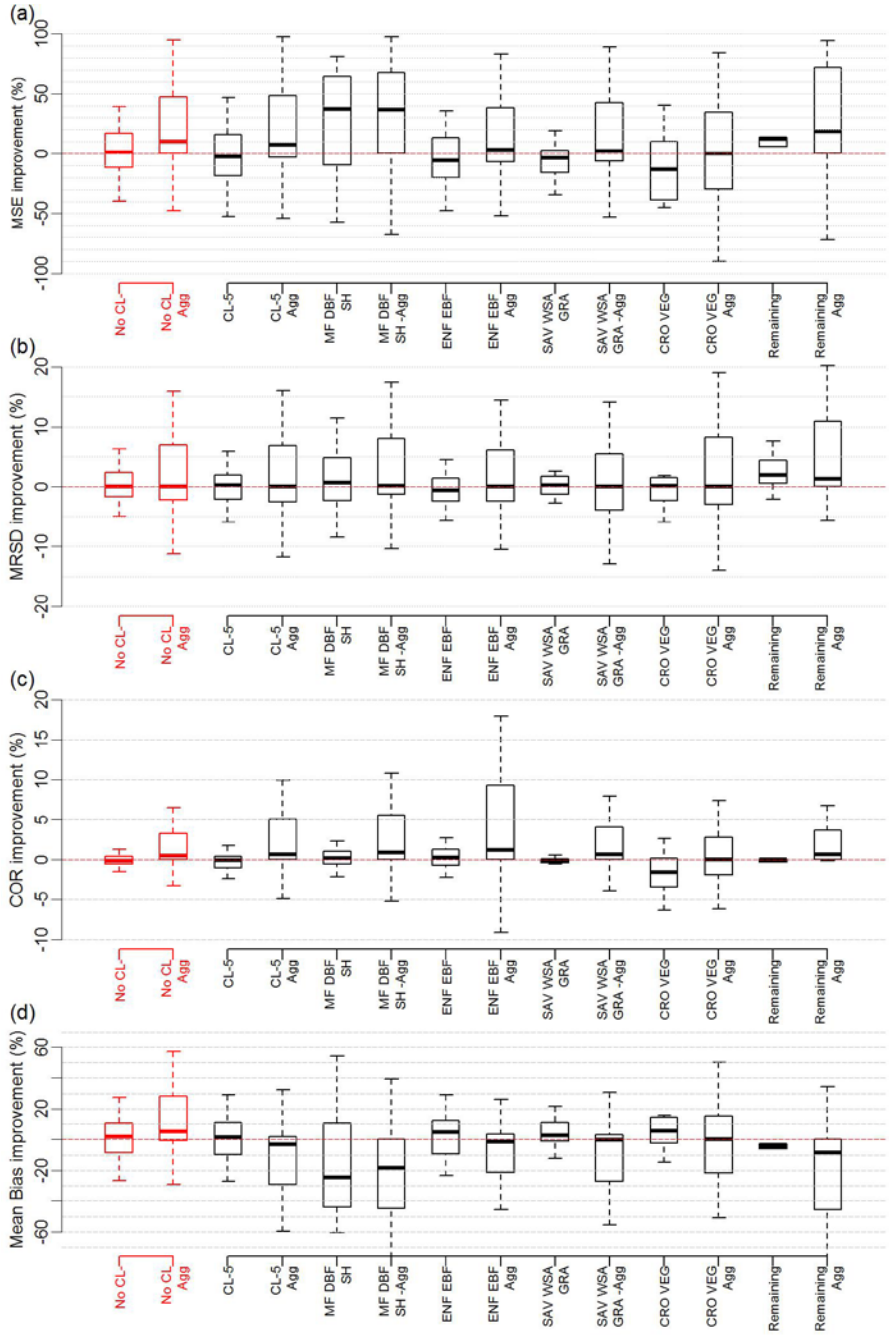

Figure S2: Box and whisker plots displaying the results of the One out-of-sample site test for the cluster independent weighting (No CL and No CL Agg boxplots), the cluster dependent weighting (Cl-5 and CL-5 Agg) and over individual biome types for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias (d).