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 #1 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 #1

General comments

1. **There is obviously a question regarding the representativeness of the FLUXNET stations to their regions (grid box average or dominant surface conditions), which the paper addresses fairly well. However, one specific issue jumps out at me: what about rare representatives, particularly in the tropics? FLUXNET is woefully thin on stations at low latitudes. The out-of-sample tests are like an OSSE, but there is little sampling in the tropics to play with here. What was the specific effect of denying low-latitude stations? Can we use this study to make case to prioritize FLUXNET expansion into the tropics, S/SW Asia and the Southern Hemisphere? I would like to see this issue discussed more, where the global maps are presented and also in the Discussion. Motivating targeted FLUXNET network expansion would be a good "broader impact" of this study.**

This is a great point. We have extended the discussion to try to address this:

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 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.

2. **What about the uncertainty/error in the FLUXNET observations? They certainly contain random errors and systematic biases. For a measured quantity that has a red-noise spectrum (following a Markov process) the random error can be estimated from the**

behavior of lagged autocorrelations (cf. http://dx.doi.org/10.1175/JHM-D-15-0196.1). Random error will systematically degrade correlation and affect other statistics (cf. http://dx.doi.org/10.1175/JHM-D-15-0063.1). Systematic error is more difficult to identify, but the energy balance corrections in FLUXNET2015 give some clue (see specific comments below). All these mean that using FLUXNET for cal/val is itself flawed and imperfect. On the other hand, model ET is inherently precise (no random error) and can be used to estimate/differentiate this type of error from others by noting its statistical differences from the instrument records. This issue should be acknowledged and discussed, including how the assumptions in DOLCE regarding errors (that they are uncorrelated) affect results. In other words, more discussion about uncertainties.

Yes. We have extended the discussion to address this point, at least to some degree:

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 Figure 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.*

3. **In the Supplement: There are captions for Fig S1 and Fig S2, but the PDF does not contain any figures!! Please recheck the rendering.**

   We thank the reviewer for spotting this, we now added the missing Figures in the Supplement.

Specific comments:

4. **P3 L6: Change "85 FLUXNET tower data" to "85 FLUXNET towers"**

   We've corrected this in the manuscript.

5. **P3 L8: "ground-truthed" is not an appropriate verbification for this context. Change '...gridded ET data sets are "ground-truthed" using flux tower data from FLUXNET...' to '...tower data from FLUXNET provide ground truth for gridded ET data sets...'**

   Thanks for picking this up, we've made the change.

6. **P5 L16: Change "where" to "were"**

   Thanks for picking this up, we've corrected this in the manuscript.

7. **P7 L22: RSD cannot convey whether the variance is too large or too small. So for the changes shown in Fig 3b - we cannot tell whether the improvement is due to an increase or decrease in standard deviation. Furthermore, what about locations where mean ET (denominator) is near zero - does that explain some of the very large values and changes?**

   The reviewer is right; as a result of this comment and a comment by Carlos Jimenez, we've replaced the relative standard deviation metric (RSD) with a modified relative standard deviation MRSD defined as $\frac{\sigma_{\text{dataset or observation}}}{max(mean(observation),\ q)}$. This addresses the two issues and removes the potential for improvement in the RSD metric simply because the mean has improved. We have also added a fourth panel to this Figure showing improvement in the mean, for reference. We added the text below to explain the new metric:

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 2nd percentile of the distribution of observed mean flux (i.e. temporal mean ET) across all sites (about 13 W/m2), 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. $|MRSD_{dataset} - MRSD_{observation}|$, and showed the performance improvement of the weighted mean.

8. **Fig 3: So the spread is across 5000, and each of those is an average of 25% out-of sample stations, right? It is not 5000x172x0.25 points. Please make clear.**

   The reviewer is right. We now clarified this in the figure caption:

9. **Also, Fig 4 suggests individual o-o-s stations frequently fare worse. Transferability of calibrations appears to be kind of weak, which is not a surprise. Calibration transferability is a very difficult enterprise. This should also be acknowledged, either here or in the Discussion section (wishing PILPS-San Pedro had been completed as it would have really shone a light on this problem).**

   Yes, this is indeed worth mentioning. We have modified the results section to read:

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 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.*

10. **P9 and Fig 4: "widespread out-of-sample improvement that this approach offers over existing gridded ET products" seems like a bit of an overstatement - there are frequently locations that fare worse, and sometimes much worse. In Fig 4, there is typically a tremendous range, and usually the central two quartiles encompass the zero line. For RMSE it appears that _20-49% of the time the estimates are worse than other products and methods, 30-55% for RSD, and 20-55% for COR. While there is definitely a net (and welcome) improvement in almost all cases, often the DOLCE estimate is worse. This should be acknowledged.**

    This is indeed a fair criticism. We have modified this to read

    "*On the basis of the aggregate out-of-sample improvement that this approach offers over existing gridded ET products*".

11. **P9 L26: "Standard Deviation (SD) difference" - this is clearly not the same as RSD defined earlier - please define, which minus which?**

    The reviewer is right, this is not the same as RSD. We've clarified this in the manuscript:

    Standard Deviation (SD) difference (i.e. $\sigma_{DOLCE} - \sigma_{observation}$)

12. **Table 2 has little accompanying discussion, and what is there is very shallow adding little to comprehension. Please discuss more or remove the Table if it does not warrant discussion.**

    We thank the reviewer for his suggestion, we removed this table from the manuscript.

13. **Fig 6: Here it could be said you use DOLCE to estimate MPI, and most places have a positive bias. But the energy-balance-corrected fluxnet data, which close the surface energy balance by construct conserving Bowen ratio, consistently increases ET compared**

to the raw measurements (at 107 of 122 FLUXNET2015 Tier-1 stations during JJA by my quick calculation). There is recent independent indication that tower sites bias low because of errors in the turbulence theory applied to estimate fluxes (see: http://dx.doi.org/10.1002/2017GL073499). Could FLUXNET instrument error (and the simplicity of the energy closure correction) contribute to systematic biases in DOLCE? Please discuss here or in the Discussion section.

Yes, of course it could. We have added to the discussion to make this clearer:

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.

14. **Re Fig 7: Much of the largest differences are at low latitudes where there is little FLUXNET data for calibration. Please discuss, as I see this as a major issue (more with**

FLUXNET station distribution than your methods, but the problem of representative ET estimates in the tropics is an ongoing concern).

We have addressed this issue in our response to the first point raised above.

15. **P10 L21: "ET doesn't exhibit any seasonal change over Greenland and the deserts in North Africa..." - not expected in the absolute, because mean values are tiny. It would be more informative to also show relative (percentage) changes, which are more relevant for local water balances.**

A good point. We have now added two extra plot in Fig.8 that show the seasonal variability of 1) ET estimates and 2) uncertainty estimates, we changed the plot titles, made the caption clearer:

and commented on the plots in the text:

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.

16. **Discussion §: Please also speculate whether some spatial variability in weighting could improve estimates further, even if only in 2 or 3 categories of weights.**

We agree this is worth exploring. That's why we performed clustered weighting where each cluster had a different set of weights. We accept this might not have been clear enough in the manuscript, we clarified this in the discussion:

In this study, we sought a single weight for each product to apply globally. But we have a 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.

17. **Fig 9: Only stations in the tropics are 2 EBF stations; the savannah stations are in southern Africa. These are not o-o-s results, right? I would have expected these types to stand out more, but perhaps the samples are biased to the extratropical stations in the categories. Thoughts?**

We have added this point to the discussion:

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.

18. **Also Fig 9: Crops are tricky. The category is a catchall that is unsatisfactory because there is such variability in phenology, seasonality, stomatal resistances, etc. Are any of the CRO stations rice (which acts very different because of seasonal flooding). Not surprising many of the biggest errors are there.**

Good point. In the site description provided by FLUXNET the crop type is not always clear. There is only one rice site that we are sure about, but we excluded this site from the weighting because it was found in the list of irrigated sites. We excluded all irrigated sites. We've added the text below to explain the reasons:

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 (Miralles, 2011; Jung, 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 expanded our analysis as suggested by reviewer #2 point 10 to test the performance of DOLCE in three irrigated sites, we've added the text below to show this:

We tested the performance of DOLCE at three irrigated sites that were excluded from the weighting, for reasons explained earlier 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.

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 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.

19. P13 L1: "...for example anthropogenic water management..." - but it was stated earlier that irrigated sites were excluded. Please expand on this comment.

We addressed this concern in the previous point.

20. Conclusion: Please also state plans for updates, future versions, perhaps (hopefully) covering a longer time period!?

We thank the reviewer for his suggestion. We have added future improvement of DOLCE in the conclusion.

21. References: There seem to be a lot of redundancies in author lists for papers with names showing up 2 or 3 times for the same entry. Please check.

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

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	*G3A*	*1980–2014*	*Satellite based inputs*	*A revised version of GLEAM*

(Martens et al., 2016)		0.25°	Multi-source precipitation	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 Remote sensing based observations from monthly AVHRR data	PT equation

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

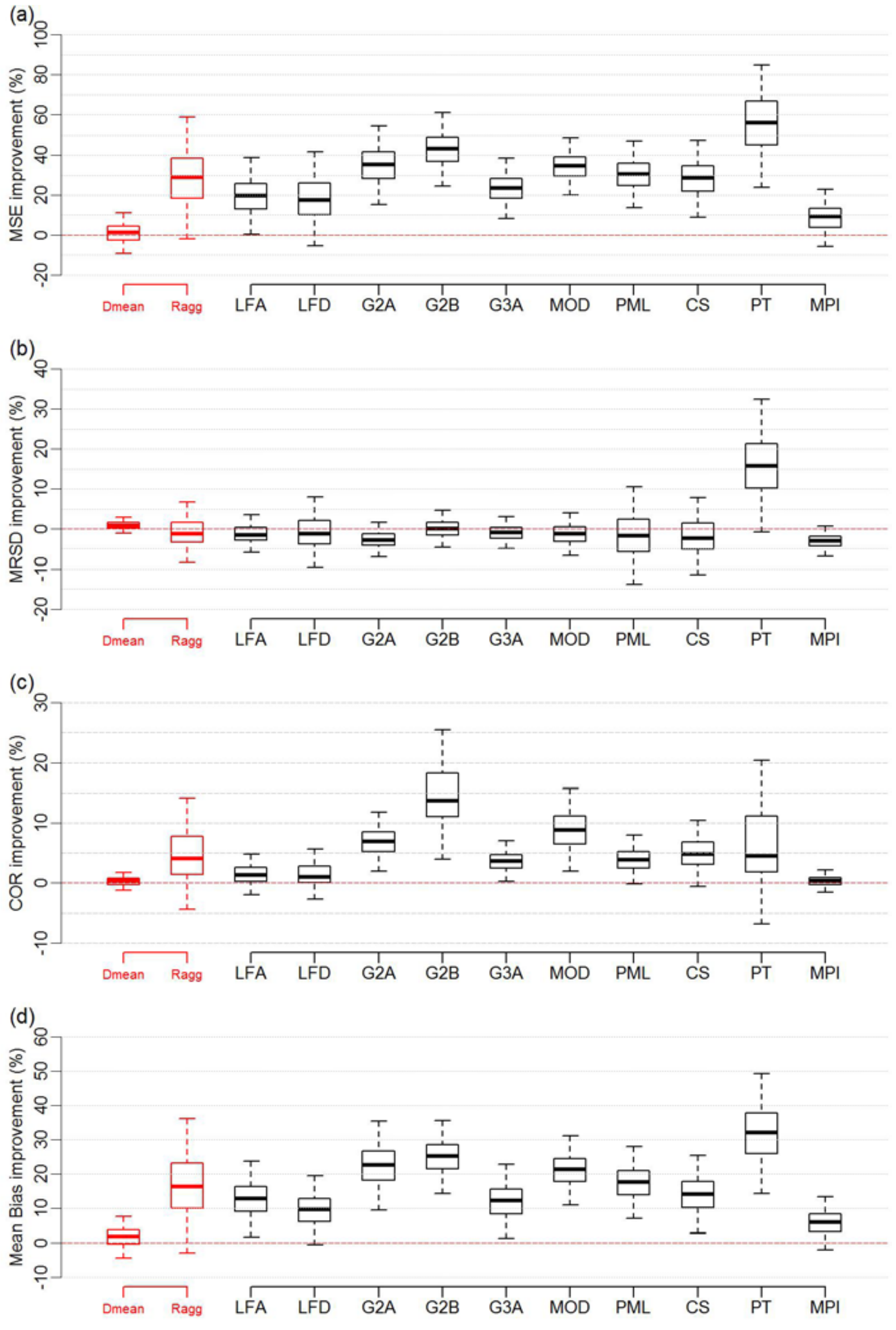

Figure 3: Box and whisker plots displaying the percentage improvement that the weighted product provides in the 25% out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean bias (d), when compared to equally weighted mean of the Diagnostic Ensemble (Dmean), aggregated Reference 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.

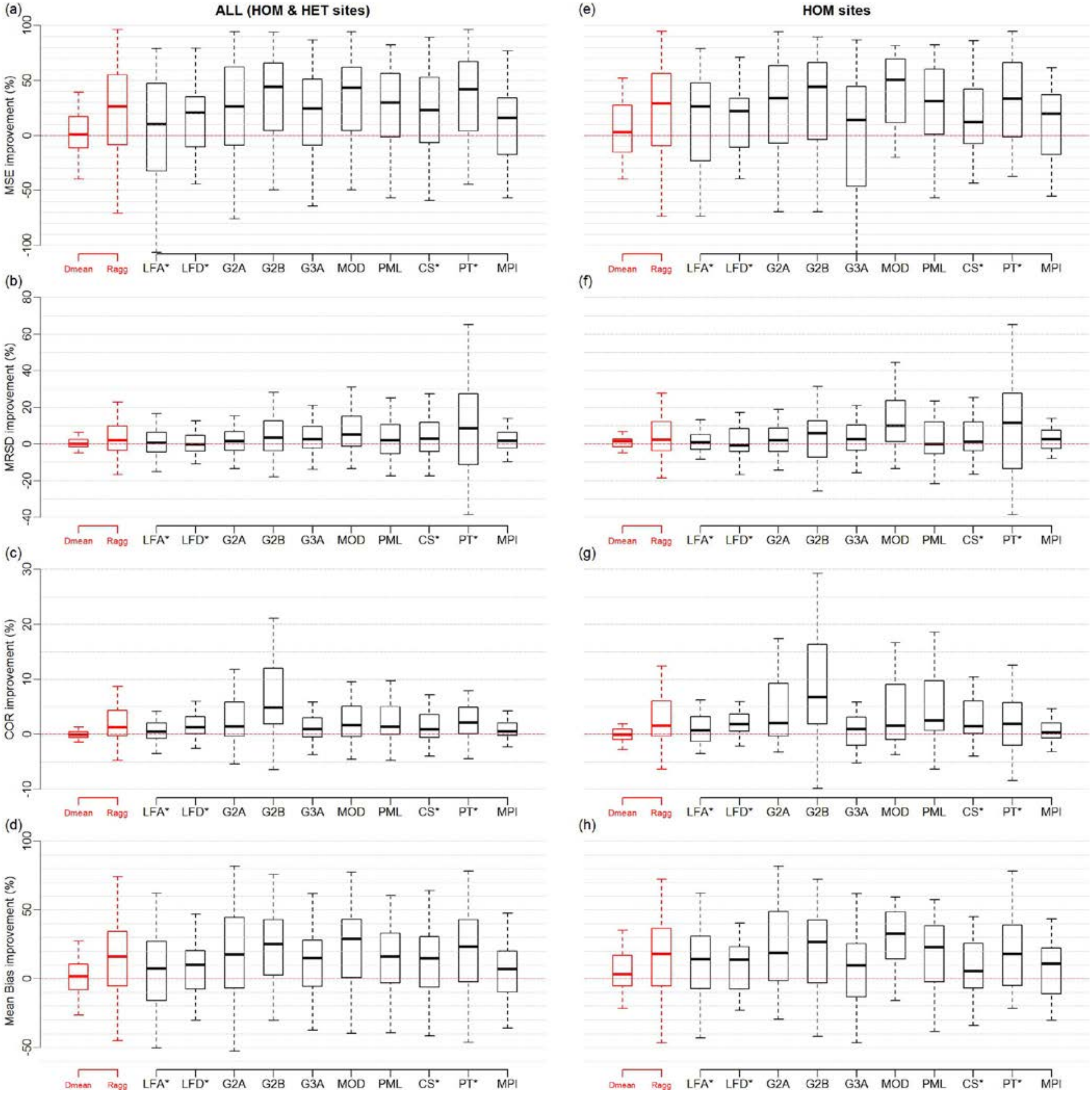

*Figure 4: In (a), (b), (c) and (d), as for Fig. 3 but showing the one site out-of-sample tests. Box and whisker plots are generated through selecting one site to be out sample and are repeated for all 138 sites. Products marked with * have limited spatiotemporal availability relative to the diagnostic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

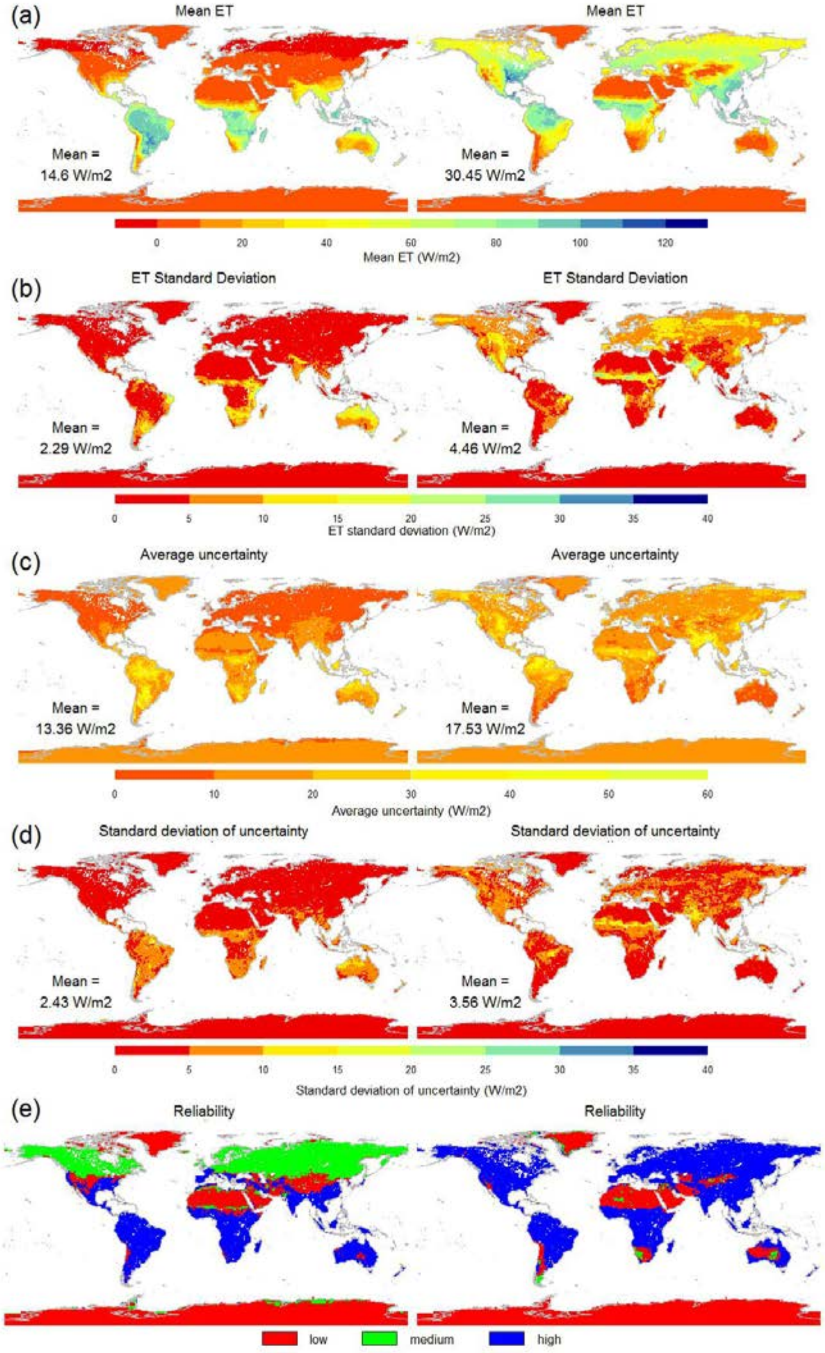

Figure 8: Seasonal (a) global mean ET and (b) its variability (standard deviation), (c) time average of uncertainty (the standard deviation uncertainty shown in Equation 7) (d) standard deviation of uncertainty over time (e) reliability, defined as high ($\frac{Uncertainty\ SD}{mean\ ET} \leq 1$ in blue), 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.

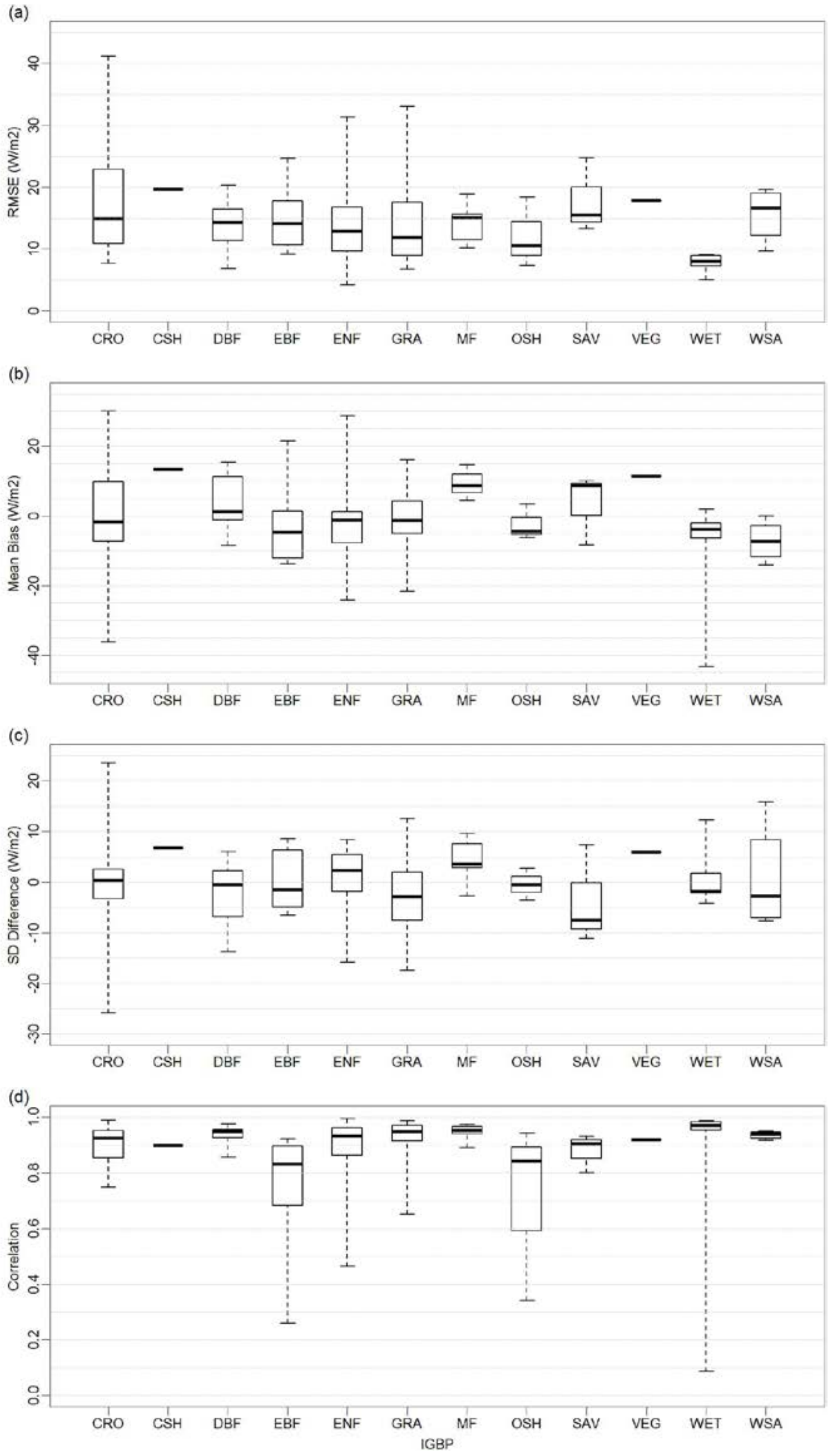

Figure 9: Four statistics, (a) RMSE, (b) Mean bias, (c) SD difference and (d) Correlation, calculated for DOLCE at 142 flux tower sites and displayed by biome types. See Fig. 2 for biome abbreviations**.**

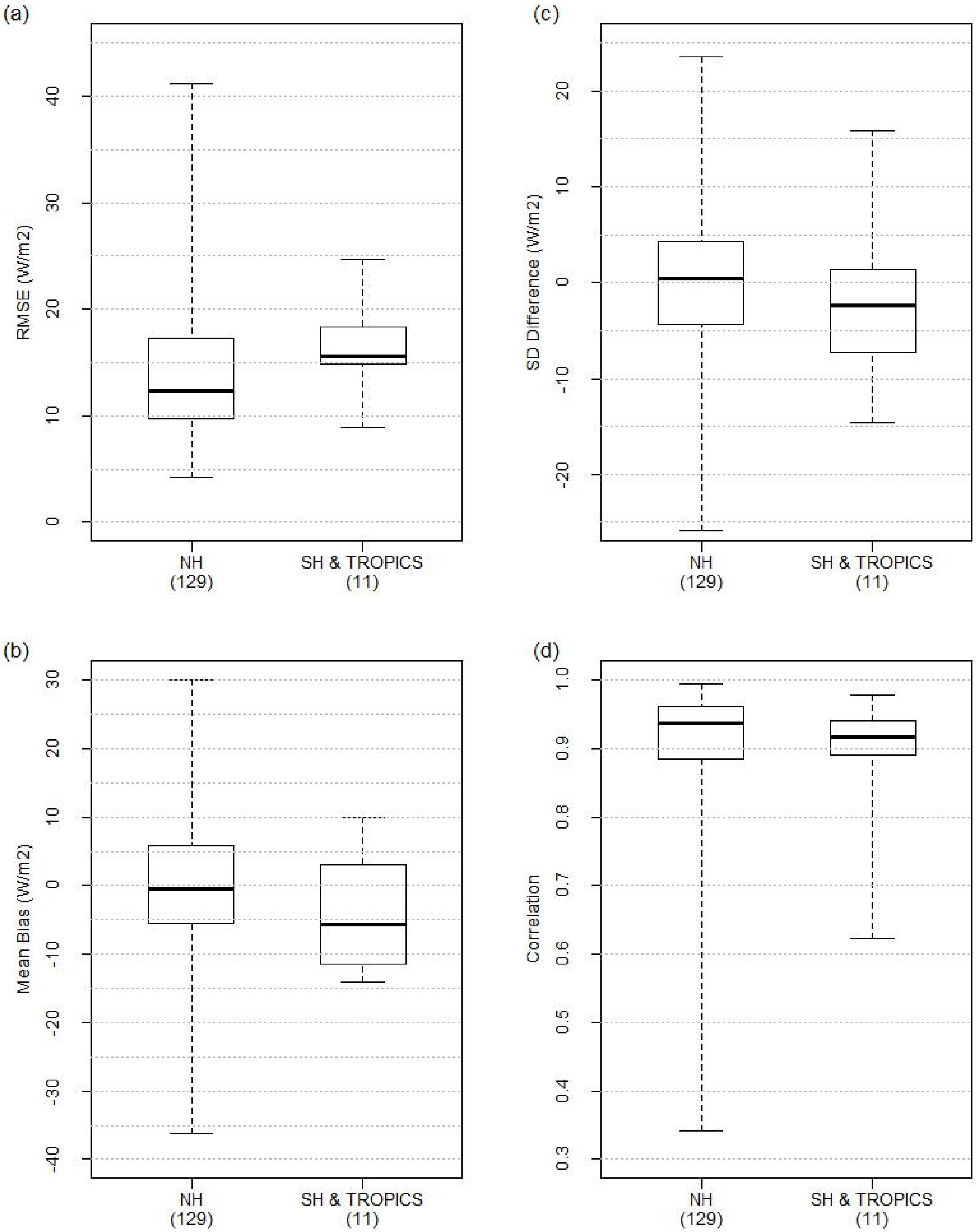

Figure 10: Four statistics, (a) RMSE, (b) Mean bias, (c) SD difference and (d) Correlation, calculated for DOLCE separately at 129 flux towers located at the Northern Hemisphere excluding the tropics (NH) and at 11 towers in the Southern Hemisphere and the tropics (SH & TROPICS)

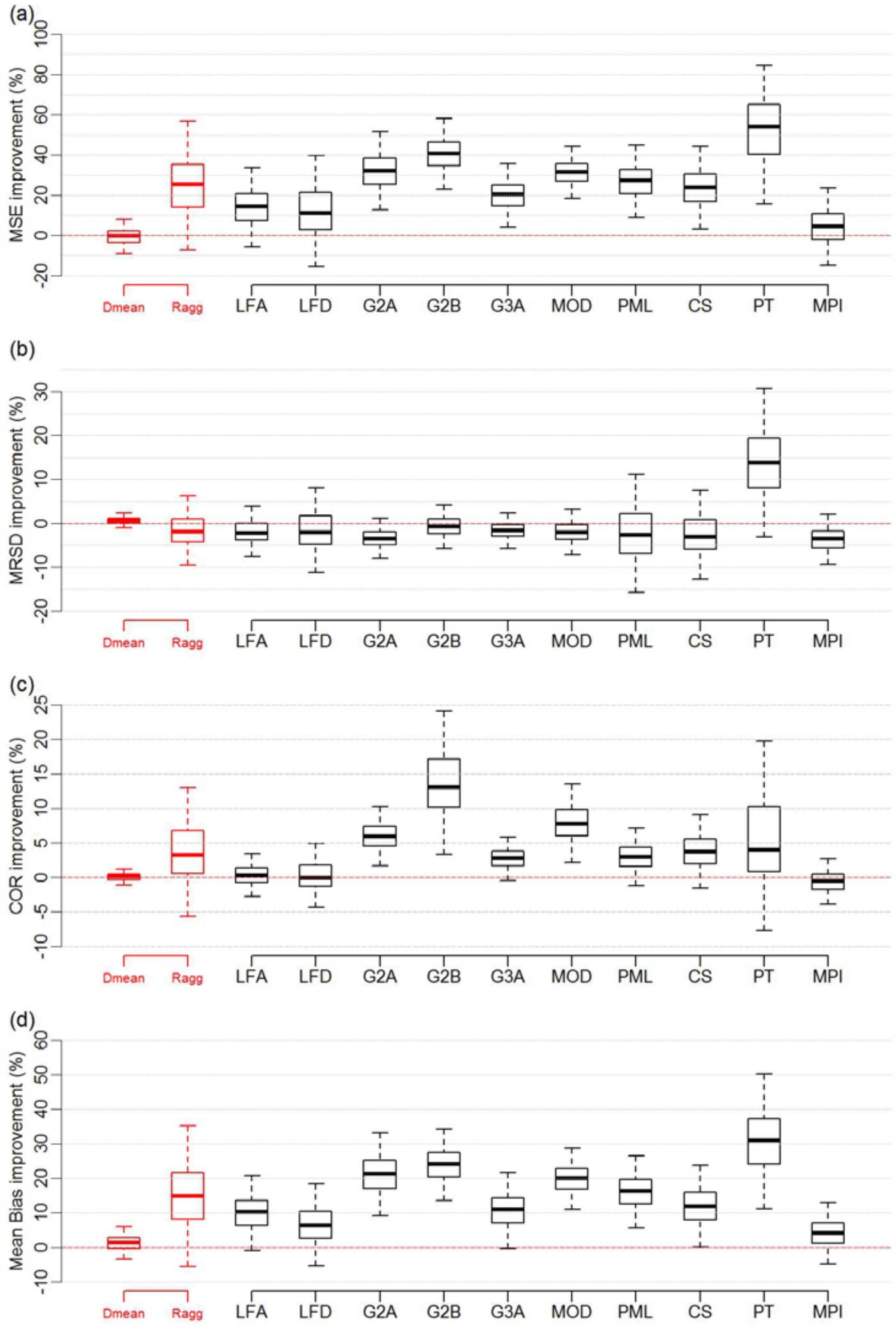

Figure 11: Box and whisker plots displaying the percentage improvement that the weighted product excluding MPIBGC provides in the 25% out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias (d), when compared to equally weighted mean (Dmean) of the Diagnostic Ensemble, aggregated 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

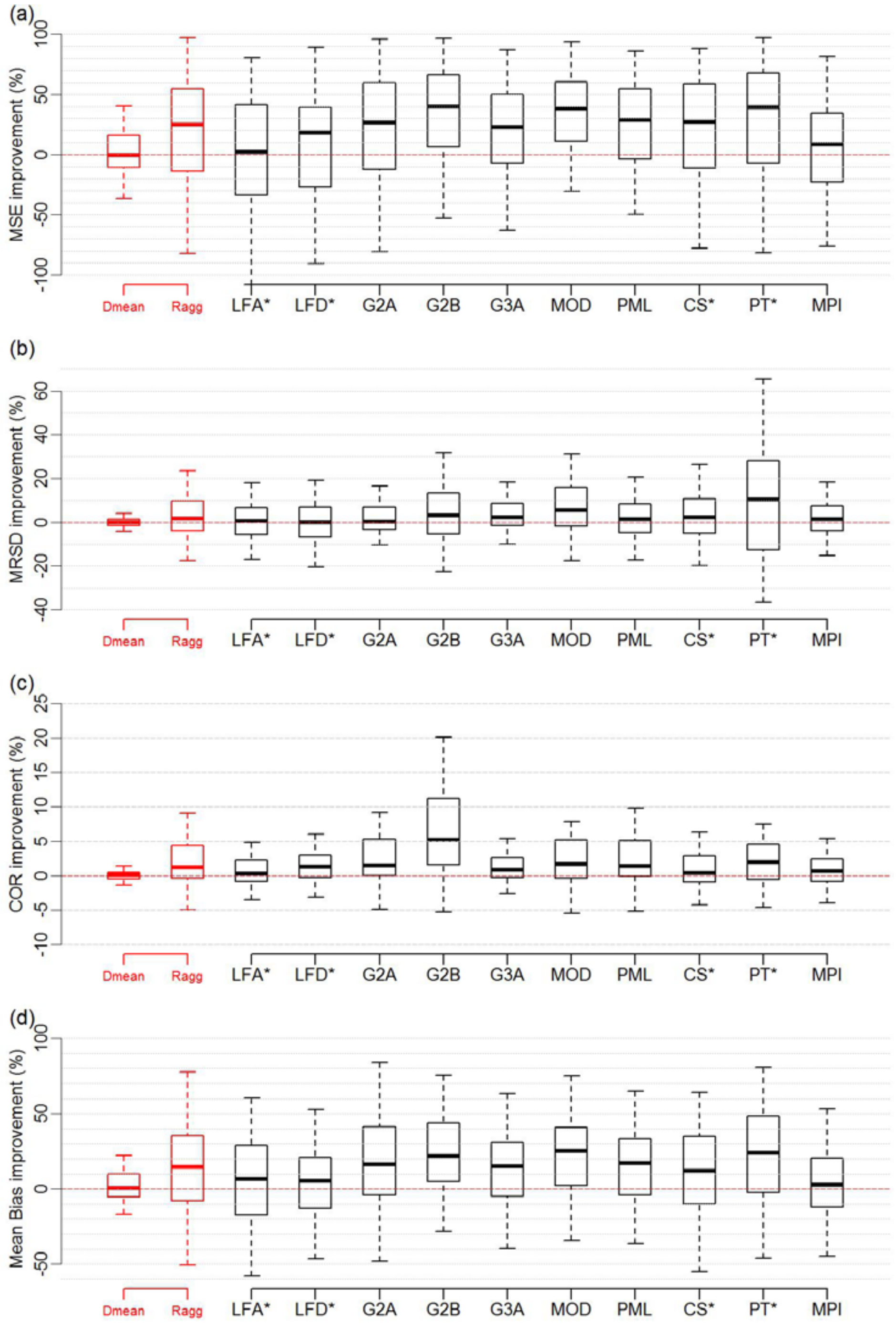

Figure 12: Box and whisker plots displaying the percentage improvement that the weighted product excluding MPIBGC provides in the one out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias (d), when compared to equally weighted mean (Dmean) of the Diagnostic Ensemble, aggregated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

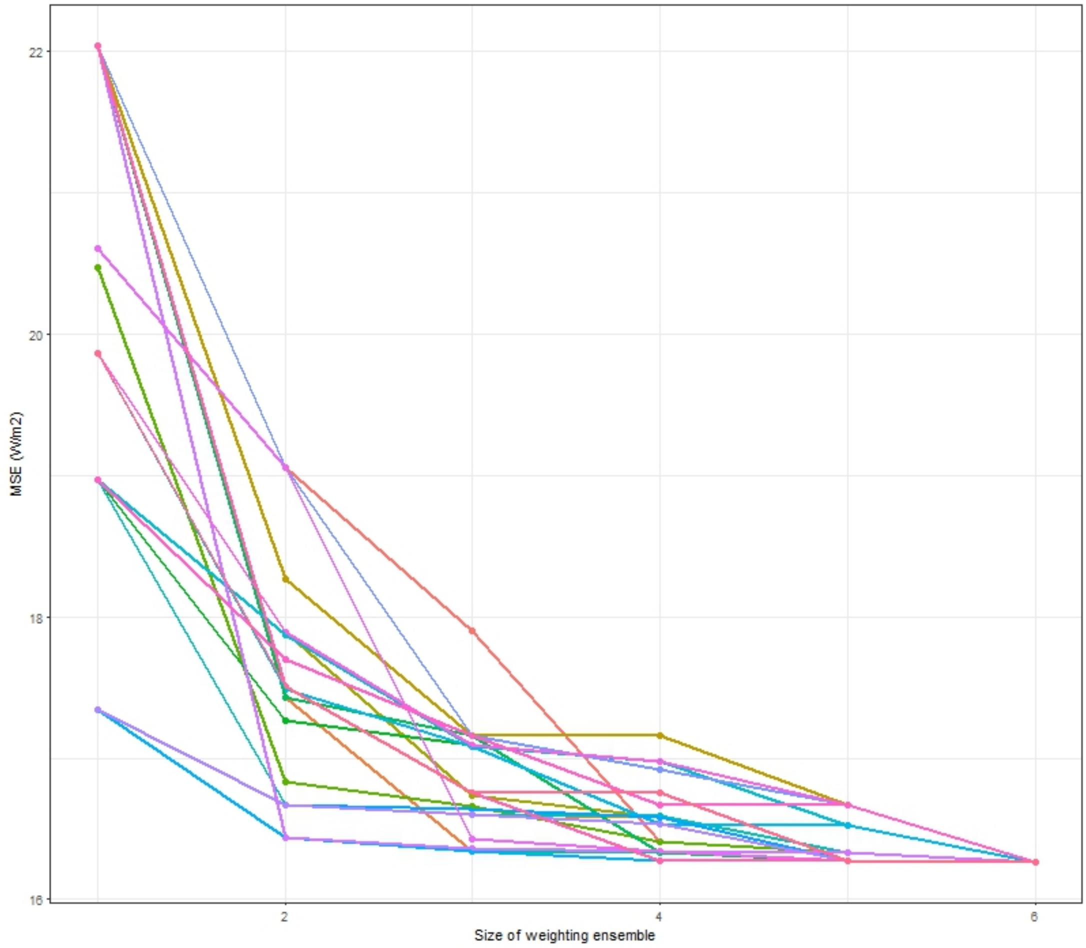

Figure S1: The results of the in-sample test showing how the Mean Square Error (MSE) of the weighting changes when we increase the number of ET products involved in the weighting from 1 to 6. The test was repeated 25 times of a random selection of n products (2 ≤ n ≤ 6).

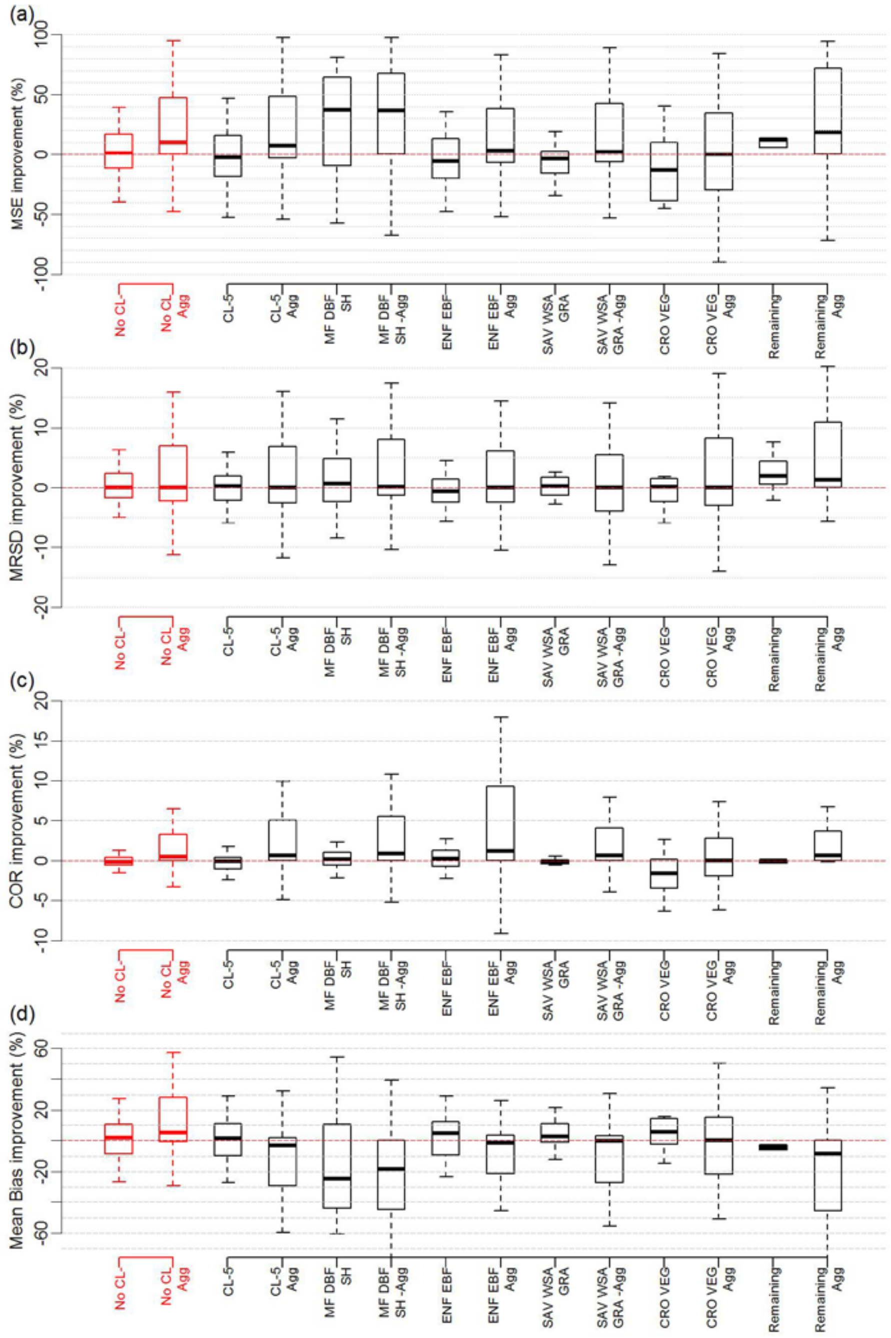

Figure S2: Box and whisker plots displaying the results of the One out-of-sample site test for the cluster independent weighting (No CL and No CL Agg boxplots), the cluster dependent weighting (Cl-5 and CL-5 Agg) and over individual biome types for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias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