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

We would like to thank Carlos Jimenez for his constructive comments on our manuscript. This document outlines our responses to his comments and the improvements made to the manuscript.

Response to Short Comments

Product Selection

While acknowledging the difficulties of finding the rights products to derive a synthesis product, I am a bit surprised about some of the choices.
As the authors state, product selection should follow the criteria of product diversity, so ideally single algorithms with different strengths and weaknesses should be combined together. In that sense, I would have not considered an already synthesis product such as LandFlux-Eval as a possible candidate for the merge.
In my opinion, a more valid alternative regarding LandFlux efforts could have been the three single products publicly available based on different ET algorithms (https://hydrology.kaust.edu.sa/Pages/GEWEX_Landflux.aspx**). Also, I do not see much interest in combining obsolete versions of products with the current, and presumably better, product, as it has been done for GLEAM. Just GLEAM V3A (and perhaps GLEAM V3B) seems to me a better option. An interesting product is MPI. This is a global extrapolation of the tower fluxes, the same tower fluxes that are used to decide on the merging weights, and quite different to the other products, which we could consider more “physically” based and less “calibrated” with the tower data. It is not surprise then that MPI is by a large margin the more weighted product. I do not deny that it can be a valid product for the merge, although much less independent than the other products with respect to the tower data. In that sense, it could have been very**

informative also to see how the merging works, and how the weights are distributed, when that product is left out

Yes, this is indeed a reasonable question. One key distinction of the weighting approach here is that it accounts not only for the performance differences between products, but also the error covariance between them (as noted in Section 2.3). So, if a product were added that were a near copy of another, it would not degrade the performance of the weighting at all. While variants of GLEAM are indeed likely to be similar, small structural differences might mean that there is in fact an advantage to using a nominally inferior version in addition to the latest version. This might explain why GLEAM-V2A was assigned a negative coefficient when the three GLEAM products participated in the weighting (tier1). By assigning a negative weight, the weighting was removing redundancies or data that was not adding any information to the weighting. It was not possible to include GLEAM v3B in the current version, due to the limitation in the covered time period which doesn't include 2000-2003.

Testing how the weighting would perform *without* MPI is an interesting idea. We therefore did exactly that - removed MPI from the weighting as a separate experiment and included this in the manuscript. Performance, perhaps surprisingly, was very similar:

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, 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, o.2 for GLEAM_v2B, 0.13 for MOD16 and 0.07 for GLEAM_v2A.

The new plots Fig.11 and 12 are shown below

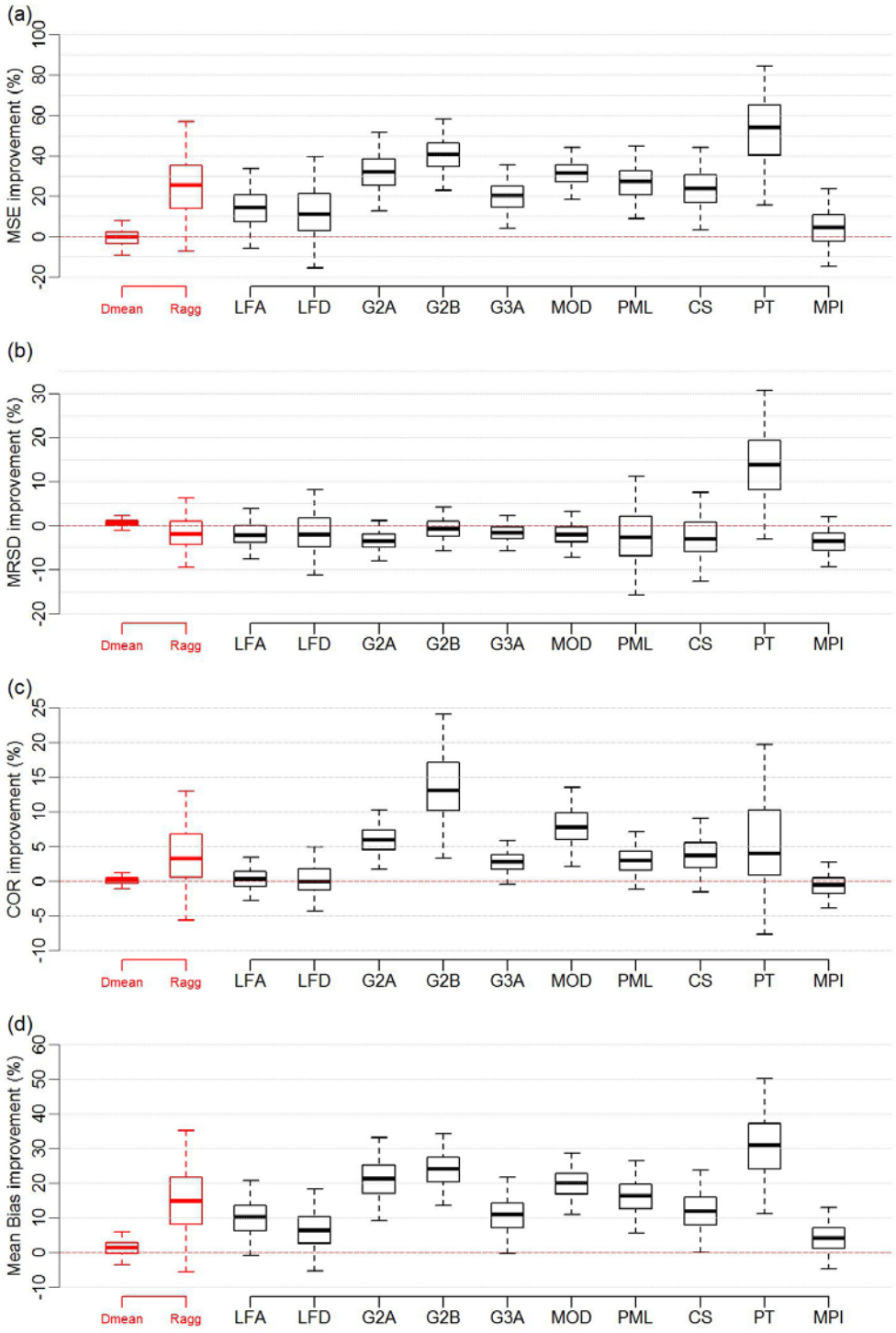

Figure 11: Box and whisker plots displaying the percentage improvement that the weighted product excluding MPIBGC provides in the 25% out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias (d), when compared to equally weighted mean (Dmean) of the Diagnostic Ensemble, aggregated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

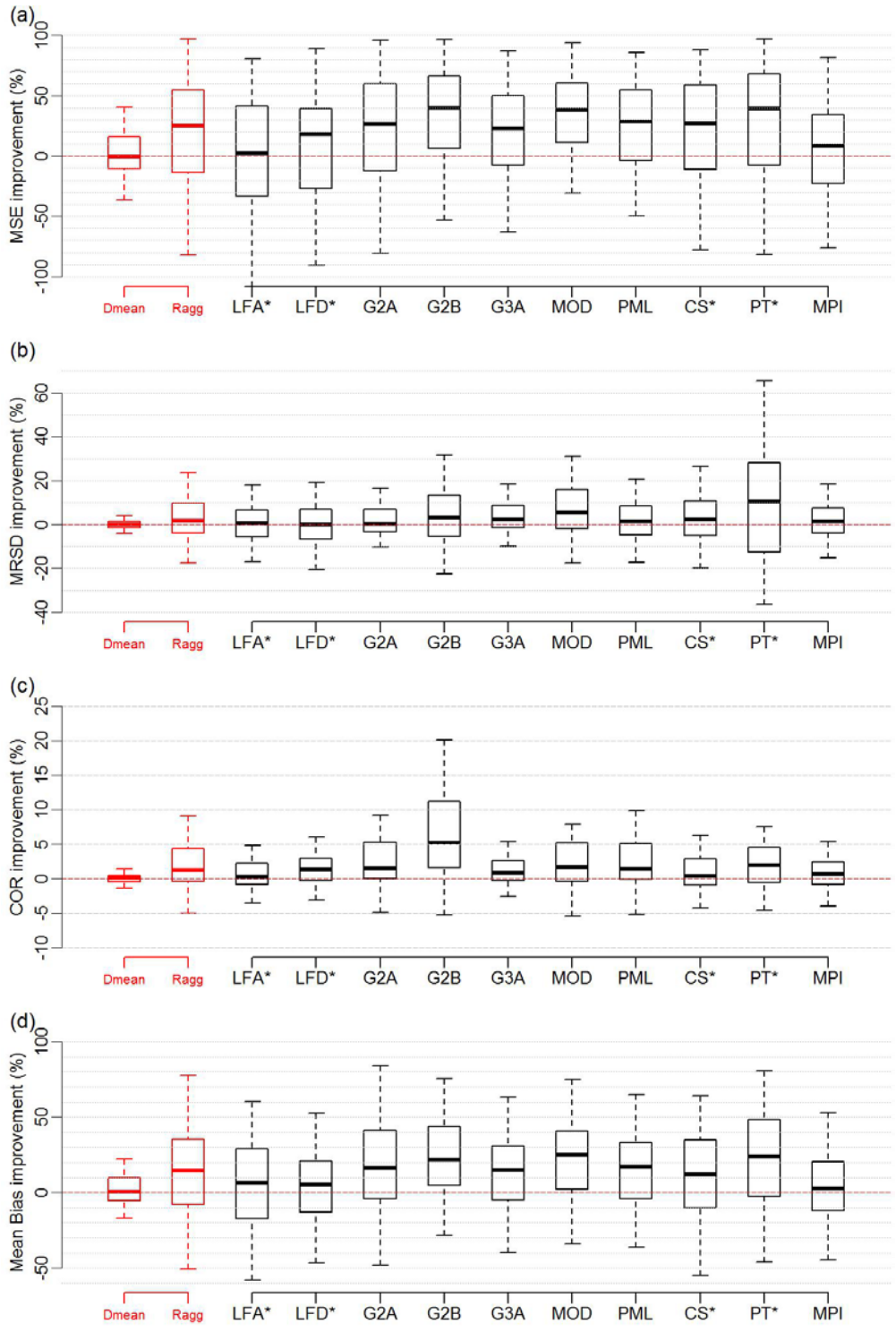

*Figure 12: Box and whisker plots displaying the percentage improvement that the weighted product excluding MPIBGC provides in the one out-of-sample sites test for four metrics: MSE (a), MRSD (b), COR (c) and Mean Bias (d), when compared to equally weighted mean (Dmean) of the Diagnostic Ensemble, aggregated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

Spatial and temporal resolution

The resolutions given in the Table 1 seem wrong for a number of products (e.g. MODIS original resolution is 1 km, Zhang 2010 is 0.05 deg, MPI and Zhang 2015 0.5 deg). Also, the periods of available data should also be revised for GLEAM (I think GLEAM V2A covered 1980-2011 and V3A 1980-2014).

Thanks for spotting this, we have corrected the errors in Table 1

As far as I can see, the only datasets limiting the study period to 2000 is MOD16 and GLEAM V2B. Perhaps the products going into merging could have been separated in two groups, a bit similar to what has been done regarding geographical coverage: from a much earlier year than 2000 will all but MODIS and GLEAM V2B, and from 2000 including MODIS and GLEAM V2B. That would have resulted in a much longer DOLCE dataset, presumably based on a larger collection of tower data and more ET products. A shorter time period than monthly will result is a more useful product. Daily will a be a better objective for future developments, although it will require a different selection of products, possibly more based on the "physically" based diagnostic ET products, where daily is a common time scale. I would suspect a more complex merging exercise, given the larger amount of time variability that needs to be captured by the merge product ET and FluxNet datasets

Great comments and suggestions. In this work, limiting DOLCE to be derived from diagnostic (as opposed to overtly model-based) products only made the choices of temporal and special resolution limited. As suggested, extending the period of DOLCE by using different products for different time slices of DOLCE is a great idea, and will likely be an objective for further development in future versions of DOLCE. Applying the same merging technique on daily diagnostic ET products requires more complex analysis at both the site scale (observation) and grid scale (gridded products), although this is worth investigating for future versions. Thanks!

Tier 3 (i.e, Greenland and Antarctica) is just a very close weighting of an obsolete version of GLEAM and a newer version. It seems a bit awkward to distribute that as

part of a synthesis product. It may have been better to just remove those regions from the synthesis product given that nearly no one dares to estimate ET over there (understandably).

Yes, agreed. We however wanted to produce DOLCE with global coverage (requiring tier1, tier2 and tier3), and did make an effort to show the level of reliability of each tier in Fig. 8. We published tier1, tier2 and tier3 in separate files in order to give the users of DOLCE the flexibility of selecting the most appropriate product.

Some of the ET datasets considered are based on algorithms that estimate separately interception, evaporation from the canopy, and evaporation from the soil (e.g, MOD16, PT-JPL, and GLEAM). Under the assumption that routine EC observations perform very poorly for rainy conditions, in principle interception is not captured by the tower observations. In some recent ET evaluations of these products the tower data has been filtered to remove rainy periods and the interception component has not been evaluated. A discussion about this could have been interesting, given that, as far as I can see, the tower data is not filtered for precipitation conditions, and the merged product is a total ET product.

Yes, we haven't filtered the sites for rainy periods, especially since we're working with monthly data. Something to explore in the future, especially if we investigate using daily data.

Regarding the energy closure issue, the text may give the impression that fluxes correction is always possible, but a large number of stations do not measure Rn and/or G. Given that only corrected fluxes are used in the study, I imagine that a number of stations have to be discarded as they corrected fluxes were not available (FLUXNET-2015) or could not be estimated (LaThuille-2007). This may be worth mentioning

We only corrected LaThuile sites for energy imbalance at sites that had measurements of all the component fluxes. We applied quality control and filtering for G, H and Rn as

highlighted in steps (2) and (3), section 2.2. This is perhaps not clear enough in the manuscript, we therefore 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.

Merging technique

If I understand the method correctly, the weights are global (i.e. one value for all pixels), time-invariant (i.e., an annual value), and the bias-correction is what Bishop, 2013 calls a "global bias" correction (i.e., is a single value per product using all towers in the in-sample training dataset). If this is true, if Product A was performing better than product B over some biomes and/or at some periods, the method cannot be used to weight more or less the products to reflect that difference in performance. If I am correct, I wonder if there is a way to modify the weighting to take into account those differences. We typically see that ET products perform differently at different biomes and/or seasons, so it may be advantageous to capture this in the weighting. Bishop, 2013 was quite illustrative about this. Given that for that temperature example the "truth" was quasi-global (i.e., not over a very few pixels like for the ET), just a "per-cell bias" correction, even without weighting, outperformed the weighted product with a prior "global bias" correction. Of course, the "per-cell bias" correction and weights cannot be be applied here, given the limited geographical coverage of the flux "truth", so the problem is more complex.

I guess your first take at this is your sites clustering based on vegetation type, presented in some detail in the Supplement. You conclude that it did not improve overall the DOLCE performance based on a global analysis, but I would be interested to see the results (in the Supplement I can download the figures summarizing the

analysis are missing). Perhaps other schemes that better cluster the flux behavior are worth investigating for future versions of DOLCE

Yes, we calculated a single bias term, and assigned one weight to each ET product to be applied globally. As you also note, we did apply clustered weighting, where we derived cluster dependent sets of bias terms and weights for weighting the ET products. We tried clustering by 1) vegetation type, and also 2) climate zone and 3) aridity index (2 and 3 not shown in this study), and we implemented the same one site out-of-sample test, but this time separately in each cluster. We calculated different sets of weights for the ET products based on their performance differences in the different clusters, but none of the clustering succeeded in deriving a better weighted product overall. We added the text below to clarify the clustered weighting.
We show the results of one site out-of-sample test where different weights were assigned to products at each biome type (in supplementary material). We apologize for not putting the figure in the supplementary material, this has now been added.

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 with different sets of weights in each cluster.

Results

It is stated that the MSE plot in Fig. 3a shows the MSE of the weighted product being better than the ensemble mean, but I do not see it. The central line of the first whisker

box in Fig 3a is at the zero line. Perhaps I am missing something regarding how to read these plots. As a side note, it may be good to say what the end of the whiskers represents. In most occasions it is used to represent the max and min of the data, but it is not always the case.

We have modified both the plots and their description in the manuscript. As noted in our response to Reviewer 1, we have modified the RSD metric and added the performance of DOLCE on mean values as well in a fourth panel. We have also clarified the description of the box plots as follows:

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(dataset) and 2) (Q_3 + IQR), while the extreme of the upper whisker is the maximum of 1) min(dataset) and 2) (Q_3+ 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 by the boxplot.

Note that (a-c) are explained in section 2.4

In the same Fig. 3a, the whisker boxes for the individual products make me think again about the MPI product. Based on the "heavy" calibration of MPI with the same tower data used to derive the weights, I would speculate that if the MPI product was removed from the merging, the percentage improvements of this new weighted product (i.e., without MPI) over the individual products will be much smaller. This may give a different perspective of the exercise regarding the skill of the tower-based merging to combine the more "tower-independent" physically based products.

Please see our response to a similar point above earlier in this document (in Product selection).

If we just concentrate on the improvements of the weighted product with respect to the equally weighted product (i.e., first whisker box in Fig 3 a-b-c), the gain in performance of the weighted product seems small. Again, if I read these plots correctly, the gain for MSE and COR is minimal, only the RSD shows some improvement. But given the definition of the RSD metric, I wonder is this is mostly associated to the bias correction. If I understand this correctly, after the bias correction mean-dataset and mean-observation will be equal, so RSD is abs(sigma-dataset –sigma-observation). In other words, I am wondering about a comparison of the equally weighted product, but with a bias-correction first, and the weighted product. I am assuming here that the equally weighted mean did not involve a prior bias correction, as nothing was stated in the paper, but I may be wrong.

Yes, the plots in Fig 3 a,b and c show that overall, the improvement of the weighted product is noticeable with respect to the reference products and marginal with respect to the equally weighted mean. As a result of this comment and those by Reviewer 1, 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 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.*

It is also true that the equally weighted mean used here doesn't involve any bias correction. Of course, some of the improvement offered by the weighting is owing to the bias correction and some comes from the weighting. We have now separated these effects in Figure S3 in the supplementary material, and referred to it in the results section:

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.

In the HOM and HET comparison, I see very small improvements in MSE, larger for RSD, and not much for COR (the median of the whisker box is for MSE and COR is at the zero percentage line). And I wonder if the separation into HOM and HET sites may have also implied a separation in land covers, so the improvements we see are more related to the weighted product working better for some specific biomes. One may think that land covers such as forested areas are more likely to be represented in the HOM class, compared with e.g. croplands. I wonder if this has been checked, i.e., that the biome representation in HOM and HET classes does not change too much.

Yes, the weighting calibrated by HOM sites offers only a marginal improvement over the weighing using all the sites (HOM and HET). We looked at the separation of Biomes in HOM and HET cases and we found a clear separation of land covers and their distribution across the HOM and HET case:

1) cropland constitutes more than 20% of the HOM-case sites and 7.6% of the HET-case sites,

2) about 23% of HOM-case sites are forests (EBF, DBF and MF), while more than half of the HET-case sites are forests, the big number of forest sites in the HET case is because most of the MF sites are located on grid boxes identified as EBF and DBF.

3) WET and WSA sites are found only in the HET-case

We added table S2 below in the supplementary material and we modified the text.

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*

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.

Regarding the boxplots of Figure 5, it is true that the end of the whiskers are larger for the HET sites, but the RMSE and correlation median and percentiles look slightly better for the HET class. I think this is also worth discussing, as it can potentially indicate again that the differences in performance between HOM and HET classes are small, with just a few HET sites having bad statistics. Given that this is based on one site out-of-sample, I wonder if the bad performance at some individual sites may be nothing to do with homogeneity, but with the fact that the site is one of a kind, so any weights derived from a sample without that site are not informative. Given the location of some the sites, it would not be a surprise.

An interesting point. It could certainly be the case that the reason for some sites in the HET group showing poor performance of DOLCE might simply be about measurement quality or site uniqueness. The box plots in Figure 5 show the performance of DOLCE in the HOM and HET cases separately. Yes, while the whiskers are larger for the HET sites but we cannot infer from these plots that DOLCE performs better in any of the two groups of sites. We modified the text in the Results and Discussion to make this point clear.

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.

And

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.

We might expect that (3) would decrease the overall performance of the HET-case sites (see Fig.3). On the other hand, it is very likely that(1) and (2) increase the performance of HET-case sites and eventually compensate the decrease of performance due the WET sites.

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 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.

MPIBGC is the larger contributor to DOLCE, with a weight close to ~0.5. Perhaps it is not surprising than the differences of MPI with DOLCE sown in Figure 6 are smaller than with LandFlux-EVAL-Diag in Figure 7, which is not part of DOLCE.
I acknowledge that it is quite difficult to come out with the right classes to try to illustrate the reliability of the weighted product. But when I look at figure 8-c what I broadly see is that arid places, Greenland and Antarctica have low reliability, snowed places medium, and the rest high. Perhaps just the uncertainty over the mean ET would have been more informative.

Great suggestion. We've expanded the plot and showed the seasonal variability of 1) ET estimates and 2) uncertainty estimates, and we've changed the plot titles and rewritten 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.*

Discussion

The discussion about a possible selection of HOM sites to construct DOLCE is quite appropriate. Being very strict, and considering that the typical tower fetch is of the order of hundreds of meters, while the grid cell has an area of ~2500 km2, possibly none of the sites will truly qualify as HOM. But if we want to keep using tower fluxes, we need to live with this. Perhaps a simple measure to reduce the effect of this tower-fetch and cell-scale mismatch is to try to work at finer resolutions in the future. GLEAM, MOD16 and Zhang, 2010 are already at resolutions 0.25 deg. PT-JPL will

be soon available at 5 km. Perhaps a couple of products will never be available at 0.25 deg (e.g., MPI), but they could be downscaled to 0.25 deg and merged with the other products (e.g., by a nearest-neighbor interpolation if we do not want to add any extra information). The lack of success of any clustering of the sites based on vegetation, climate, etc, possibly is more indicative of the limitations of the tower flux, rather than limitation of the conceptual idea. Even if not terribly successful, a look at how the relative weights of the different products change for different clusters can be informative regarding the products performance for different conditions.

Yes, both are good suggestions for future work. As noted in our response to Reviewer 2, when globally gridded ET products allow us to derive DOLCE at higher resolutions, there are obviously many benefits. More tower data will also obviously help, especially with this second point – understanding which gridded products are better suited to different conditions.

As I mentioned before, using MPI as part of the weighted product is perfectly valid. But I think the main feature defining MPI for this study is not the fact that it is a statistical product, but more that it is a global extrapolation of the same tower fluxes used to derive the weights. As clearly stated in the paper, the tower fluxes have limitations, which will have an impact on the weighted product. Presumably, the MPI product will suffer from the same limitations than the tower fluxes, but these limitations will not become apparent when looking at differences with the tower fluxes, so it will not be reflected in the weights. An example could be the ET estimation for those biomes and moments when interception can be a relevant component of the fluxes. If the tower fluxes are not properly capturing this component, the physical methods that try to do so (e.g., GLEAM, PT-JPL, MOD16) may be penalized in the weight derivation, compared with MPI, but their fluxes may be more correct. Of course, the problem is how to merge or validate the ET products away from the tower flux data. Given that DOLCE is a global product with a relatively long time series, at least the annually integrated ET could be compared with basin-integrated differences of precipitation and runoff, which may shed more light into the merits of the individual products, the equally weighted mean product, and DOLCE.

Yes. We hope that our addition of DOLCE without the MPI product to the manuscript (as described above), and the fact that it still performs well, goes some way to alleviating this concern. At the annual scale, we could indeed neglect the change in water storage, and compare evapotranspiration with the difference (Precipitation - Runoff). This difference could serve as a test dataset for DOLCE, the equally weighted mean and the reference ensemble, away from flux data, and is independent from both MPI and DOLCE. However, this validation exercise requires observational datasets for precipitation and runoff. While there is indeed a big network for observed precipitation, gridded products can vary by surprising amounts, and currently there are no global time varying observational estimates of runoff that we are aware of. Year-to-year storage changes are also not necessarily negligible. Nevertheless, validating DOLCE as a component of the water balance is part of our next piece of research… so we clearly agree this is an aim worth pursuing.