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

Dear Dr Su,

Thank you for your response and taking the time to seek additional reviews. Below we discuss both of the points you raise, as well as the additional points raised by the reviewer, and some additional investigation we did to allay these concerns. We start by attempting to give a clearer justification for the position we have taken on these in the manuscript, before outlining possible ways forward.

Reviewer #2 pointed out that it is not convincing to compare 0.5 degree grid data to in-situ flux tower measurement. I would suggest that you consider a scaling technique based on land cover so that the two are comparable. A simple foot-print consideration should also be very useful.

We wholeheartedly agree that point-scale tower measurements, with a footprint no larger than 1km^2, cannot be assumed to be representative of each 0.5 degree grid cell. However we feel that the reviewer's assertion "to assess 0.5 deg. ET product with flux towers is a kind of silly problem" misrepresents the fact that we explicitly addressed this issue in great detail and also does not acknowledge that almost all existing gridded ET products have been tested in this way in existing published research (e.g. McCabe et al., 2016).

We dedicated a significant part of the introduction to establish that this is a problem with existing published approaches and developed an experimental approach to explicitly test the validity of this assumption. Indeed we feel it is one of the most novel aspects of what is presented here, since almost all existing studies effectively ignore this issue.

We have made changes to the manuscript to try to make this clearer, which we will outline below, but first restate our approach as it was submitted to demonstrate this issue is far from ignored. In the introduction we note that:

"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 km^2. The fetch of flux tower measurements varies depending on terrain, vegetation and weather, but is typically under 1 km^2 (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."

"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."

In Section 2.4 we detail how we can test whether point scale measurements do, on average, actually provide any information about the 0.5 degree scale. This is done by testing whether the weighted combination of existing products, that clearly will improve results against site data with which it is trained, can still deliver improvements at sites that were not included in its training data set. If it does better than no weighting at these unseen sites, then there must be information content about the larger scales in the site data. This is precisely what is shown in Figures 3, 4, 11 and 12 – all of these performance results are ONLY for sites that were not used to train the weights (i.e. they are out-of-sample tests).

While the results are not spectacularly good, they are solid:

"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 then showed that restricting the application to sites where the tower was more representative of the 0.5 degree grid cell did indeed improve results further. We also feel that we are clear about caveats in this relationship:

"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."

And feel that our conclusion of this part of the investigation clearly reflected the findings:

"It was shown that despite the scale mismatch between the flux tower and the grid cell, the ensemble of flux towers as a whole can provide information about the grid cells that contain them. 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."

In short, we feel that the reviewer's one sentence dismissal is an unfair portrayal of what the manuscript contains. We note that the other reviewer, Paul Dirmeyer, explicitly endorsed the approach we took regarding this issue.

Nevertheless, we have clarified this further in the results section :

> *" 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.."*

And in the conclusion:

> *"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."*

2. Another issue is your use of available ET data - I am in the same opinion as reviewer #2 that you should consider more available data. I do not think your response to this concern is convincing as such.

We are absolutely interested in using as many products as possible, and can incorporate more relatively easily. Our primary aim for DOLCE is land surface model evaluation, and as such, we have avoided including any products that are derived using similar model structures, such as reanalyses.

We also thought of including other products, in particular SEBS, CSIRO-global and PTJPL, however none of them have a full coverage of the period 2000-2009. This period was in fact chosen to maximise the number of products we could include while still having a final product that was at least a decade in length.

One potential way to address your concern is to derive DOLCE with different component products in different time periods, however we feared that doing so would lead to temporal discontinuities in the derived product. We have now stated as much in section 2.1:

> *“ 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.”*

If you have a suggestion for how this could be made clearer please let us know.

Response to Reviewer:

We also address the reviewer’s concerns individually here:

1. *To assess 0.5 deg. ET product with flux towers is a kind of silly problem.*

Please see our detailed response above – in short, we feel that this comment ignores the very extensive treatment of this issue in the manuscript, particularly that we clearly showed as part of the results that this was not a ‘silly problem’, that flux towers did in fact reliably provide information at the 0.5 degree scale. We also feel that the comment ignores that most constituent products used here were evaluated, and published, in this way.

2. *Why creating ET at a finer resolution (0.05) was not possible, by using the ensemble weighting and rescaling technique? Indeed, if there is a way for higher resolution, I suggest to rethink on this. As in next years, fine resolution ET will come out.*

As noted in our previous response to this same point, this is not the aim of our study, and only one product has information at this resolution, so even if we wished to apply the novel methodology in this paper to this scale, it would simply not be possible.

3. *In addition there are other ET dataset, such as PTJPL, SEBS, GLDAS, ERA-Interim, which should be used to enlarge your ET input source, if you contact the groups.*

As explained in more detail above, we would love to have more products in DOLCE, indeed our approach to component product selection was precisely to maximise the number of products we could include, but for the given time window, and our desire to keep the product applicable to land surface model evaluation, this is the maximum set currently available to contribute.

4. *If the author doesn`t expand the time coverage or spatial resolution before the paper is published, it is mostly likely not possible for DOLCE ET updated after the publication.*

This seems an odd comment. Why would a change in time coverage or spatial resolution affect our ability to update the product in the future? At least to us, this is incorrect.

5. *As bother reviewer and editor have suggested using other global ET, but no real reaction is adopted.*

We assume this is the same concern raised in point 3 above – please see our response to (3) and to the editorial comment above.

6. *Don`t agree that there is not enough flux towers for each land cover and biome types.*

This is simply about statistics. While we can fit several ET products to a small amount of flux tower data, this will result in overfitting and poor out of sample performance. It is an extremely widespread rule of thumb in statistical literature that the bare minimum number of data points required for a meaningful regression relationship is approximately ten times the number of predictors. We can provide references if needed, but feel this is fairly rudimentary.

7. *Even the clustering by biome type doesn`t improve the weighting, however, it can help you derive a high resolution ET.*

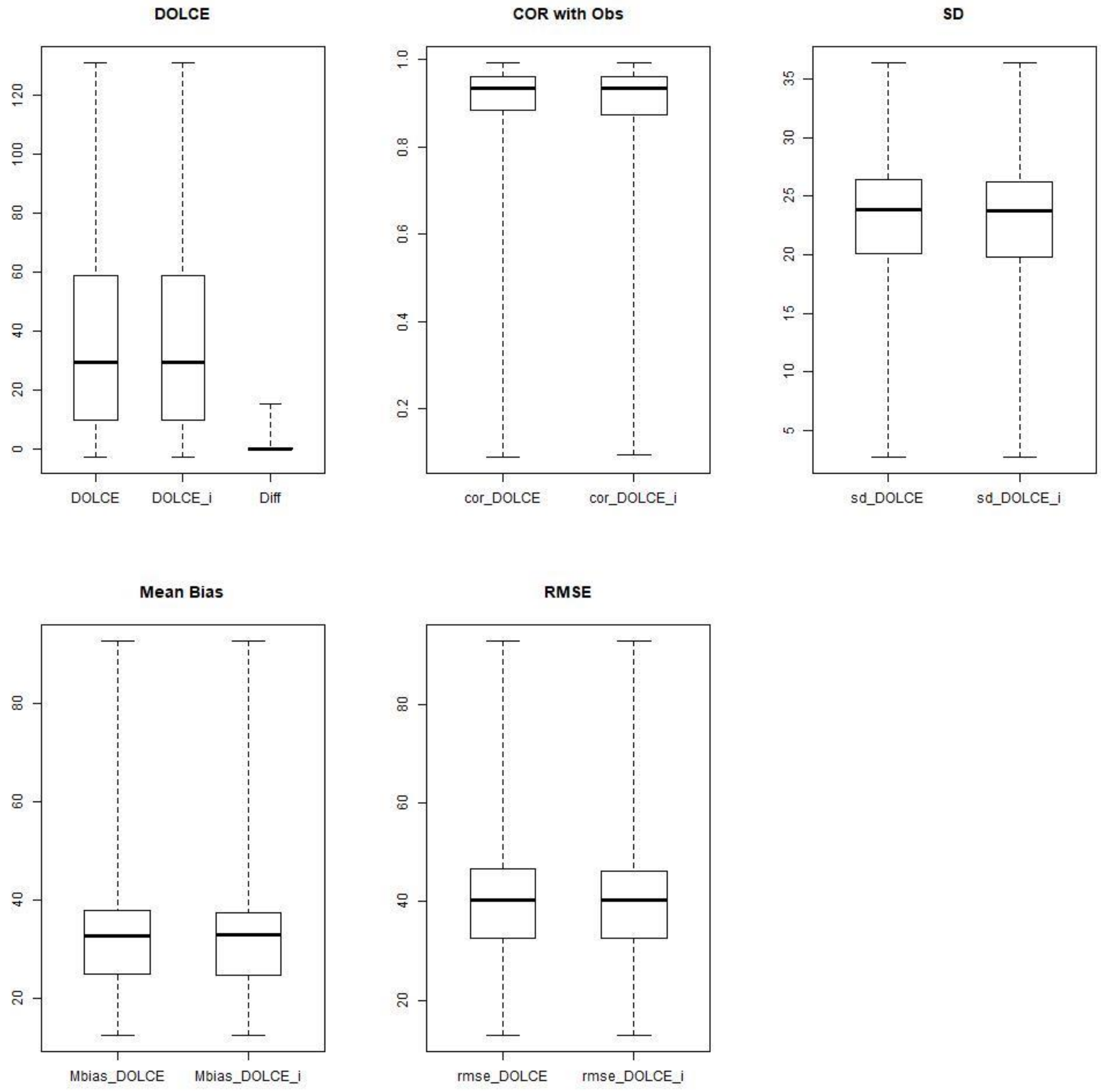
Please see our response above, and in our first set of responses as to why we do not derive a higher resolution product.

8. *Answers to comment 7, whether to do spatial interpolate or not does not influence your flux tower evaluation result? Please re-check this. If I am right, here you are selecting the pixel value where the flux tower be located in to match with flux observation. However, you can also use 2-d spatial interpolation to get the point ET value with the geo- location of the tower. Please check if this will influence your weight, mean bias, and SD. Then you can say it`s not necessary to calibrate weighting ET at higher spatial resolution with flux observations.*

Firstly, as explained in our response to (2) above, we note that a higher spatial resolution product using the novel methodology in this paper is simply not possible, as only one component product contains information at an appropriate scale. It is also not the aim of this study.

Spatial interpolation at 0.5 degree to compare tower and grid cell data is of course possible, and is indeed a sensible suggestion. To understand whether this was likely to have any qualitative effect on our results, we compared the grid cell values of DOLCE that contained each flux tower with DOLCE values at the towers using bilinear interpolation. Results are shown below. The first panel compares the ET values of grid cells (‘DOLCE’) with interpolated values ('DOLCE_i'), as well as box and whisker plots of the differences between these two. The remaining panels show comparisons to flux tower data using the key metrics used in the paper – as is shown in Figure 5 of the manuscript. As you can see this makes no qualitative difference to the results. As such we see no reason to further complicate the methodology of the paper, but have noted that this avenue was explored in the discussion section of the manuscript:

> *"We also investigated using bilinear interpolation instead of direct grid cell to tower comparison (not shown), but found no qualitative differences."*


DOLCE
DOLCE
DOLCE_i
Diff
COR with Obs
cor_DOLCE
cor_DOLCE_i
SD
sd_DOLCE
sd_DOLCE_i
Mean Bias
Mbias_DOLCE
Mbias_DOLCE_i
RMSE
rmse_DOLCE
rmse_DOLCE_i