Reviewer 3

This manuscript describes work to combine two ET products, PT-JPL and GLEAM, using a weighted average, with weight determined by fit to tower observations. The resulting product is limited to the locations of the towers, and no attempt is made at extrapolation to other sites. While the manuscript is well written, and the analysis sound, the work itself is not well motivated and, as currently presented, does not represent a significant contribution.

R. We thank the reviewer for taking his/her time for a detailed review of our paper, but we certainly disagree about his/her rating of our paper. We also clarify that the resulting merging method is a first and necessary step towards a global merger that is certainly envisaged.

The merged product presented in this manuscript does not add any value to the ET products that are already available. The motivation seems to be to merge the two ET products (PT-JPL and GLEAM) to produce a new product that is as close to the tower ET time series as possible. How is this new merged product then any more useful than the original tower ET time series?

R. The local merge of GLEAM and PT-JPL at the selected towers was the first step to produce a merge product. Perhaps we were not clear in the motivation and objectives. We fully agree with the reviewer that the merge product would be useful outside the locations of the towers, but not where we already have the tower estimates. But the first step is to prove that the merge product fits the tower data better than the individual products at the tower sites, and this is mainly what this paper is about. It is not an obvious exercise, as the tests carried out in the paper show, and we definitely think that it is worth publishing.

The merged ET product has not been independently evaluated. It is shown to be closer to the tower observations, but this is by design. Given that the tower observations also have errors (and given how closely the merged product has been fit to the tower obs), it does not follow that the merged product is necessarily more accurate. I am concerned that the merged product is over-fitting to the tower obs (weights calculated independently at each location, using a moving temporal window).

R. Over-fitting is always a concern with these statistical approaches, but for the moment we are just trying to show that at each specific site the optimal

estimator can result in a product better fitting the tower ET. Now, for the global merger over-fitting definitivel needs to be tested. If the merge product were doing well at a specific location, but poorly in a similar but different location, then we would have over-fitting, or poor generalization issuess. The usual approaches to test this could then be applied, such as the cross validation techniques where stations are grouped by land cover and each land cover dataset stratified in independent parts to derive the weights and test the weight performance. Still, these are in a way "self-contained" tests and not independent evaluations, as we keep using tower data, and we can only prove that the merged product is more "accurate" with respect to the tower data. At the individual site scale, we do not see how else the merge product accuracy can be tested, and we would be happy to hear suggestions from the reviewer in this sense. If spatial integrations are allowed, then more "independent" evaluations can be carried out by comparing with "inferred" ET.

Certainly the tower ET also has errors, as described in the paper, and any methodology that tries to fit to the tower ET is likely to inherit those. Nevertheless, there is some consensus in the ET community that the tower fluxes are our best shot for ground "truth" at ecosystem scale. The optimal linear estimator applied here tries to minimize the error variance of the merged product with respect to a reference, in this case the tower observations, and in that sense certainly by definition the merge product tries to get closer to the tower observations, compared with the original ET estimates.

The work is not very well motivated. Why merge just these two products? Why not merge as many as are available, or as many as meet some pre-defined standard? The selection of just these two products is particularly awkward given that they are not independent.

R. We stated in the Introduction the reasons behind merging GLEAM and PT-JPL. In short, after years of testing different methodologies to derive "satellite-based" ET products, GLEAM and PT-JPL showed more skills than others tested methodologies (Michel et al., 2016, Miralles et al., 2016, McCabbe et al., 2016), so we wanted to see if we could merge them to produce a better product. We think that this is a legitimate objective in the framework of our WACMOS-ET project and connected initiatives, such as the GEWEX LandFlux initiative (https://halo.kaust.edu.sa/Pages/GEWEX_Landflux.aspx), and we do not see anything awkward here even if the products are not completely independent. So, although merging a large number of products is also a valid objective (e.g., the

recent Yao et al, 2017), this is not our objective in this study.

To be publishable this work must i) provide a product that adds value in some way to the original products., and ii) the resulting data set must also be independently verified.

R. We disagree that research on this topic can only be published if it results in a new product. We believe that what we learned about merging these two specific products is of broad interest for other colleagues working on these topics, even if it is just a firs step for a successful merger.

The most obvious way to achieve this would be to spatially extrapolate the weightings. This could potentially provide a new product with (near-) global coverage that is more accurate than either of the original gridded ET data sets, and would also allow independent verification against withheld tower observations.

R. Extrapolating the weights is certainly required to produce a global merger. It was already mention in the Conclusions, and it is something we have already worked on. Although we do not think that this specific merge was distinctively better than the simple average of GLEAM and PT-JPL, we already did some tests with the current weights in preparation for future merging efforts. So far we have used a multilayer-perceptron driven with a selection of the ET model inputs and trained to reproduce the current 84 station weights. This first setup seemed to show some skills to extrapolate the weights globally, with the predicted weights averaged over all stations correlating 0.7 with the original weights.

A figure showing preliminary globally seasonally averaged weights is displayed below, followed by a second figure showing seasonally averaged ET differences between the global average of GLEAM and PTJPL and the global product derived from applying the globally extrapolated weights, normalized by the seasonal average ET. Some clear seasonal and geographical patterns are observed in the extrapolated global weights and ET differences, which, as the reviewer indicated, would need to be assessed by confronting the global merge product with independent estimates.

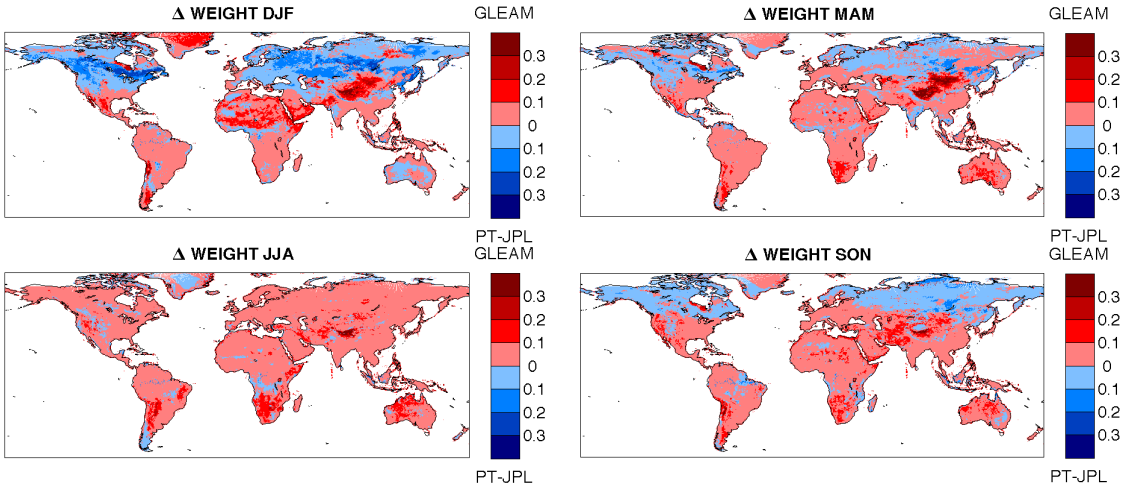

Figure A. Global seasonally averaged weights. GLEAM weighted more than PT-JPL is indicated in red. PT-JPL weighted more than GLEAM is indicated in blue. The seasonal weight is the value shown in the colour bar plus 0.5.

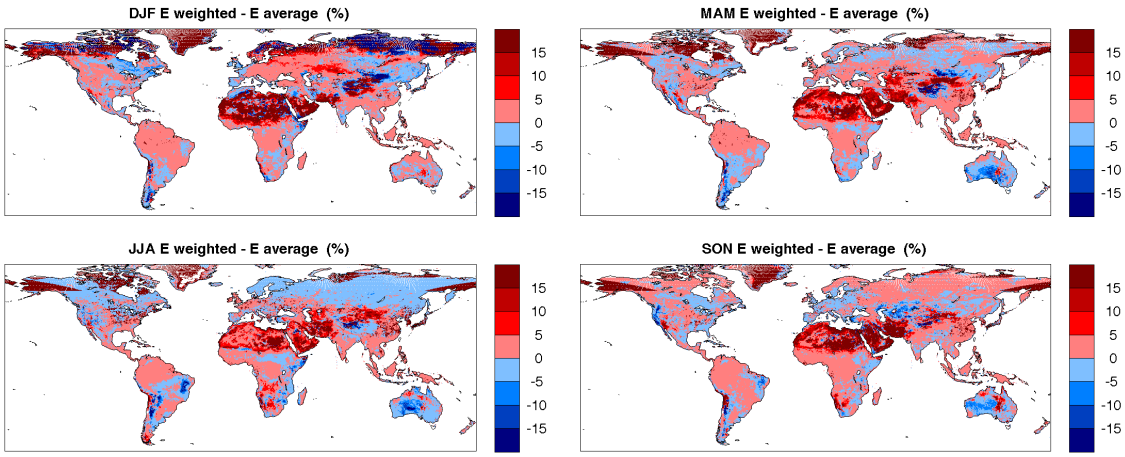

Figure B. Seasonally averaged E differences between the simple-average and the weighted product normalized by the seasonal average E.

If this is not possible, I suggest that the manuscript be re-submitted and re-written (with additional discussion and conclusions) to focus on evaluating the GLEAM and PT-JPL products against tower obs.

R. We have already published GLEAM and PT-JPL evaluations against tower observations and independent ET products, both at the tower and global scale, with the model using a variety of forcing datasets (Michel et al., 2016, Miralles et al., 2016, McCabbe et al., 2016). For us, this is a step further in a quest for a global merger of these types of ET methodologies, and we firmly believe that reporting these our first findings will be well received by interested colleagues.

Our next attempt to provide a more successful merge product will be based on using GLEAM and PT-JPL run with less common forcings, and for more extended periods. It is likely that the more independent model forcings will result in a more diverse performance at the towers. This, together with a larger pool of tower stations from the more extended period, could make the tower fluxes more informative, contributing more to the merged product. We will also be looking at adding ET estimates from the other two models run by WACMOS-ET, such as a newer global version of SEBS at 5km and daily resolution currently available from the SEBS developer team, or the new MODIS ET product version 6. We hope that a more extended dataset of tower locations and the more diverse weights of the future merge will be working on will also help to improve the global extrapolation of the weights. This will be reported in our next paper, together with independent evaluations of the merged product by comparing with other ET products and with inferred E derived from long time averages of basin precipitation and river runoff, as we have done in Miralles et al., 2016.

To address the reviewer concerns, we will make the following changes to the original manuscript:

(1) As suggested by another reviewer, we will change the title to make clear that the paper is more about the process of merging the products rather than about providing a successful merge product. We are considering different options, e.g., "Exploring the merging of two global land evaporation datasets based on local measurements".

(2) We will add to the literature review the statistical approaches that regress directly the tower ET on explanatory variables: "These efforts range from simply averaging a number of ET products (Mueller et al., 2003) to more complex approaches, such as fusion algorithms where the original ET products are combined to reproduce flux observations (Yao et al., :2017), or integration methodologies that seek consistency between ET products and related products

of the water cycle (Aires, 2014, Munier et al., 2014). Notice that there are also ET products based on a direct regression of tower ET on a set of explanatory variables (Jung et al., 2009, Wang et al., 2010)".

(3) We will focus more the paper objectives by adding: "Nevertheless, substantial differences were found between both model products. As such, we pose the question: can a product combining the GLEAM and PT-JPL estimates result in a more accurate ET estimate? We start here by investigating a weighted combination of GLEAM and PT-JPL estimates at some selected locations. Ideally, the weight assigned to each product during their merging should be based on an accurate description of the specific product uncertainties. However, even if some attempts to derive model uncertainty exist (Miralles et al., 2011a; Badgley et al., 2015; Loew et al., 2016), the complexity to derive estimates of ET from remote sensing data means that reliable quality assessment is only attained through validation against tower flux measurements. Therefore, here we propose a flux tower-based weighting of GLEAM and PT-JPL and an investigation of the performance of the resulting merger over a selection of tower sites".

(4) At the end of the section describing the merging technique we will add: "It should be noticed that the merging technique described would also need weights derivation outside the tower locations for the merge product to be useful. This will require an extrapolation technique, which can potentially be based on a regression of the weights on some explanatory variables. Likely candidates for the latter can be the inputs used to drive the ET models, assuming that relationships between the weights and the model inputs exist. Here we limit our study to investigate the performance of the merging technique at the tower sites, but this needs to be addressed in forthcoming efforts to provide merged estimates outside the location where the weights are derived.

MINOR COMMENTS:

Section 2: There is not enough information here for the reader to understand how the two products are calculated and what their main differences are. Please provide full details of the methodology of each product, rather than relying on previous work.

R. This is the third paper of the WACMOS-ET project, the first two ones also published in this journal. The GLEAM and PT-JPL models were described in more detailed, including their main equations, in the first paper, while for the second and this third one we only describe the main characteristics of the models. We are certain that any reader interested in this work would need to glance through the previous papers to follow this one, so we are not sure that fully describing the models here will be that useful. The same applies to the model forcings, which we described in detail in the first paper, and that we only summarized in the second and this third paper. We will consult with the editor about this, as we already had plagiarism complains precisely by mentioning again in this paper project elements already described in the first papers.

P5, L24: give the specific resolutions.

R. We will rephrase as: "Notice that the WACMOS-ET runs were done at 3-hourly and daily time resolutions, while only daily estimates are calculated for this study".

P8, L5: mention that the station coverage is not globally uniform, with nearly all stations in Europe and the US.

R. We will rephrase as: "This processing selects 84 stations for the 2002-2007 study period, with nearly all stations in Europe and US".

P8, L20: 'corrected fluxes are preferred". Provide citation. Also, for the results provided in this paragraph for the corrected fluxes, how were they corrected?

R. The citation is Foken et al., 2006, already provided. We will rephrase as:

"Techniques to correct this exist (Foken et al., 2006), and corrected fluxes applying these techniques are typically preferred over the original uncorrected observations."

Equation 1: add a sentence to describe what this metric is measuring (something like "the first term is the mismatch between the land cover at the tower and at the grid cell level, and the remaining terms are the net mismatch in land cover types across the two resolutions").

R. We will add as suggested: " …where the tower is situated. The first term is the mismatch between the land cover at the tower and at the grid cell level, and the remaining terms are the net mismatch in land cover types across the two resolutions. It takes the value …".

P14, paragraph from line 10: the text here implies that the motivation is to match the tower ET as closely as possible, but the tower ET will also include errors. This paragraph should be re-written to acknowledge that the tower ET will also include errors (and the methodology perhaps adjusted to not over-fit to the ET data)

R. We will improve this paragraph to make it clearer and add a note concerning the tower errors at the end: "When improving a product and comparing with a reference, the common targets are correlation unity and zero RMSD. Here, instead, we define a product that minimizes the RMSD with the tower ET, and refer to it, in the context of our merging strategy, the optimum product. At the days when the tower-measured ET is between the GLEAM and PT-JPL estimates, the optimum estimate equals the tower-measured ET. However, when both GLEAM and PT-JPL estimates are below or above the tower ET, the merge product will never be the tower-measured ET because it is bounded by the two model estimates. This is the main reason why correlation unity and zero RMSD can never be achieved here. For the optimum product, the closest model ET to the tower ET is the value that minimizes the RMSD with the tower ET, and this is the value selected for the optimum product in this case. For the 84 towers considered here this is the most common situation, happening on average over

all stations for around 80% of the days. In addition, as discussed in Section 2.2.2, the tower ET is also subject to errors, so the optimum product will inherit those errors at the instants where it takes the tower value. Therefore, there is not claim that the optimum product at those instants represents the unknown true ET, but the ET at the tower footprint as measured by the eddy-covariance instruments".

P15, L10. The use of the full seasonal cycle concerns me. In general, different ET products agree reasonably well in terms of the seasonal cycle (Jimenez et al. 2011; Mueller et al. 2011; Miralles et al. 2011). It is the anomalies that have more disagreement, and should then be the focus of efforts to improve / combine ET products. Also, using anomalies would be consistent with the assumption in the methodology that there are no biases. The reason given for not using anomalies is that there is insufficient tower data - if there really is insufficient data, this implies that ET cannot be trained on tower obs.

Jimenez, C., and Coauthors, 2011: Global intercomparison of 12 land surface heat flux estimates. Journal of Geophysical Research: Atmospheres, 116, D02 102, doi:10.1029/2010JD014545. Miralles, D., T. Holmes, R. de Jeu, J. Gash, A. Meesters, and A. Dolman, 2011: Global land- surface evaporation estimated from satellite-based observations. Hydrology and Earth System Sciences, 15, 453–469, doi:10.5194/hess- 15-453-2011. Mueller, B., and Coauthors, 2011: Evaluation of global observations- based evapotranspiration datasets and IPCC AR4 simulations. Geophysical Research Letters, 38, L06 402, doi:10.1029/ 2010GL046230.

R. The better agreement of ET products when the full seasonal cycle is considered is just the result of correlating two variables with marked lows and highs. In general, the more pronounced the seasonal cycle, the better the agreement in terms of correlation. At locations when the seasonal cycle is smaller, such as tropical forests, the agreement of the absolute ET values is much poorer in terms of correlation. This is not exclusive of ET estimates, it is also the case for other variable with strong seasonal cycles (e.g., radiation, temperature, precipitation).

Certainly, working with the anomalies would have been interesting, but this cannot be reliably achieved with our present data records. To work with

anomalies, a robust calculation of the seasonal cycle at the tower locations is needed. How many years would be acceptable? If we take the whole 1980-2015 FLUXNET2015 synthesis data set, and we conserve stations having at least 10 years of data, we are left with ~25% of the stations. If we take 5 years, which can be disputed as a sufficient number of years for a climatology, we still remove ~50% of the stations. As mentioned by the reviewer, the tower dataset is already severely limited in terms of geographical coverage, so such dramatic cuts in the number of stations is not very helpful for any merging methodology.

As mentioned in the paper, even if we do not work with anomalies, we try to show seasonal agreement between the different products, not just annual values, to reduce somehow the effect of the seasonal cycle in the analyses. Also, as the weights are calculated over a 61-day running window, there will be times of the year where the inter-seasonality within the 61-day running windows is small. At least for those times of the year there should not be much difference between the weights working in an absolute or anomaly space.

Nevertheless, independent of all this measures to mitigate the impacts of the seasonal cycle, we are certainly not working in an anomaly space. But for us, stating that "this implies that ET cannot be trained in tower observations" is not justified. It is still a challenge to reproduce the absolute ET values, as shown in the references given by the reviewer, or Figures 2 and 3 for this particular case, and as far as we can see most ET product merging efforts based on tower data work with absolute values.

P18, L18: EC is known to under-estimate the fluxes. Using the sum of LH and SH as the incoming energy will almost certainly give an underestimate.

R. True. We have been very clear about it in the same paragraph, stating the 6.1% underestimation when averaged over all the stations. We would argue than this underestimation has a smaller impact here, compared with other statistical approaches directly targeting the tower ET, such as the MTE product suggested later on by the reviewer. This is because we are not directly reproducing the tower ET, but weighting the original GLEAM and PT estimates. There can be an effect when deriving the error variance, as the relative differences of GLEAM and PT-JPL with the tower ET can change if corrected or uncorrected tower fluxes are used, but the merge product still remains bound by the original estimates.

Figure 5: what is causing the sudden changes in the time series? The 91 day windows used shouldn't suddenly change like this.

R. The reviewer is right, and continuous 61-day windows should not produce abrupt changes in this plot. However, at some locations the weights do not exist for all days. This happens at a few stations, as we impose that there should be at least 20 well spread daily values in the 61-day running window to derive the weights. For instance, in the right panel of Figure 5 the maximum values before the sudden decrease at day 180 correspond to the station CN-Dan. Due to observations quality and rain episodes there are not enough daily values to derive the weights for this station for the next few days, and the next maximum value comes from a new station with a lower value, producing the discontinuity.

We will be adding to the text describing the weight calculation: "A number of 30 days before and after each calendar day was found to provide a good compromise between the smoothness of weights and the number of samples required, so a 61-day running window was used to provide the daily weights. Notice that due to the masking of the tower data at very few occasions the 61 daily estimates are present in the running window, and 20 daily values reasonably spread in the running window are required as a minimum number of ET estimates to derive the weights. At most stations weight values exist for nearly all days, but at 8 stations there are larger gaps. The worst case is the tropical BR-Sa3 station, where the frequent rainy episodes complicate the derivation of the weights."

We will also replace the maximum and minimum curves in the Figure with the 5% and 95% percentiles, which are very close to the maximum and minimum values, but less sensitive to the discontinuities caused by the gaps in the time series of the weights at a few stations. The new figure can be found below.

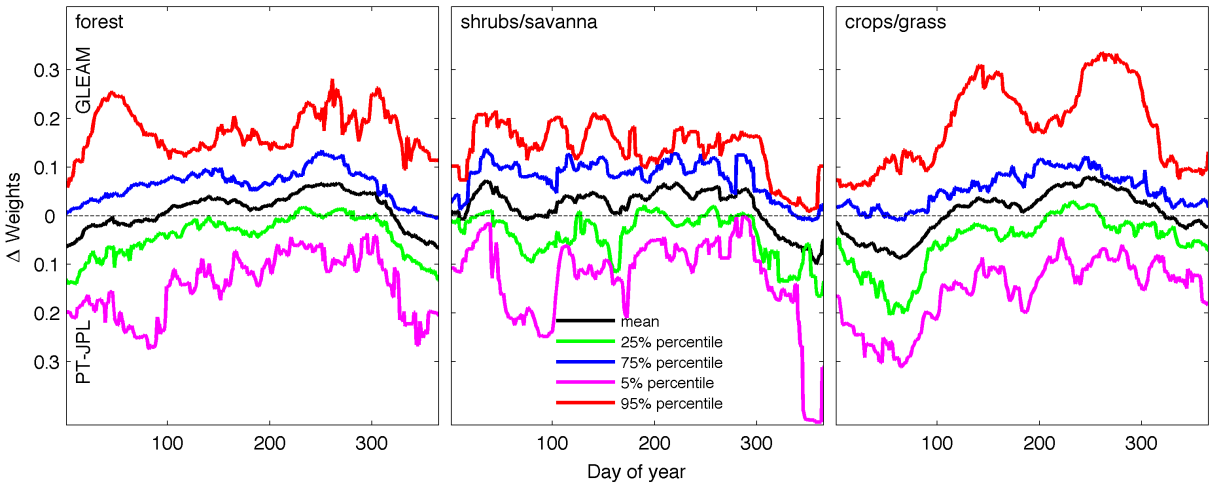

Figure 5. Daily statistics of weights over all forest (left), shrub/savanna (middle), and crop/grass (right) sites, plotted as the difference of the absolute weights and 0.5 (referred to as weight devia- tions, Δweight). A GLEAM Δweight of a means that GLEAM ET is weighted 0.5 + a, and PT-JPL ET 0.5 − a, while a PT-JPL Δweight of b means that GLEAM ET is weighted 0.5 − b, and PT-JPL ET 0.5 + b. Displayed are the mean, and the 5%, 25%, 75% and 95% percentiles.

Figure 7: This sudden increase in the tower ET in the upper panels look incorrect (and seem to occur at the same time each year - unless these are preceded by significant rain events, this don't look right). This time series needs to be checked, carefully QC- ed, and unusual features like this should be explained in the text.

R. As described in Section 2.2.2, the tower data was quality-controlled using the provided quality flags, and the represented fluxes were not marked as problematic. This site is a semi-arid savannah where vegetation development and associated fluxes are tightly linked to precipitation and humidity conditions. Station precipitation and soil moisture measurements in the 2004-2007 period can be found in Scott et al., 2009 (J. Geophys. Res., 114, G04004, doi:10.1029/2008JG000900), and match the general behaviour of the fluxes.

However, we plotted the ET estimates used for the derivation of the weights with the rainy episodes removed. This together with the running window used to smooth the lines produced the abrupt changes at the arrival of the summer rainfall, when many ET estimate are removed to derive the weights. To remove any confusion, we will be re-plotting the full time series, with the rainy days

marked also in the figure. We can have the merge product for all days, as the weights exist for all days, although to study their agreement with the tower ET we only consider the non-rainy days, as discussed in the article. The new figure can be found below.

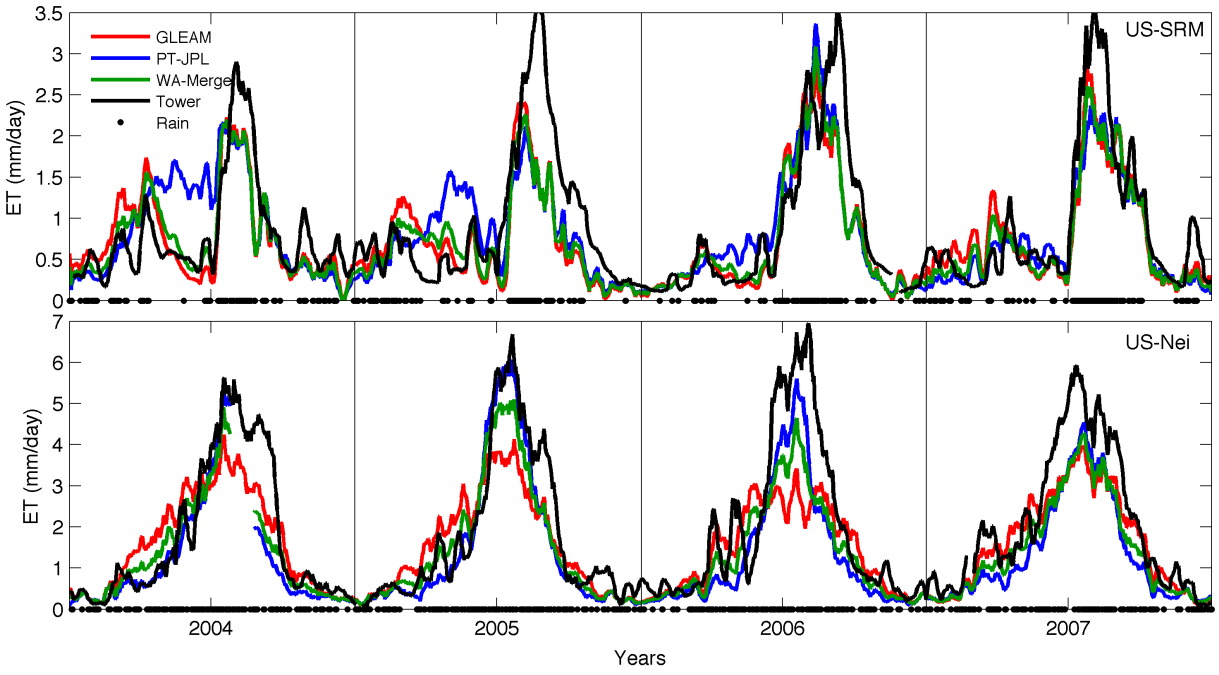

Figure 7. *ET time series at the sites US-SRM (top) and US-Nei (bottom). The daily values are time smoothed with a 10-day moving averaged window to better display the more persistent temporal features. Rainfall is marked with black circles in the x-axis, showing dry and rainy periods for US-SRM, and a more evenly distributed rain along the year at US-Nei.*

We will also modify the text to reflect these changes: "At the semi-arid grassland US-SRM site there are relatively large ET seasonal variability, with the ET tightly linked to the precipitation and associated increased in soil moisture (Scott et al., 2019). In terms of correlation, GLEAM and PT-JPL do not agree very well with the tower ET, with correlations of 0.31 and 0.24 for GLEAM and PT-JPL, respectively, calculated over the non-rainy days. Correlations are higher when calculated over all days, with values of 0.81 and 0.74 for GLEAM and PT-JPL, respectively, but as discussed in Section 2.2.2, we remove rainy episodes when analysing the data. GLEAM seems to capture better the spring ET decrease associated with the increase in temperatures and decrease in soil moisture, and both GLEAM and PT-JPL capture well the sudden increase in ET values at the beginning of summer related to the rainfall coming

from the North American monsoon. In this case, the larger weighting for GLEAM results in a weighted product that seems closer to the observed values, although the 0.37 correlation of the SA-merged product is not significantly higher than the 0.31 correlation of the WA-merged product".

The work would benefit from being placed within the context of other efforts to estimate ET with tower data / statistical methods. In particular the MTE product should be mentioned somewhere, as an example of using tower EC obs to estimate global ET.

Jung, M., M. Reichstein, and A. Bondeau, 2009: Towards global empirical upscal- ing of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. Biogeosciences, 6, 2001–2013, doi:10.5194/bg-6- 2001-2009.

R. *We are very familiar with the MTE product, and compared the WACMOS-ET estimates with the MTE product in e.g. Miralles et al., 2016. This product is now cited as discussed above.*