Answer to Reviewer 2 ESSD-2021-61

We thank Reviewer 2 for the comments. We provide here our responses to those comments and describe how we addressed them in the revised manuscript. The original reviewer comments are in normal black font while our answers appear in blue font. The corresponding edit in the manuscript are included in red font.

Lu et al., 2021 present a new merged global land evaporation dataset based on three existing products. While the paper is interesting for publication is this journal, my current recommendation is for a major revision for the following reasons. First, there is the issue of considering GLEAM as an observational dataset. GLEAM is really just another data source, and the authors even show that the individual models they are comparing outperform GLEAM in terms of R and RMSD compared to flux sites. I think the comparisons with GLEAM are okay, as long as authors specifically mention that it is not used for validation.

Response:

Thank you for your very thoughtful comment and suggestions. We have added the clarification of the role of GLEAM as an independent reference data set in the data section. GLEAM is a long sequence data set predominantly based on remote sensing observations, and on occasion, reanalysis data. GLEAM is unlike traditional land models, such as found in ERA5, MERRA2 and GLDAS, in that it is driven by satellite observations to obtain evaporation estimates. The version of GLEAM here relies very little on reanalysis datasets (only radiation and temperature of ERA-Interim). Therefore, GLEAM has the most independence relative to the model-based products.

We have added the description of the relative independence of GLEAM in the revised manuscript. The text there reads as: "*In addition, GLEAM is a long sequence data set predominantly based on remote sensing observations, and on occasion, reanalysis data. GLEAM is unlike traditional land models, such as found in ERA5, MERRA2 and GLDAS, in that it is driven by satellite observations to obtain evaporation estimates. The version of GLEAM here relies very little on reanalysis datasets (only radiation and temperature of ERA-Interim). Therefore, GLEAM has the most independence relative to the model-based products, which is selected as the reference data due to its relative independence.*".

GLEAM was included in preliminary evaluations of the ET estimates. By including GLEAM in the evaluations, we aim to assess regions where the reference data will be potentially less reliable. This also provides the information on regions of high uncertainties with respect to the reference data, which becomes very useful with applications. Thus, we obtained a good understanding of the skill of GLEAM prior to

its use as the reference data. In addition, an aim of the study is to leverage the uniqueness of GLEAM (as discussed above) to combine the model-based products. It is expected that GLEAM's over-reliance on observations states would serve as some sort of benchmark to estimate the weights of the model-based products. Thus, the goal is not based on a superior skill of GLEAM but its added value due to its uniqueness relative to the model-based products, which we believe, does have merits.

My other major concerns are some inconsistencies I see in figures, that need to be reevaluated or at least better explained (see more detailed comments below). One example in Figure 7, high NDVI values are incorrectly linked solely to humidity conditions, and the authors do not discuss the potential saturation issues at high NDVI values often seen in remote sensing datasets. The authors omit some necessary details (i.e. the use and source of NDVI is never explained in methods). Lastly, there are some grammar errors which should be addressed.

Response:

Thank you very much for your comment. We have made more specific explanations of the corresponding problems. Thereinto, we have corrected the one-sided expression about high NDVI values in the revised manuscript. The text there reads as: "*As shown in Fig. 7, the quality of each data set is relatively low and shows a rapid decline with the increase of vegetation density when NDVI is greater than 0.7, the case of optimal conditions for vegetation growth.*".

We have added the comments on the potential issue of vegetation index saturation at high amounts of NDVI to the manuscript. Vegetation index saturation at high amounts of NDVI poses potential issues. Generally speaking, NDVI is likely to become saturated over a dense canopy for forested areas, and becomes saturated rapidly for vegetation with a nearly closed canopy (Liu et al., 2011). Based on the analysis of hyperspectral data, it is found that there is an obvious saturation problem in the relationship between LAI and NDVI, that is, when LAI exceeds 2, NDVI asymptotically reaches the saturation level (Haboudane et al., 2004). When biomass reaches a certain level, NDVI is not sensitive to changes in biomass (Huang et al., 2021). Dynamic vegetation is not used in these models, resulting in lower data quality with dense vegetation. Therefore, vegetation index saturation at high amounts of NDVI results in a decrease in the quality of these datasets at high vegetation density. As shown in Fig. 7, the quality of each data set is relatively low and shows a rapid decline with the increase of vegetation density when NDVI is greater than 0.7, the case of optimal conditions for vegetation growth. In addition, a lot of remote sensing data have been used in GLEAM, such as satellite soil moisture, which is not of high quality when the vegetation density is high, affecting the quality of the final

output. Further, errors in GLEAM will affect the merged product because GLEAM acts as the reference data. The text in the revised manuscript reads as: "*It is well known that vegetation index saturation poses potential issues. Generally speaking, NDVI is likely to become saturated over a dense canopy for forested areas, and becomes saturated rapidly for vegetation with a nearly closed canopy (Liu et al., 2011). Based on the analysis of hyperspectral data, it is found that there is an obvious saturation problem in the relationship between LAI and NDVI, that is, when LAI exceeds 2, NDVI asymptotically reaches the saturation level (Haboudane et al., 2004). When biomass reaches a certain level, NDVI is not sensitive to changes in biomass (Huang et al., 2021). Dynamic vegetation is not used in these models, resulting in lower data quality with dense vegetation. Therefore, vegetation index saturation at high amounts of NDVI results in a decrease in the quality of these datasets at high vegetation density. As shown in Fig. 7, the quality of each data set is relatively low and shows a rapid decline with the increase of vegetation density when NDVI is greater than 0.7, the case of optimal conditions for vegetation growth. In addition, a lot of remote sensing data have been used in GLEAM, such as satellite soil moisture, which is not of high quality when the vegetation density is high, affecting the quality of the final output. Further, errors in GLEAM will affect the merged product because GLEAM acts as the reference data.*".

Monthly GIMMS NDVI3g data with a spatial resolution of 0.25° was used for the analysis and we have added the description of the product in the revised manuscript. The text there reads as: "*Monthly GIMMS NDVI3g data with a spatial resolution of 0.25° from the Global Inventory Modeling and Mapping Studies (GIMMS) was used in our study (Pinzon & Tucker 2014), with the time span from 1982 to 2014, which is available from http://ecocast.arc.nasa.gov/data/pub/gimms/3g/.*".

We have corrected the grammar errors.

Reference:

Liu, Y. Y., de Jeu, R. A. M., McCabe, M. F., Evans, J. P., and van Dijk, A. I. J. M.: Global long-term passive microwave satellite-based retrievals of vegetation optical depth, Geophys. Res. Lett., 38, L18402, https://doi.org/10.1029/2011GL048684, 2011.

Haboudanea, D., Miller, J. R., Pattey, E., Zarco-Tejada, P. J., and Strachan, I. B.: Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture, Remote Sens. Environ., 90, 337–352, https://doi.org/10.1016/j.rse.2003.12.013, 2004.

Huang, S., Tang, L., Hupy, J. P., Wang, Y., and Shao, G.: A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing, J. For. Res., 32, 1–6, https://doi.org/10.1007/s11676-020-01155-1, 2021.

Line 36: Since ‘the’ land surface

Response:

Thank you for pointing this out. We have corrected it.

Line 37: resulted should be resulting

Response:

Thank you for pointing this out. We have corrected it.

Line 46: Should say flux tower data

Response:

Thank you for pointing this out. We have corrected it.

Line 74: Pixel should not be capitalized

Response:

Thank you for pointing this out. We have corrected it.

Line 82: Lately for Lastly

Response:

Thank you for pointing this out. We have corrected it.

Line 88-90: What is the reference for this?

Response:

Thank you for that comment. We have added the reference in the revised manuscript. The text there reads as: “*Compared with the simple average method, Reliability Ensemble Average method (REA) extracts the most reliable information from each model by minimizing the impact of "outliers" or underperforming models, subsequently reducing the uncertainty range in simulated changes, which also stands out in terms of computational efficiency (Giorgi & Mearns, 2002).*”

Table 1: If using GLEAMv3, should include the Martens et al., 2017 reference (https://doi.org/10.5194/gmd-10-1903-2017)

Response:

Thank you very much for your comment. We have added the reference in the revised manuscript. The text there reads as: "

Table 1. Summary of ET data sets involved in merging.

Name	*ET schemes/ land-surface schemes*	*Spatial resolution (degree)*	*Temporal resolution*	*Time span*	*Reference*
GLEAM3.2a	*Priestley-Taylor*	*0.25×0.25*	*daily*	*1980-2017*	*Miralles et al. (2011b) Martens et al. (2017)*
ERA5	*IFS*	*0.25×0.25*	*1-hour*	*1980-2017*	*Hersbach et al. (2020)*
MERRA2	*GEOS-5*	*0.625×0.5*	*daily*	*1980-2017*	*Gelaro et al. (2017)*
GLDAS2.0 & 2.1	*Noah*	*0.25×0.25*	*3-hour*	*1980-1999& 2000-2017*	*Sheffield & Wood (2007)*

".

Reference:

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, Diego., Beck, H. E., Dorigo, W. A., Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geosci. Model Dev., 10, http://doi.org/1903–1925, 2017.

Line 119: Can you give some examples of the empirical parameters you mean? If not, might be best to remove.

Response:

Thank you for your cogent advice. There are some empirical parameters such as the evaporation stress factor (S), the latent heat of evaporation (λ) and the slope of the saturated water vapour-temperature curve (Δ).

Phenological constraints, heat stress or water availability affecting evaporation are usually combined in a empirical stress factor (Sellers et al., 2007). In GLEAM, an empirical parameter called evaporation stress factor (S) is defined, which ranges between 0 (maximum stress and no evaporation) and 1(no stress and potential evaporation).

In the Priestley and Taylor (1972) equation, the latent heat of evaporation (λ) and the slope of the saturated water vapour-temperature curve (Δ) are estimated from an empirical relationship with temperature (Henderson-Sellers, 1984; Maidment, 1993).

$$\lambda E_{\mathrm{p}} = \alpha \frac{\Delta}{\Delta + \psi}(R_{\mathrm{n}} - G)$$

We have added several specific empirical parameters in the revised manuscript. The text there reads as: "*The empirical parameters contained in this algorithm such as the evaporation stress factor, the latent heat of evaporation and the slope of the saturated water vapour-temperature curve have been obtained from the findings in different fields.*".

Section 2.1.2. Perhaps it's worth to mention that ERA-5 still appears to overestimate the latent heat flux, https://gmd.copernicus.org/articles/13/4159/2020/

Response:

Thank you very much for your comment. We have added it in the revised manuscript. The text there reads as: "*Martens et al. (2020) evaluated surface energy partitioning in ERA5 especially including the latent heat fluxes using different reference datasets and modeling tools, with the analysis showing that there is lower absolute biases in the surface latent heat flux of ERA5 than ERA-Interim, though ERA5 still appears to overestimate the latent heat flux in most catchments.*".

Line 168: three is spelt out and then indicated by numeric.

Response:

Thank you for pointing this out. We have corrected it.

Section 2.1.5: was any masking done to the dataset? I.e. removal of snowy, frozen or rainy days?

Response:

Thank you very much for your comment. We masked measurements using the provided quality flags. The removal of snowy, frozen and rainy days was not applied apart from the quality control applied. Generally, eddy-covariance measurements are less reliable during these days. As can be seen from the validation results, monthly scale validation results of all the datasets were better than the daily scale, indicating the reduced reliability has a greater impact on daily scale datasets. Nevertheless, the quality of the merged product is higher than other datasets at the daily scale. Therefore, the validation results are representative. We have added the description of the process of EC ET data preprocessing to section 2.1.5 in the revised manuscript. The text there reads as: "*Measurements are masked with the provided quality flags in the data set archives.*".

Figure 2&3: While I can understand Figure 2, I am not sure how to interpret it relative to Figure 3. It appears in Figure 2, that over some regions (i.e. the Amazon), the three datasets are in good agreement (CV close to 0). I would think this translates to evenly distributing the weight of each dataset on the merged product, but Figure 3 shows that MERRA2 is much less considered. Some patterns do make sense, for example in northern latitudes, it appears ERA5 is most closely related to the other datasets, despite their being greater CV, so it's more weighted.

Response:

Thank you for your very thoughtful comment. The CV analysis aims to evaluate the relative systematic differences between the three model products. As a consequence, relative deviations can be obtained. Since the CV is not computed relative to the reference data, GLEAM, it does not directly translate into the merging approach. This is why these differences between Figure 2a-c & Figure 2d mainly exist. While Figure 2a-c describe the products' contributions due to it a computed weight relative to the GLEAM, CV aims to understand the relative systematic differences apart from the reference. Nonetheless, what we attempt to achieve with the CV analysis is to identify the regions of significant differences between the products even apart from the reference. This serves as some sort of dual check for higher consistencies in the merging scheme.

We have added this explanation to the discussion in the revised manuscript. The text there reads as: "*The CV analysis aims to evaluate the relative systematic differences between the three model products. Since it does not take the reference data into account, it does not directly translate into the merging scheme. Nonetheless, it serves as an added check to obtain optimum consistencies in the merging process for higher skill in the merged data.*"

Figure 4:How can the authors explain the nearly symmetrical -50 to 50 mm/year the GLEAM model shows, versus the anomalies in the other products never going below 0?

Response:

Thank you for your cogent advice. The bandwidth of the kernel smoothing window was set to 10 to smooth the curve, making the GLEAM model show the nearly symmetrical -50 to 50 mm/year. The anomalies of all the five datasets were obtained by subtracting the climatology of GLEAM ET rather than their own from the original data to highlight the differences between them as a whole. However, we have found that it made more sense to subtract their own climatology than to subtract GLEAM's. Therefore, we have modified Figure 4c and added the bandwidth of the kernel smoothing window to the caption of figure in the revised manuscript.

The text there reads as: "

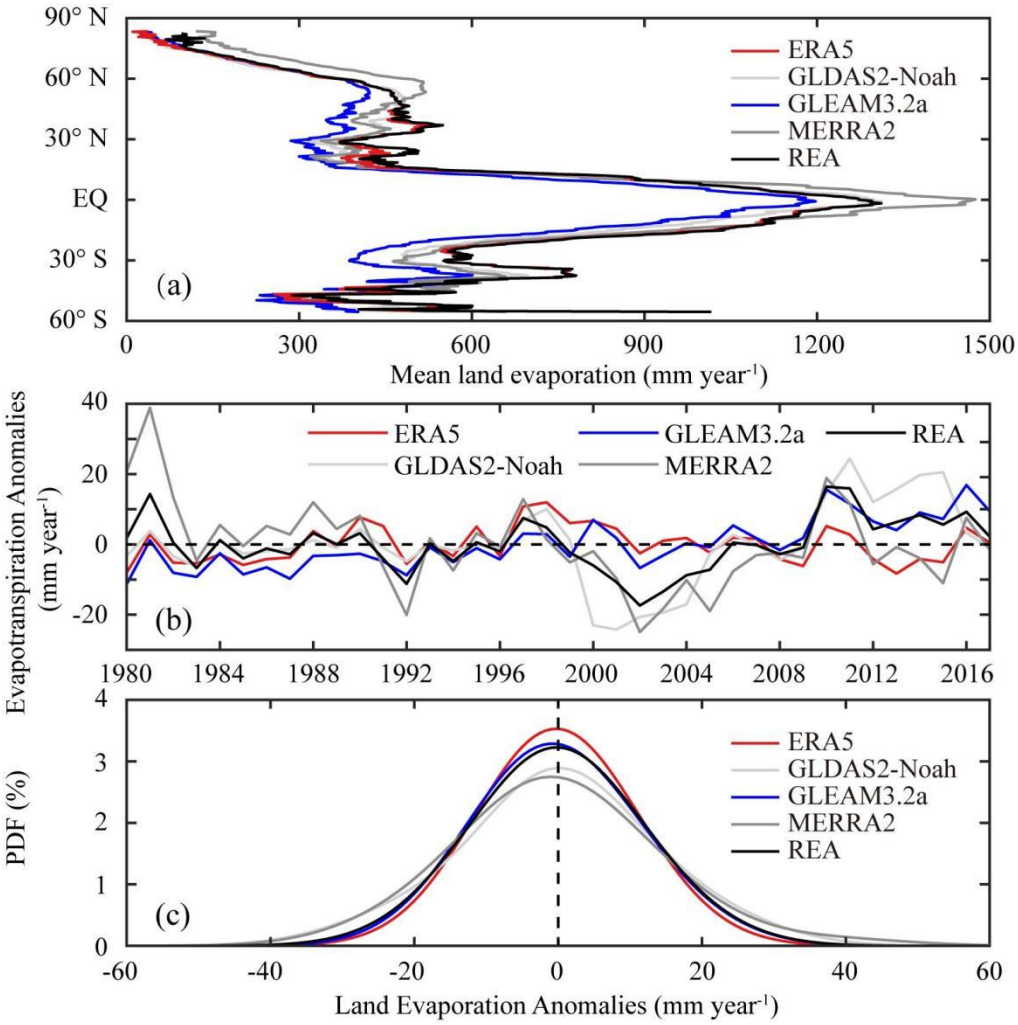

Figure 4. a) Latitudinal distribution of mean land evaporation from five data sets, b) time series (1980-2017), and c) probability distribution of annual land evaporation anomalies from five ET products. The bandwidth of the kernel smoothing window was

set to 10.".

We have modified the analysis of Figure 4c in the revised manuscript. The text there reads as: "*The obvious differences between the probability density distributions of multiple data sets are clearly visible. In general, the consistency between the merged product and GLEAM ET is relatively better, which may be greatly related to GLEAM as the reference data in the merging process. Due to the discrepancies in the driving data and calculation formulations for land evaporation, anomalies vary from data to data.*".

Figure 5: This figure highlights one point which is that GLEAM does not even perform better as some of the individual models for tower comparisons. If it is only used for comparisons, and not used as a validation source, that is still acceptable.

Response:

Thank you for that comment. GLEAM was included in preliminary evaluations of the ET estimates. By including GLEAM in the evaluations, we aim to assess regions where the reference data will be potentially less reliable. This also provides the information on regions of high uncertainties with respect to the reference data, which becomes very useful with applications. Thus, we obtained a good understanding of the skill of GLEAM prior to its use as the reference data.

Figure 7: The authors do not state where NDVI was obtained from and at what resolution.

Response:

Thank you very much for your comment. Monthly GIMMS NDVI3g data with a spatial resolution of 0.25° was used for the analysis and we have added the description of the product in the revised manuscript. The text there reads as: "*Monthly GIMMS NDVI3g data with a spatial resolution of 0.25° from the Global Inventory Modeling and Mapping Studies (GIMMS) was used in our study (Pinzon & Tucker 2014), with the time span from 1982 to 2014, which is available from http://ecocast.arc.nasa.gov/data/pub/gimms/3g/.*".

Line 303: NDVI >0.7 is not only under humid conditions, rather just optimal conditions for vegetation growth which widely vary depending on the ecosystem.

Response:

Thank you for your cogent advice. We have corrected the one-sided expression in the revised manuscript. The text there reads as: "*As shown in Fig. 7, the quality of each data set is relatively low and shows a rapid decline with the increase of vegetation density when NDVI is greater than 0.7, the case of optimal conditions for vegetation growth.*".

Figure 7: Can the authors comment on the potential issue of vegetation index saturation at high amounts of NDVI or LAI? Could that also explain some of these patterns?

Response:

Thank you for your very thoughtful comment. We have added the comments on the potential issue of vegetation index saturation at high amounts of NDVI to the manuscript. Vegetation index saturation at high amounts of NDVI poses potential issues. Generally speaking, NDVI is likely to become saturated over a dense canopy for forested areas, and becomes saturated rapidly for vegetation with a nearly closed canopy (Liu et al., 2011). Based on the analysis of hyperspectral data, it is found that there is an obvious saturation problem in the relationship between LAI and NDVI, that is, when LAI exceeds 2, NDVI asymptotically reaches the saturation level (Haboudane et al., 2004). When biomass reaches a certain level, NDVI is not sensitive to changes in biomass (Huang et al., 2021). Dynamic vegetation is not used in these models, resulting in lower data quality with dense vegetation. Therefore, vegetation index saturation at high amounts of NDVI results in a decrease in the quality of these datasets at high vegetation density. As shown in Fig. 7, the quality of each data set is relatively low and shows a rapid decline with the increase of vegetation density when NDVI is greater than 0.7, the case of optimal conditions for vegetation growth. In addition, a lot of remote sensing data have been used in GLEAM, such as satellite soil moisture, which is not of high quality when the vegetation density is high, affecting the quality of the final output. Further, errors in GLEAM will affect the merged product because GLEAM acts as the reference data. The text in the revised manuscript reads as: "*It is well known that vegetation index saturation poses potential issues. Generally speaking, NDVI is likely to become saturated over a dense canopy for forested areas, and becomes saturated rapidly for vegetation with a nearly closed canopy (Liu et al., 2011). Based on the analysis of hyperspectral data, it is found that there is an obvious saturation problem in the relationship between LAI and NDVI, that is, when LAI exceeds 2, NDVI asymptotically reaches the saturation level (Haboudane et al., 2004). When biomass reaches a certain level, NDVI is not sensitive to changes in biomass (Huang et al., 2021). Dynamic vegetation is not used in these models, resulting in lower data quality*

with dense vegetation. Therefore, vegetation index saturation at high amounts of NDVI results in a decrease in the quality of these datasets at high vegetation density. As shown in Fig. 7, the quality of each data set is relatively low and shows a rapid decline with the increase of vegetation density when NDVI is greater than 0.7, the case of optimal conditions for vegetation growth. In addition, a lot of remote sensing data have been used in GLEAM, such as satellite soil moisture, which is not of high quality when the vegetation density is high, affecting the quality of the final output. Further, errors in GLEAM will affect the merged product because GLEAM acts as the reference data.".

Reference:

Liu, Y. Y., de Jeu, R. A. M., McCabe, M. F., Evans, J. P., and van Dijk, A. I. J. M.: Global long-term passive microwave satellite-based retrievals of vegetation optical depth, Geophys. Res. Lett., 38, L18402, https://doi.org/10.1029/2011GL048684, 2011.

Haboudanea, D., Miller, J. R., Pattey, E., Zarco-Tejada, P. J., and Strachan, I. B.: Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture, Remote Sens. Environ., 90, 337–352, https://doi.org/10.1016/j.rse.2003.12.013, 2004.

Huang, S., Tang, L., Hupy, J. P., Wang, Y., and Shao, G.: A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing, J. For. Res., 32, 1–6, https://doi.org/10.1007/s11676-020-01155-1, 2021.

Line 306: This is true, also it brings the question of what land cover classifications are assigned for each model. If some models for example are using MODIS IGBP versus another data source, this could be a huge reason for discrepancies.

Response:

Thank you very much for your comment. We have added this consideration in the revised manuscript. The text there reads as: "*There are unique advantages and limitations of the existing land ET data sets for specific land cover types, however, quite few are globally suitable for meteorology and hydrology. Specific land cover classifications are assigned for each model, leading to the use of land cover classification from different sources bringing about discrepancies in the estimation of land ET.*".

Line 322: GLEAM is not the only product from even this study which considers soil moisture estimates from satellites.

Response:

Thank you for that comment. Indeed, GLEAM is not the only one that contains soil moisture, however, GLEAM is the only product that uses satellite retrieved soil moisture to drive the model. The two reanalysis products, ERA5 and MERRA2, depend on atmospheric based observations from satellites and ground observations assimilated into their atmospheric models. GLDAS, on the other hand, is a result of a free model run forced with atmospheric observations from satellites and ground observations. We have deleted the incorrect description and its subsequent paragraph.

Figure 8: Why are there missing areas in REA which is not observed in the other datasets? Especially in Northern Africa and Asia?

Response:

Thank you very much for your comment. We used the coefficient of variation (CV) as the indicator to select the merging regions with high data consistency, and the regions with low consistency were excluded from the merging scope, including the north of North America, west of South America, desert regions of mid-latitude Africa and Asia. The CV analysis aims to evaluate the relative systematic differences between the three model products. As a consequence, relative deviations can be obtained. Nonetheless, what we attempt to achieve with the CV analysis is to identify the regions of significant differences between the products even apart from the reference. This serves as some sort of dual check for higher consistencies in the merging scheme.
We have added this explanation to the discussion in the revised manuscript. The text there reads as: "*The CV analysis aims to evaluate the relative systematic differences between the three model products. Since it does not take the reference data into account, it does not directly translate into the merging scheme. Nonetheless, it serves as an added check to obtain optimum consistencies in the merging process for higher skill in the merged data.*"

Line 405: 0.5 degree or 0.25 degree?

Response:

Thank you for pointing this out. We have changed "0.5 degree" to "0.25 degree" in the revised manuscript. The text there reads as: "*We merged three land ET data sets, ERA5, GLDAS and MERRA2, respectively using REA method to generate a set of long sequence global daily ET data with a spatial resolution of 0.25 degree and a time span of 38 years.*".