A DETAILED LIST OF RESPONSES

TO REVIEWER #2

Anonymous Reviewer #2 comments

General comments:

The work "Multi-spatial scale assessment and multi-dataset fusion of global terrestrial evapotranspiration (ET) datasets" presents a detailed comparison of 30 global-scale evapotranspiration datasets and uses Bayesian Model Averaging to create a new weighted ensemble dataset. The paper is logically structured and clear overall. The comparison of such a large sample of ET datasets and the evaluation and comparison at a range of scales alone are interesting, novel, and valuable. For the dataset to be useful, more methodological details are needed about the pre-processing methods used align all datasets to a consistent spatial and temporal basis beyond the descriptions provided in the supplementary material. In addition, more detail is needed regarding the robustness of the BMA approach to key assumptions, namely land cover change, land cover classification uncertainty at the resolutions presented, and BMA model validation. Many such questions would be far easier to review and provide feedback on if (annotated) code used to generate the dataset were provided. These additions as well as a correction of the 1 degree resolution dataset are recommended before publication in ESSD.

Response: We greatly appreciate your careful reading of the manuscript, insightful comments, and valuable suggestions. Your thoughtful review has enhanced our paper considerably. The manuscript has been revised thoroughly according to your comments and those of the individual reviewers, with our point-by-point responses detailed below.

Specific comments:

1. There is a problem with the 1 degree dataset starting at approximately timestep 262.

Response: First of all, we thank you for your comments. We apologize that due to our carelessness, an incorrect version of the 1 degree spatial resolution BMA-ET dataset was uploaded here, for which we have updated and corrected.

2. Figure 7: What resolution and spatial interpolation methods were used to fill in data gaps in producing Figure 7? Below is a detail mean annual derived from 0.5 degree data (see attached notebook for full size image, the 500px restriction on image attachments is rather limiting here).

Response: Thank you for your comments. We apologize that we did not explain this clearly in the paper. In performing the plotting Figure 6 (corresponding to Figure 7 in the original version), we assigned the grid points with no values to 0 mm and did not use any spatial interpolation to fill in the data gaps. We have corrected Figure 6.

3. Line 178: how was training/validation split done for evaluating BMA performance? Given the sparsity of flux sites, why wasn't cross-validation considered? How sensitive are model weights to the training sample?

Response: Many thanks for highlighting this point. In this study, we use 60% of station data for training and 40% of station data for validation. In order to verify the robustness of the results, we performed 2 additional sets of experiments, 70% of data for training and 30% of data for validation, 80% of data for training and 20% of data for validation, respectively. The results show that the data accuracy of the fusion product BMA-ET is not sensitive to FLUXNET station split. The accuracy evaluation results showed remarkable consistency across different training set proportions (60%, 70%, and 80%), with BMA-ET demonstrating correlation coefficients of 0.68, 0.67, and 0.65 respectively when compared to FLUXNET-ET (Figure R1). In addition, we also evaluate the accuracy of BMA-ET under different vegetation types. The results show that the accuracy of BMA-ET under each vegetation type is not sensitive to the split ratio of the training set (Figures R2–R4).

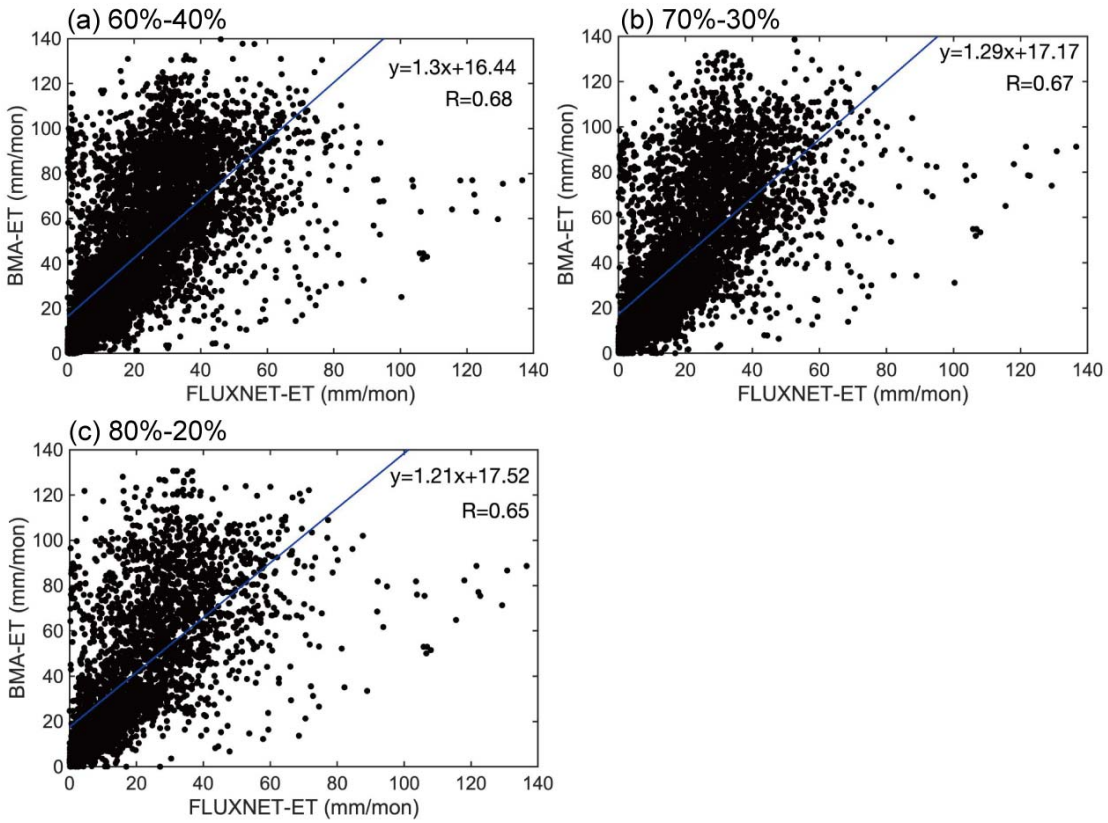

Figure R1 (Corresponding to Figure S29). Accuracy evaluation of BMA-ET. (a) 60% of data for training and 40% of data for validation, (b) 70% of data for training and 30% of data for validation, (c) 80% of data for training and 20% of data for validation.

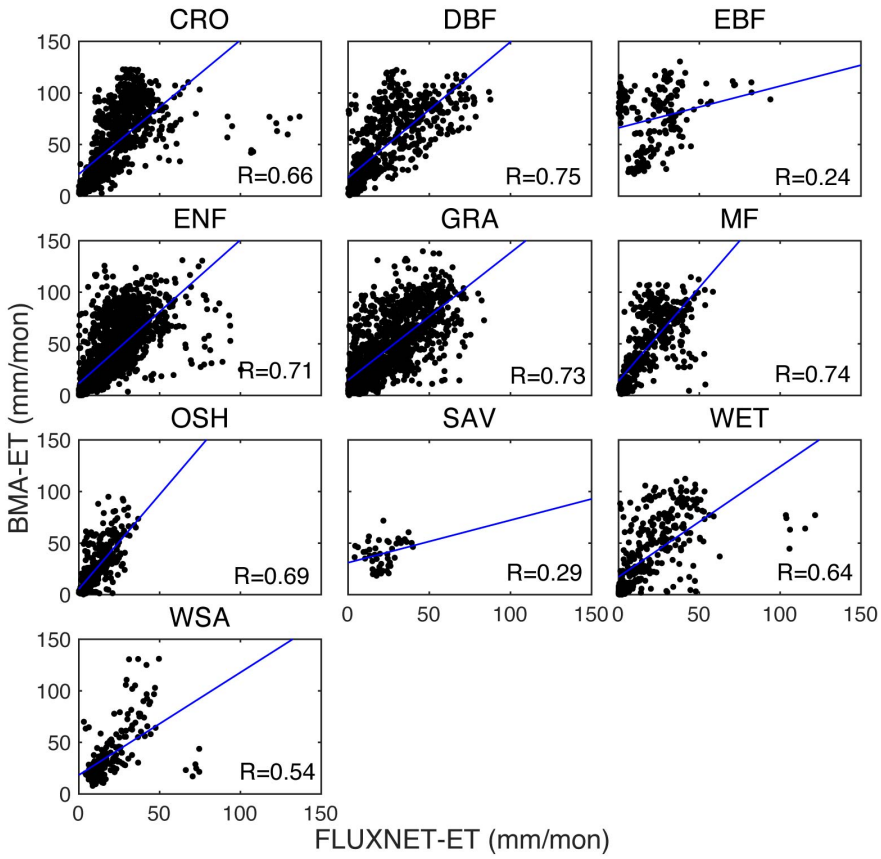

Figure R2 (Corresponding to Figure S30). Accuracy evaluation of BMA-ET over different vegetation types. 60% of data for training and 40% of data for validation.

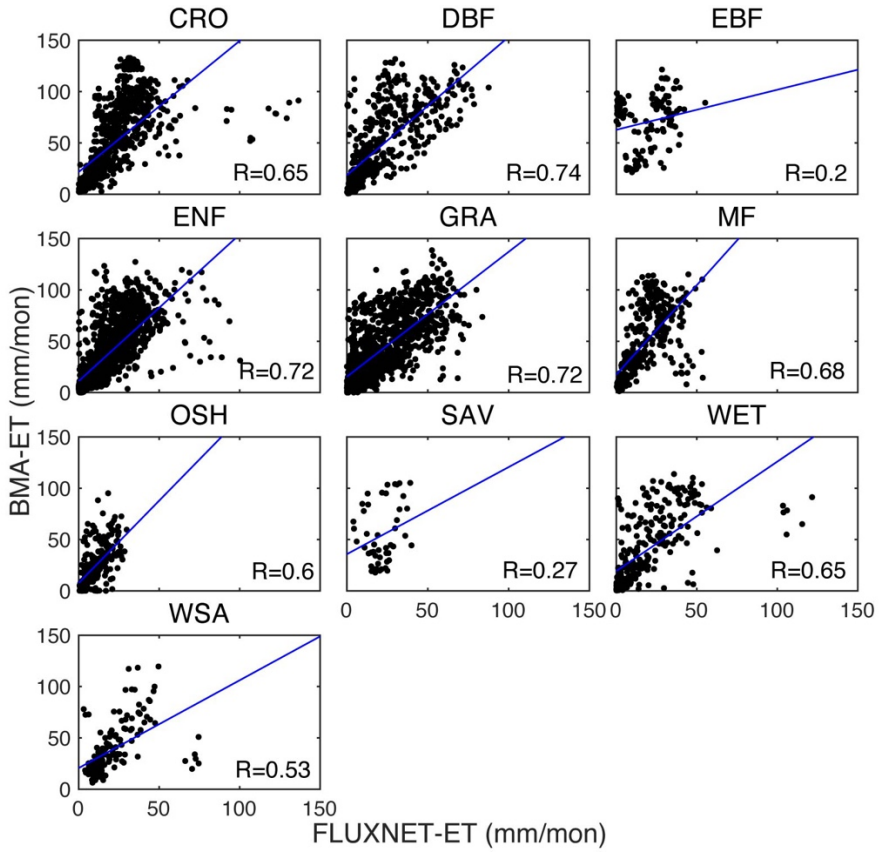

Figure R3 (Corresponding to Figure S31). Accuracy evaluation of BMA-ET over different vegetation types. 70% of data for training and 30% of data for validation.

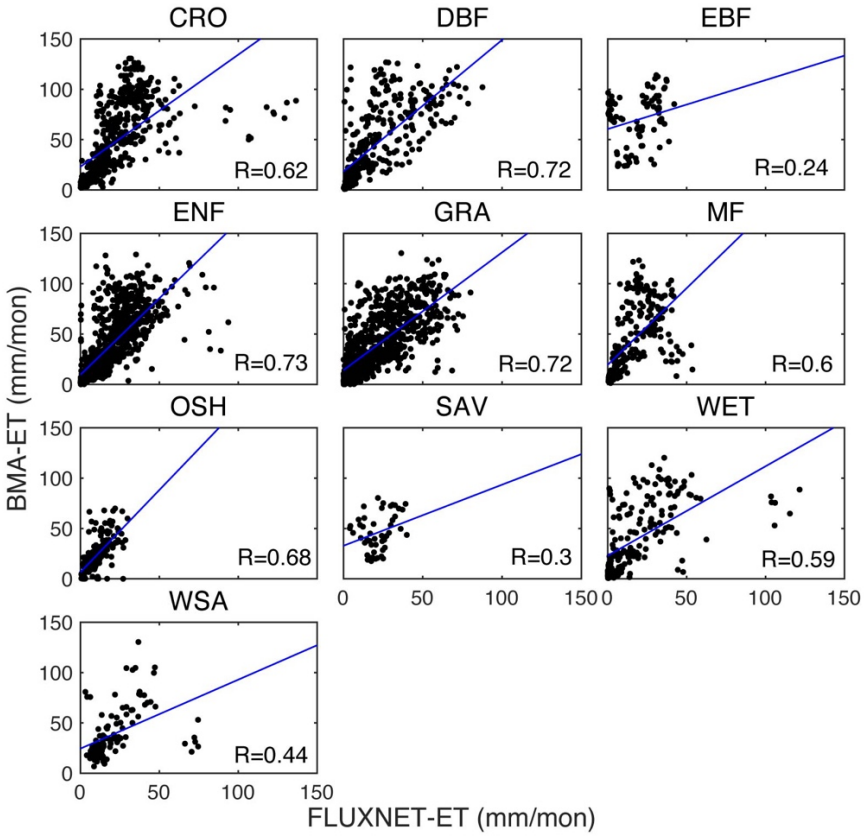

Figure R4 (Corresponding to Figure S32). Accuracy evaluation of BMA-ET over different vegetation types. 80% of data for training and 20% of data for validation.

We have made the following additions in section 4.1:

"In this study, we use 60% of station data for training and 40% of station data for validation. In order to verify the robustness of the results, we performed 2 additional sets of experiments, 70% of data for training and 30% of data for validation, 80% of data for training and 20% of data for validation, respectively. The results show that the data accuracy of the fusion product BMA-ET is not sensitive to FLUXNET2015 station split. The accuracy evaluation results showed remarkable consistency across different training set proportions (60%, 70%, and 80%), with BMA-ET demonstrating correlation coefficients of 0.68, 0.67, and 0.65 respectively when compared to FLUXNET2015 ET (Fig. S29). In addition, we also evaluate the accuracy of BMA-ET under different vegetation types. The results show that the accuracy of

BMA-ET under each vegetation type is not sensitive to the split ratio of the training set (Fig. S30, S31 and S32)."

4. Line 191: Can you quantify or estimate distributions of typical land cover changes at the appropriate dataset resolution as a basic test of model sensitivity to the stationary land cover assumption?

Response: We sincerely thank the reviewer for noting this. We performed a dynamic comparison of the data set MOD12Q1 for the three periods of 2001, 2010 and 2020 (Figure R5). We found that the consistency of the land cover types between the years was high, with the proportions of consistency between 2001 and 2010, 2010 and 2020, and 2001 and 2020 being 0.80,0.86 and 0.78, respectively (Figure R6). In addition, we also analyzed the percentage of global land area covered by 12 vegetation types (Figure R7). The results show that the proportion of area covered by various vegetation types does not vary much between years, especially for the four main vegetation types, OSH, WSA, SAV and GRA, which account for a larger proportion of the area.

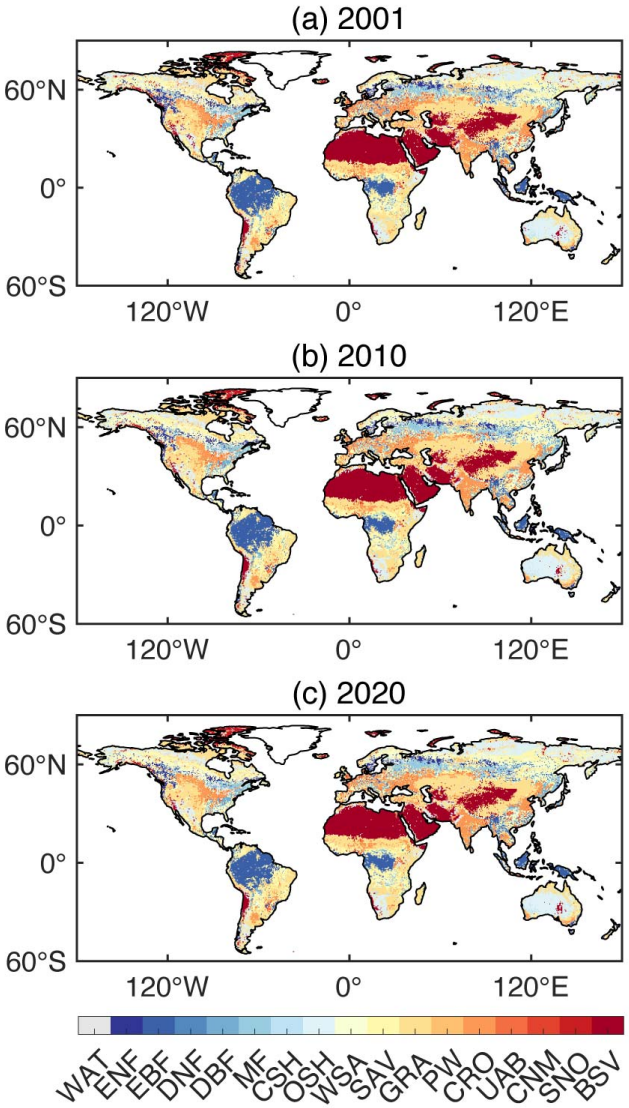

Figure R5 (Corresponding to Figure S33). Spatial distribution of land cover changes in three periods (2001, 2010, 2020) based on MOD12Q1.

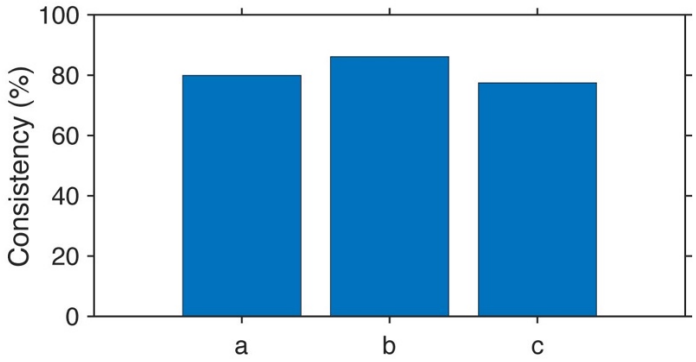

Figure R6 (Corresponding to Figure S34). Consistency between MOD12Q1 land cover types across years. Subfigures a-c show the level of consistency between land cover types for 2001 and 2010, 2010 and 2020, and 2001 and 2020, respectively. The

level of consistency is characterized by the ratio of the number of grid points with consistent vegetation types to the total number of grid points on land.

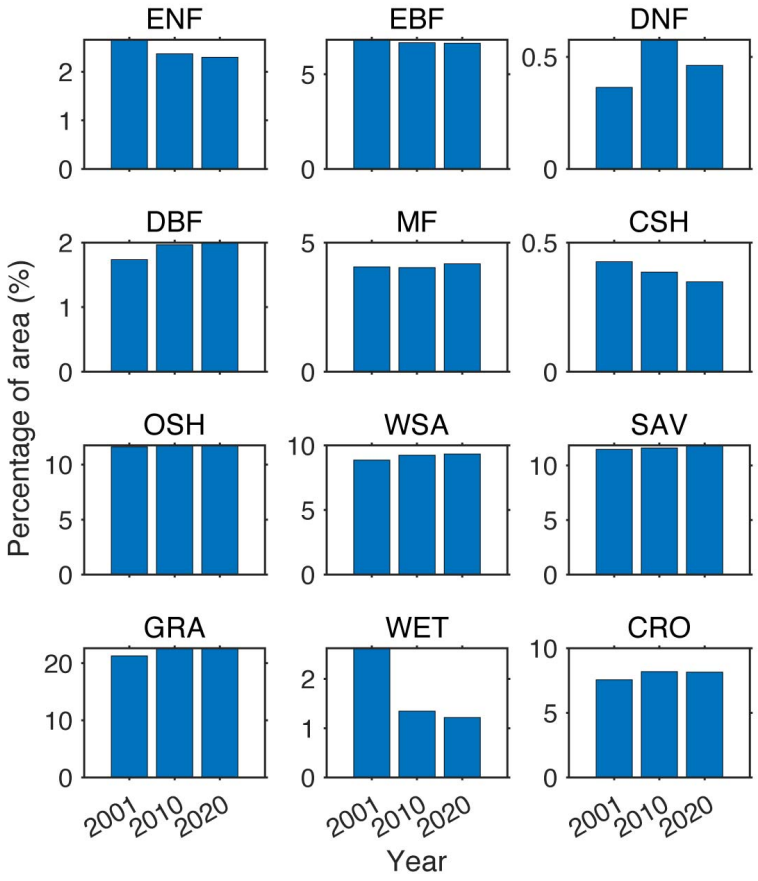

Figure R7 (Corresponding to Figure S35). Percentage of global land area covered by 12 vegetation types based on MOD12Q1. The bars in each subfigure represent the proportion of vegetation types in 2001, 2010 and 2020, respectively.

We have made the following additions in section 4.1.

"To validate model sensitivity to the stationary land cover assumption, we performed a dynamic comparison of the data set MOD12Q1 for the three periods of 2001, 2010 and 2020 (Fig. 33). We found that the consistency of the land cover types between the years was high, with the proportions of consistency between 2001 and 2010, 2010 and 2020, and 2001 and 2020 being 0.80, 0.86 and 0.78, respectively (Fig. 34). In addition, we also analyzed the percentage of global land area covered by 12 vegetation types (Fig. 35). The results show that the proportion of area covered by various vegetation

types does not vary much between years, especially for the four main vegetation types, OSH, WSA, SAV and GRA, which account for a larger proportion of the area."

5. Line 245: It isn't clear why correlations here are based on mean annual values and elsewhere (Figure S18) based on monthly data, making it more difficult to interpret the different comparisons presented (i.e. site, basin, global scales).

Response: Thank you for your comments. We apologize for not explaining this clearly in the paper. In Figure S18, each small circle represents a basin. This figure is not a comparison of ET datasets at any time scale, and it is a comparison of ET for all basins. Therefore, each small circle represents a multi-year average ET value for that basin. In this figure, even when replaced with monthly average ET values for each watershed, the comparison of multi-year average evapotranspiration from 30 ET datasets and multi-year average observed evapotranspiration from basin water balance are unchanged.

We have made the additions in the title of Figure S20.

"Fig. S20: Comparison of multi-year average evapotranspiration from 30 ET datasets and multi-year average observed evapotranspiration from basin water balance. Each small circle in the figure represents a basin."

6. Line 284 -- comparing typical MAE values with the stated trend, what is the uncertainty in the 0.21mm/yr trend line? How significant is the magnitude (and precision) of this trend compared to typical variability due to error?

Response: Thank you for your comments. Based on the fused dataset BMA-ET, we calculated the change trend of ET from 1980 to 2020. The global average ET change trend and its uncertainty is 0.65 (0.51–0.78) mm/yr (Figure R8). By calculating the signal-to-noise ratio (SNR) of the year-by-year ET data, it is found that the SNR is less than 1, i.e., the ET trend is smaller than the interannual variability, which indicates that the ET trend signal is weak or noisy. The SNR only reflects the ratio of the trend to the interannual noise, and cannot distinguish between a weak true climatic

signal (the trend is small) and a large noise in the data (e.g., fluctuations introduced by the observation error or the interpolation of the missing values). The Mann-Kendall test is a nonparametric test that does not require data distribution assumptions (Kendall, 1948; Mann, 1945); it is effective in detecting linear or nonlinear monotonic trends (continuously rising/declining) and is suitable for long-term climate change analyses; and it is insensitive to extreme values and robust. Therefore, we used the Mann-Kendall test to verify the significance of the trend. Despite the low SNR of the ET series, the Mann-Kendall test showed that the trend was significant (p<0.01), suggesting a persistent upward trend in the ET series.

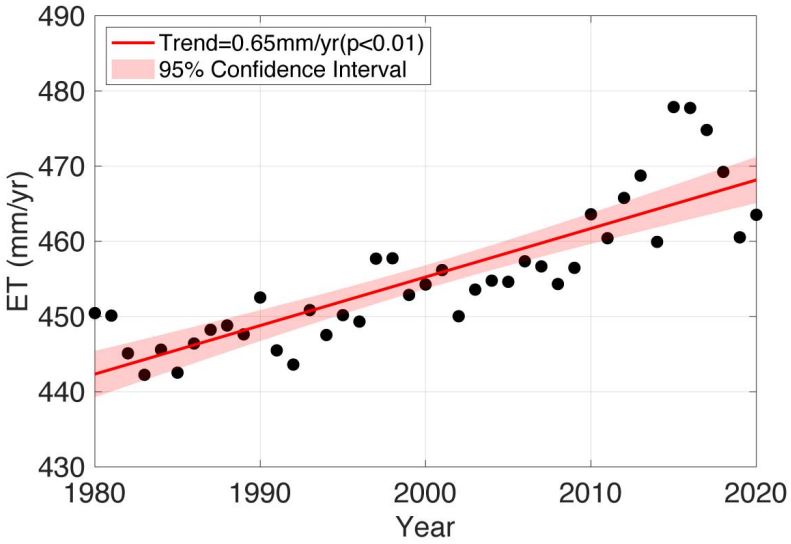

Figure R8 (Corresponding to Figure S25). Interannual variations of BMA-ET during 1980–2020. The global land average results are calculated based on a weighted average of the global land area.

7.A basic attempt to replicate Figure S23 was unsuccessful. There is likely a simple explanation for the substantial offset (~20mm) but it is much more laborious to investigate without the full replication code. A copy of the code used to generate the figures presented in this review is attached.

Response: I'm sorry we didn't make that clear. We used the area-weighted average method in calculating the global average ET. Data are available in a regular

longitude-latitude grid. Therefore, grid cells do not have an equal size, with smaller grid cells at higher latitudes. Hence, for calculating the land area fractions, we must assign a weight to each grid cell on the basis of size. Here we compute these weights (w_i) as the size of each grid cell at latitude lat_i relative to the size of the largest grid cells located at the Equator, given by:

$$W_i = \left|\frac{\sin(lat_i+(0.5\times res))-\sin(lat_i-(0.5\times res))}{\sin(0.5\times res)-\sin(-0.5\times res)}\right| \quad (1)$$

where lat is the vector indicating the latitude of each grid cell center, and ranges from $-90+(0.5\times res)$ to $90-(0.5\times res)$ with increasing step or resolution denoted by res.

We have made the additions in Figure S25.

8.Figure 8: What are the units these models are compared on? i.e. is standard deviation mm/year? Was some kind of normalization/standardization done to make the reference dataset standard deviation exactly 1?

Response: Thank you. The Taylor diagram was first proposed by Taylor and is mainly used to evaluate the ability of different models to simulate a variable. The plot combines three evaluation metrics: the correlation coefficient, the root mean square error, and the ratio of the standard deviations of the simulated and observed fields on a single polar plot. See Taylor for specific formulas. Standardized Taylor diagram normalize the standard deviation and root mean square error to eliminate their physical units of measure. When the correlation coefficient is larger, the root-mean-square error is smaller, and the ratio of the standard deviation of the simulated values to that of the observed values tends to be closer to 1, it indicates that the simulation results are in good agreement with the measured data, i.e., the model simulation results are highly reliable. In this study, a standardized Taylor diagram was used for the comprehensive assessment of the global ET dataset.

We have made the following additions in section 2.2.

"We also used standardized Taylor diagrams for comprehensive evaluation of ET datasets (Supplementary information Text S5)."

We have made the following additions in Text S5.

"Text S5. Taylor diagram

The Taylor diagram was first proposed by Taylor and is mainly used to evaluate the ability of different models to simulate a variable. The plot combines three evaluation metrics: the correlation coefficient, the root mean square error, and the ratio of the standard deviations of the simulated and observed fields on a single polar plot. See Taylor for specific formulas. Standardized Taylor diagram normalize the standard deviation and root mean square error to eliminate their physical units of measure. When the correlation coefficient is larger, the root-mean-square error is smaller, and the ratio of the standard deviation of the simulated values to that of the observed values tends to be closer to 1, it indicates that the simulation results are in good agreement with the measured data, i.e., the model simulation results are highly reliable. In this study, a standardized Taylor diagram was used for the comprehensive assessment of the global ET dataset."

We have made the following additions in Figure 7:

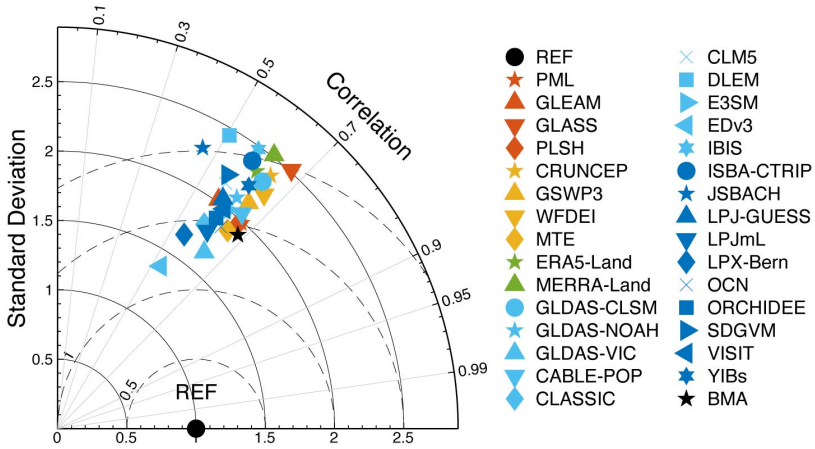

Figure 7: Standardized Taylor diagram of ET datasets at all stations from 1991 to 2011. The observation data are ET from FLUXNET2015.

9.Figure 8: What is the advantage of the BMA-ET dataset over the GLDAS-VIC dataset, or other datasets with similar correlation, lower RMSE, and standard deviation closer to the reference dataset?

Response: Thank you for your comments. In Figure R9 (Corresponding to Figure 7), BMA-ET has the highest correlation coefficient with FLUXNET2015. While the RMSE of BMA-ET vs. FLUXNET 2015 site ET was not the lowest, the ET dataset with the lower RMSE had a much lower correlation coefficient than BMA-ET, less than 0.6. In addition, the BMA-ET dataset has a longer time period of coverage than other datasets with similar correlations, lower RMSEs and standard deviations closer to the reference dataset. Therefore, BMA-ET is more suitable for long-term climate change studies.

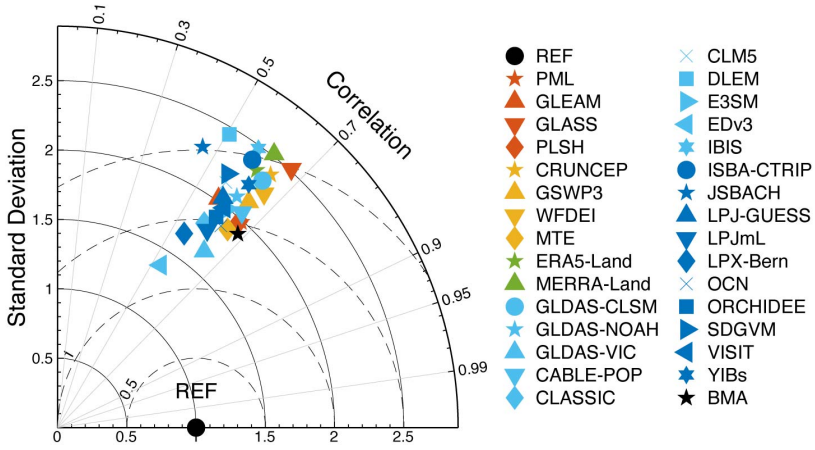

Figure R9 (Corresponding to Figure 7). Standardized Taylor diagram of ET datasets at all stations from 1991 to 2011. The observation data are ET from FLUXNET2015.

10.Line 313: What is the sensitivity of model performance to typical differences / uncertainties introduced by spatial scale mismatch?

Response: Since the original spatial resolutions of the 30 datasets are different, and this study uses a bilinear interpolation method to unify them to the same resolution, this resampling process introduces some uncertainty. Therefore, this study compared the differences in ET data at 1° and 0.5° globally and under each vegetation type, respectively. The results show that correlation coefficients for ET datasets at 1° and

0.5° spatial resolution globally are more than 0.9 among all ET datasets (Figure R10). For each vegetation type, correlation coefficients for ET datasets at 1° and 0.5° spatial resolution are over 0.8. This suggests that the uncertainty due to differences in spatial resolution is relatively small.

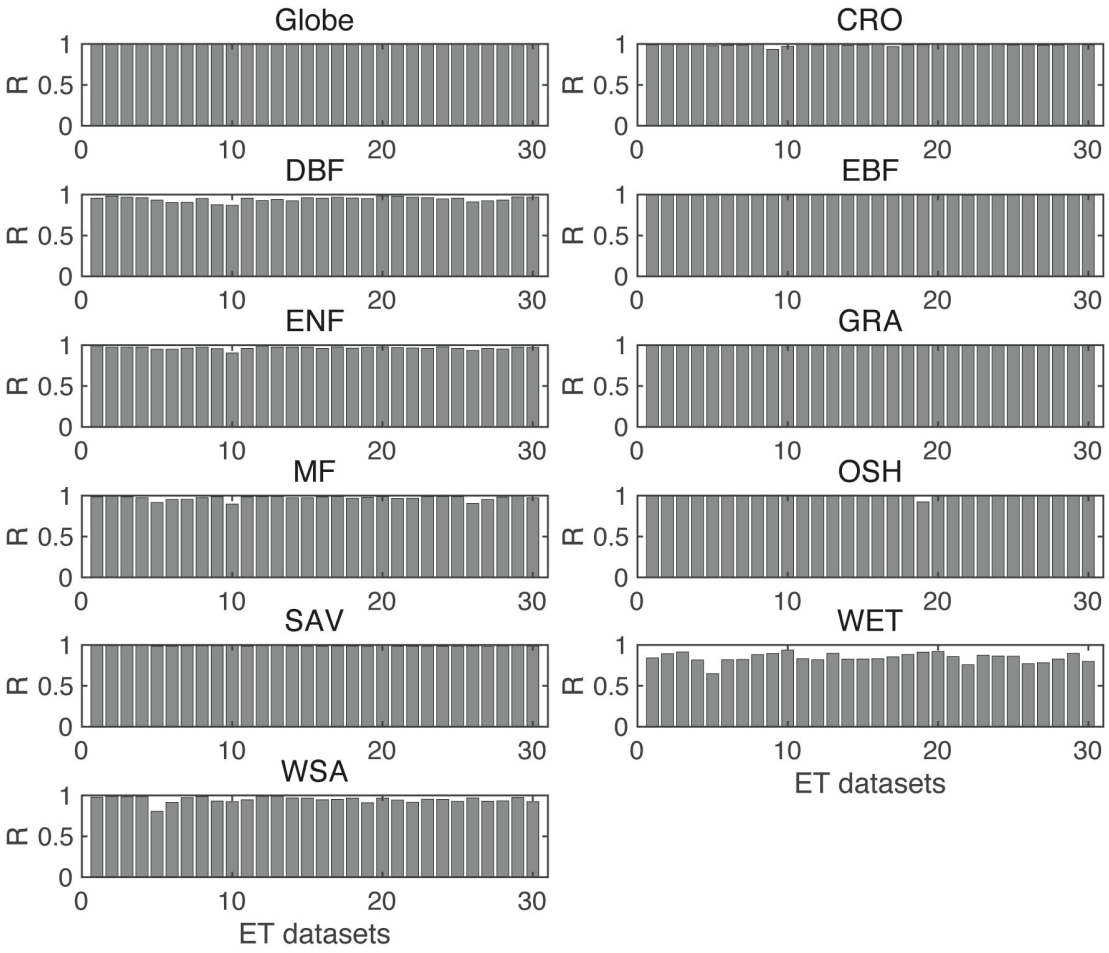

Figure R10 (Corresponding to Figure S4). Correlation coefficients for ET datasets at 1° and 0.5° spatial resolution globally and for each vegetation type.

We have made the following additions in section 2.2 and Figure S4.

"This study compared the differences in ET data at 1° and 0.5° globally and under each vegetation type, respectively. The results show that the sensitivity to spatial resolution is low for each ET dataset (Fig. S4)."

11.Section 4.2: More discussion of how data leakage was avoided is needed. How is training data independent of validation data in each comparison?

Response: We thank the reviewer for your valuable suggestion. To address the above issues, we completed the following two components.

(1) 30 datasets are clustered and then fused

First, we calculated a Pearson correlation coefficient matrix using the residuals from the 30 sets of ET datasets with observations from FLUXNET2015 sites. Second, the 30 sets of ET datasets were clustered based on the residual correlation coefficient matrix. Third, for each vegetation type, the BMA fusion of the ET data within each cluster was performed first, and then the BMA fusion of the fused data for each cluster was performed.

(2) Independent data source validation

We used independent data sources to validate the fusion dataset BMA-ET, specifically including AmeriFlux, ChinaFlux, and ICOS. The site information of AmeriFlux, ChinaFlux, and ICOS list in Tables S5–S7. Additionally, we also evaluated the accuracy of BMA-ET using FLUXNET2015 data from 2012 to 2015. The results demonstrate that BMA-ET outperforms other external datasets, achieving correlation coefficients of 0.61, 0.72, and 0.74 with site-level ET measurements from AmeriFlux, ChinaFlux, and ICOS, respectively (Figure R11). Using FLUXNET2015 as reference observations, BMA-ET showed a correlation coefficient of 0.58 with FLUXNET2015 site-level ET during 2012–2015 (Figure R11), while also demonstrating high accuracy across various vegetation types (Figure R12).

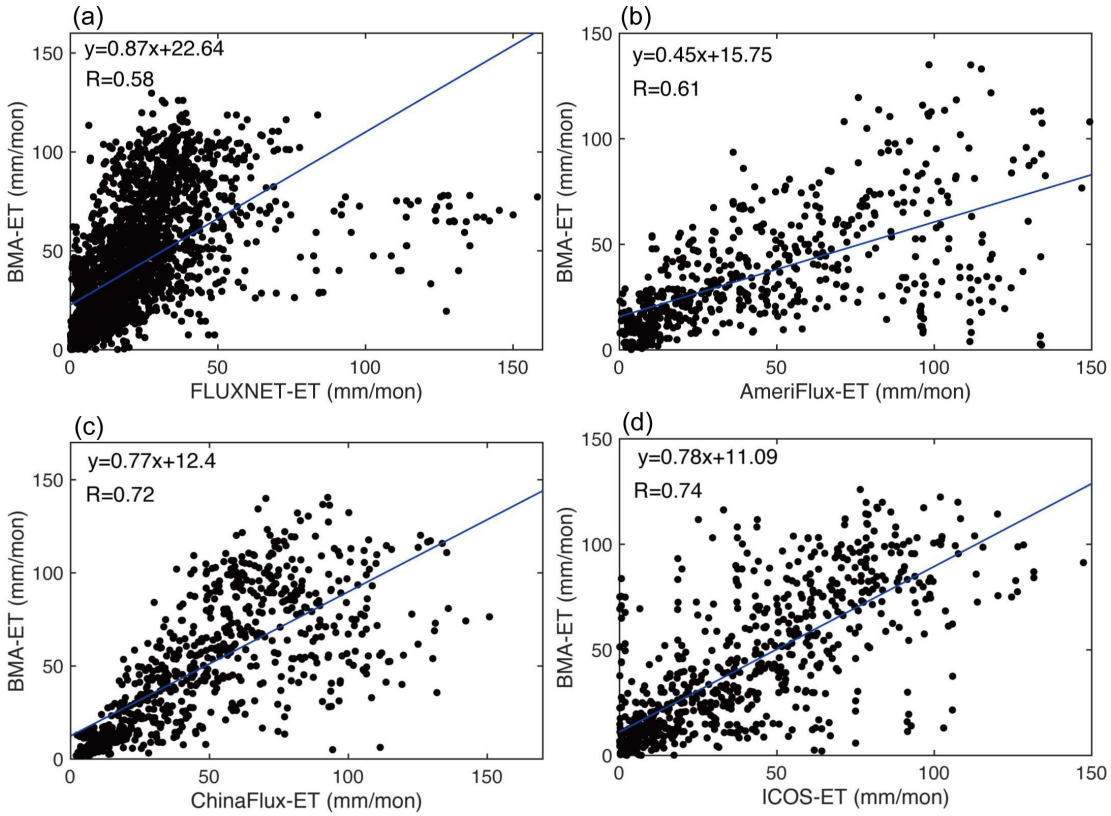

Figure R11 (Corresponding to Figure 8). Accuracy evaluation of BMA-ET. The observation data is ET from (a) FLUXNET2015 during the period 2012–2015, (b) AmeriFlux during the period 1994–2020, (c) ChinaFlux during the period 2003–2010, (d) ICOS during the period 2003–2010.

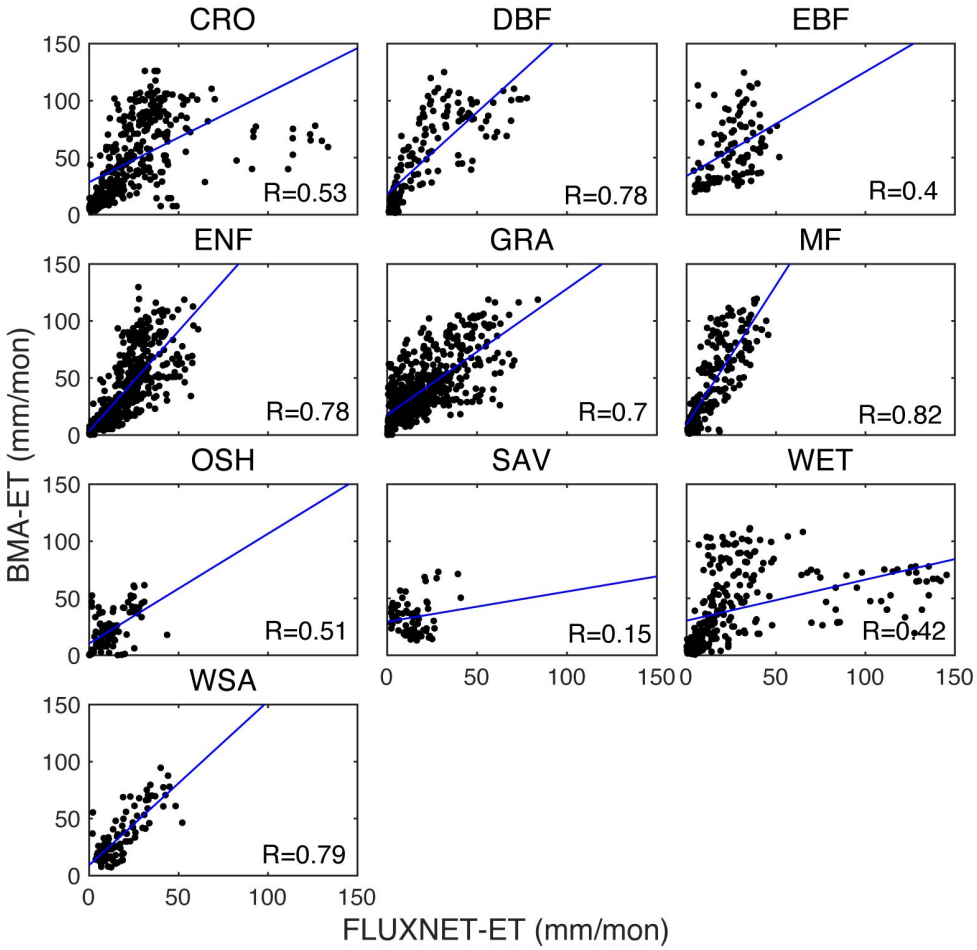

Figure R12 (Corresponding to Figure S28). Accuracy evaluation of BMA-ET under different vegetation types during the period 2012–2015. The observation data is ET from FLUXNET2015.

We have made the following additions in section 2.2:

"Finally, evapotranspiration fusion was completed using a BMA method. The 30 sets of ET datasets are clustered and then fused. First, we calculated a Pearson correlation coefficient matrix using the residuals from the 30 sets of ET datasets with observations from FLUXNET2015 sites. Second, the 30 sets of ET datasets were clustered based on the residual correlation coefficient matrix (Table S9). Third, for each vegetation type, the BMA fusion of the ET data within each cluster was performed first, and then the BMA fusion of the fused data for each cluster was performed."

We have made the following additions in section 3.2:

"We used independent data sources to validate the fusion dataset BMA-ET, specifically including AmeriFlux, ChinaFlux, and ICOS. Additionally, we also evaluated the accuracy of BMA-ET using FLUXNET2015 data from 2012 to 2015. The results demonstrate that BMA-ET outperforms other external datasets, achieving correlation coefficients of 0.61, 0.72, and 0.74 with site-level ET measurements from AmeriFlux, ChinaFlux, and ICOS, respectively (Fig. 8). Using FLUXNET2015 as reference observations, BMA-ET showed a correlation coefficient of 0.58 with FLUXNET2015 site-level ET during 2012–2015 (Fig. 8), while also demonstrating high accuracy across various vegetation types (Fig. S28)."

---------- end line----------

In order to make the review of our revision more convenient, we have marked all changes using the "Track Changes" function in Microsoft Word and have uploaded the "tracked changes" version as Supplementary Material.

References

Kendall, M. G.: Rank correlation methods, Griffin, Oxford, England, 1948.

Mann, H. B.: Nonparametric Tests Against Trend, Econometrica, 13, 245–259, https://doi.org/10.2307/1907187, 1945.