A DETAILED LIST OF RESPONSES

TO REVIEWER #1

Anonymous Reviewer #1 comments

General comments:

Reviewer #1: The research entitled "Multi-spatial Scale Assessment and Multi-dataset Fusion of Global Terrestrial Evapotranspiration Datasets" meticulously evaluated the accuracy and uncertainty inherent in thirty ET datasets at multiple spatial scales. These datasets encompass a variety of methodologies, including those derived from remote sensing‐based, machine learning‐based, reanalysis‐based, and land‐surface‐model‐based. Then the study produced a fusion ET dataset (BMA-ET) using BMA method and dynamic weighting scheme for different vegetation types. The article is well-written and demonstrates strong logical coherence. However, I am doubt about the purpose of this study. As the authors have pointed out, "there are large discrepancies among ET estimates from different methods", I am wondering how does the research handle the uncertainty between different types of ET datasets. Due to differences in algorithm frameworks and input data, the uncertainty of estimation results varies. The ET Fusion not only combines the advantages of different models, but also integrates uncertainty and even enhances errors. Regarding this, the author did not provide a solution. For a global ET dataset, data availability is more important than validation accuracy, and the results and novelty do not reach the desired level, which I do not think meet the requirements of ESSD. Thus, I recommend rejection. Please see my specific comments below.

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. I think the most significant problem with this research is that all the machine learning ET models and some other models (GLASS, PML, etc.) have been calibrated by ground observations from FLUXNET. The BMA-ET generated in this study used FLUXNET observations to fuse thirty ET datasets, which poses a problem of data reuse, and the estimated results may even overfit.

Response: We thank the reviewer for your valuable suggestion. Indeed, the use of FLUXNET2015 data for calibration or validation in some datasets has affected the results to some extent. 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 R1). Using FLUXNET2015 as reference observations, BMA-ET showed a correlation coefficient of 0.58 with FLUXNET2015 site-level ET during 2012–2015 (Figure R1), while also demonstrating high accuracy across various vegetation types (Figure R2).

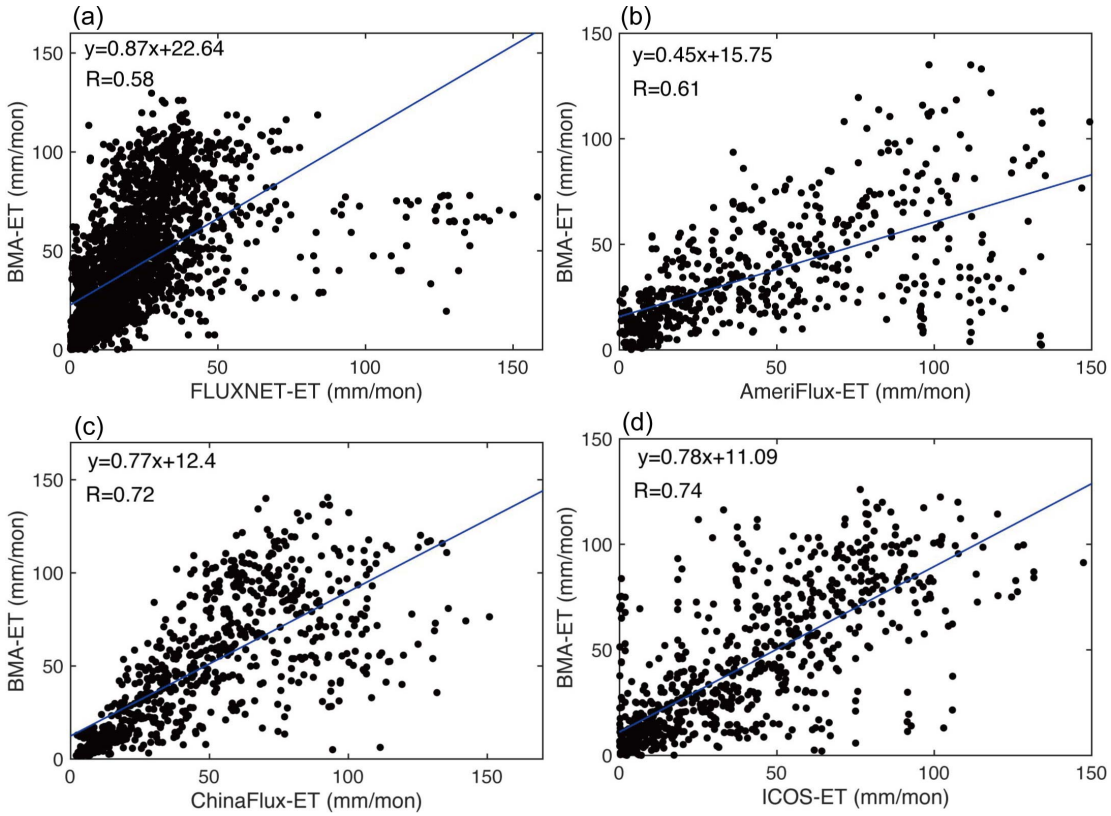

Figure R1 (Corresponding to Figure 8). Accuracy evaluation of BMA-ET. The observation data is ET from (a) FLUXNET2015 during the period 2012–2015, (b) AmeriFlux during the period 1994–2020, (c) ChinaFlux during the period 2003–2010, (d) ICOS during the period 2003–2010.

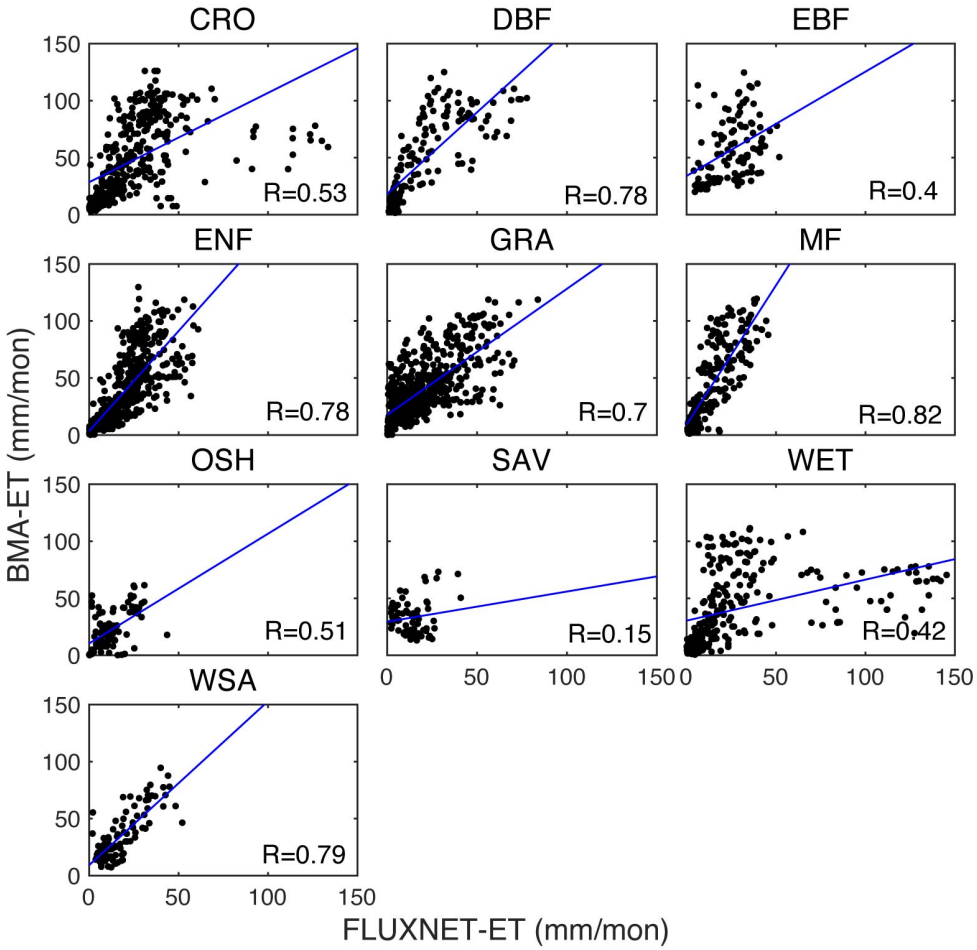

Figure R2 (Corresponding to Figure S28). Accuracy evaluation of BMA-ET under different vegetation types during the period 2012–2015. The observation data is ET from FLUXNET2015.

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)."

2. How did the authors handle the estimation accuracy of sparse areas such as South America and Africa during the fusion process?

Response: Thank you for this pertinent advice. In this study, ET data were fused not by continent, but also by vegetation type. It is important to acknowledge that the number of sites is small in some vegetation types, and because of the overall short time period covered by the flux sites, this study spliced the time series of all the sites in each vegetation type, resulting in a longer time-span data set of site observations to be used as observations for the BMA analysis. The relevant content has been described in detail in the section 2.2.

3. The BMA is not an advanced fusion algorithm. The GLASS v4.0 integrated five ET algorithms using BMA in 2014 and upgraded to v5.0 using a deep learning algorithm in 2022. Which version of GLASS product was fused in this study? Why don't the authors consider using deep learning fusion algorithms?

Response: Thank you for this comment. We considered GLASS v5, but since its temporal coverage is 2001–2015 and the fusion process requires the introduction of an evapotranspiration product with a longer coverage period, GLASS v4 was finally selected for this study. The GLASS v4 evapotranspiration product uses a Bayesian approach to estimate global land surface latent heat fluxes by combining five

traditional latent heat flux algorithms (MOD16 algorithm, improved PM, PT-JPL, MS-PT, and Semi-Empirical Penman Algorithm), using observations from 240 flux sites around the globe as a reference to determine the weighting values for each algorithm. In this study, the Bayesian model averaging method was chosen over deep learning algorithms for the fusion of multiple sets of ET products for the following reasons:

(1) Explicit uncertainty quantification and probabilistic outputs

BMA provides probabilistic outputs (e.g., confidence intervals) by directly quantifying the uncertainty of the model weights and the error distribution of the input data through a Bayesian framework (Vrugt et al., 2008). For example, when fusing multi-source ET data, BMA can explicitly give the contribution weights of different data sources and their uncertainties, which facilitates the assessment of the reliability of the fusion results (Vrugt and Robinson, 2007). Deep learning methods typically output deterministic results and require the additional introduction of a probabilistic framework to quantify uncertainty, but are computationally complex and weakly interpretable (Zhang and Zhu, 2018).

(2) Physical Interpretability and Model Transparency

The BMA approach preserves the physical meaning of each input model (e.g. Penman-Monteith equations, remote sensing inversion algorithms), and reflects the applicability of different models under specific conditions through the assignment of weights (Vrugt et al., 2008). The results can be directly related to the physical mechanisms of the input models, compounding the need for mechanism interpretation in geoscientific studies. Deep learning methods, as ‘black box’ models, have internal feature representations that are disconnected from the physical process, making it difficult to explain the logic of weight assignment or to correct for sources of model bias (Castelvecchi, 2016).

(3) Robustness for small sample data

When the amount of data is limited (e.g., sparse ground validation sites), the BMA method can avoid overfitting through Bayesian prior distributions and model weight optimization, and is particularly suitable for regional-scale ET fusion (Vrugt et al., 2008). For example, when fusing a small amount of vorticity covariance flux data with multi-source ET products, BMA can constrain the range of weights through prior knowledge. Deep learning methods, on the other hand, are prone to overfitting or underfitting when there are insufficient samples (Zhang et al., 2019), and data augmentation strategies are limited by physical plausibility in geomatics.

4. Table 2 shows that the spatial resolutions of the 30 ET datasets are different. How did the author solve the problem of spatial scale mismatch during the fusion process?

Response: Indeed, as you mentioned, the 30 ET datasets have different spatial resolutions. Therefore, we used a bilinear interpolation method to unify the spatial resolution of all the ET datasets to 0.5°×0.5° and 1°×1° before performing the fusion of the multi-source ET datasets.

We have made the following additions in section 2.2.

"Prior to the evaluation and fusion of the ET datasets, the spatial resolution of all ET datasets was standardized to 0.5°×0.5° and 1°×1° using a bilinear interpolation method. 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)."

5. The 30 ET datasets cover different time ranges. How to carry out ET fusion for years with missing ET data?

Response: We sincerely thank the reviewer for noting this. Since different ET datasets cover different years, for the common coverage years 1982–2011, the weights of each ET dataset under each vegetation type were obtained by performing BMA analyses based on 30 ET datasets, and these weights were applied to all the years 1982–2011. For the non-common coverage years 1980–1981 and 2012–2020,

the weights for each year were obtained by filtering all ET datasets covering that year for BMA analysis to obtain the corresponding weights. The BMA analysis process mentioned above is all based on the years 1991–2011, except that the number of ET products involved in the BMA analysis is changing.

We have made the following additions in section Text S7 and Figure S5.

"Text S7. Weighting scheme for ET dataset fusion

Analysis was performed based on the grid scale of each ET dataset. Among these datasets, 1982–2011 are the common coverage years for all ET datasets, and their weights are calculated as detailed in section 2.2 of the main text. As for the non-common coverage years 1980–1981 and 2012–2020, the weights for each year are obtained by filtering all ET datasets covering that year for BMA analyses to obtain the corresponding weights (see Fig. S5). For example, there are a total of 27 ET datasets covering the year 1980, and the weights obtained from the BMA analysis based on these 27 ET datasets are used as the weights of these 27 datasets in 1980; there are a total of 19 ET datasets covering the year 2019, and the weights obtained from the BMA analysis based on these 19 ET datasets are used as the weights of these 19 datasets in 2019."

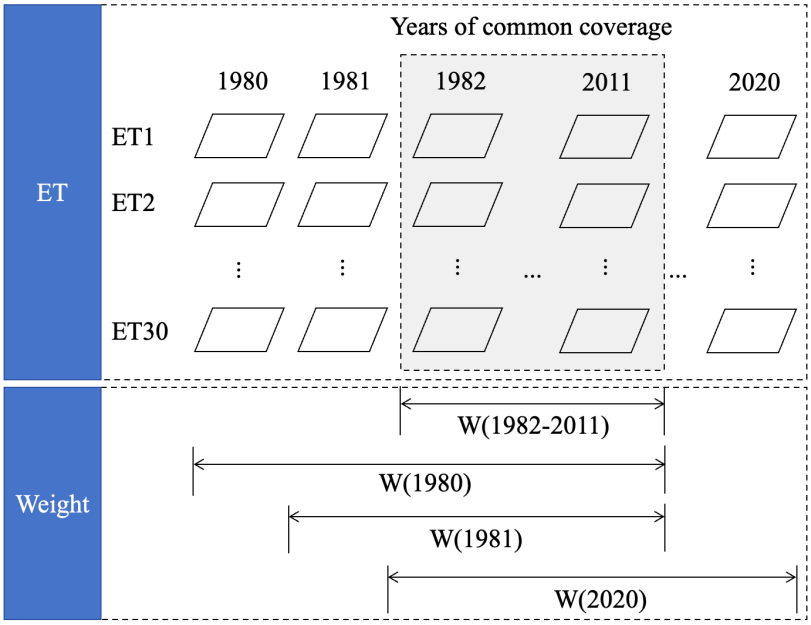

Fig. S5: Weighting scheme for ET datasets during non-common coverage years.

6. What are the spatial and temporal resolution of BMA-ET? How to handle the mismatches with 30 ET input datasets?

Response: In order to meet the needs of research at different spatial scales, we produced 2 sets of BMA-ET datasets with spatial resolutions of 0.5°×0.5° and 1°×1°. At the spatial scale, we used bilinear interpolation to unify the spatial resolution of the 30 sets of ET datasets to 0.5°×0.5° and 1°×1°. In turn, the subsequent ET evaluation and fusion were performed. At the temporal scale, the response content of the fifth comment was attended.

We have made the following additions in section 2.2:

"Prior to the evaluation and fusion of the ET datasets, the spatial resolution of all ET datasets was standardized to 0.5°×0.5° and 1°×1° using a bilinear interpolation method."

We have made the following additions in Text S7:

"Text S7. Weighting scheme for ET dataset fusion

Analysis was performed based on the grid scale of each ET dataset. Among these datasets, 1982–2011 are the common coverage years for all ET datasets, and their weights are calculated as detailed in section 2.2 of the main text. As for the non-common coverage years 1980–1981 and 2012–2020, the weights for each year are obtained by filtering all ET datasets covering that year for BMA analyses to obtain the corresponding weights (see Fig. S5). For example, there are a total of 27 ET datasets covering the year 1980, and the weights obtained from the BMA analysis based on these 27 ET datasets are used as the weights of these 27 datasets in 1980; there are a total of 19 ET datasets covering the year 2019, and the weights obtained from the BMA analysis based on these 19 ET datasets are used as the weights of these 19 datasets in 2019."

7. Is the observation interval of the ground measurements from FLUXNET half an hour? How to process observation data into monthly scale? Is nighttime observation data used?

Response: In this study, half-hourly data from FLUXNET2015 were selected as ET site observations. The half-hourly scale ET data were aggregated to the monthly scale, resulting in monthly scale ET site data for subsequent analyses. Given that flux data from nighttime eddy covariance measurements are usually subject to large deficiencies and errors, only daytime latent heat flux data were selected for this study. Daytime was defined as the period from 07:00 to 19:00 local time. The study did not use incident shortwave radiation to define the daytime period because occasionally large values of shortwave radiation occur at night.

We have made the following additions in section Text S1.

"Because flux data from nighttime eddy covariance measurements are missing and have large errors (Mahrt, 1999; Yuan et al., 2021), the study extracted only daytime latent heat flux data. Daytime is defined as 7:00 to 19:00 local time (McGloin et al., 2019; Yuan et al., 2021). The study did not use incident shortwave radiation to define daytime, as nighttime occasionally has larger values."

8. In line 181: What do 10 sites refer to? Does it refer to 60% of CRO sites? Please explain more clearly.

Response: I'm sorry that we didn't make this clear in the paper. Here, 10 sites refer to 60% of the CRO sites.

We have made the following additions in section 2.2.

"The sites were partitioned according to land cover type. We took cropland (CRO) as an example, a total of 16 CRO sites were involved, and 60% of the sites (corresponding to 10 sites) were selected to participate in the BMA fusion, while 40% of the sites (corresponding to 6 sites) were used for validation."

9. In section 2.2 (lines 176-195), “The ET fusion datasets for each vegetation type were spliced to obtain the final global ET fusion dataset”. How to obtain the boundaries of vegetation types at the regional scale? What is the accuracy? Have authors considered the fusion errors caused by land cover classification errors?

Response: Thank you for this pertinent advice. In general, the vegetation cover type boundaries are curves, i.e., Shapefile vector boundaries. However, in this study, the vegetation cover type we use is grid point data. And the vegetation cover type grid point data is produced by the previous study. The MCD12Q1 data are considered in this study. The MCD12Q1 data is a fusion of data from sensors Terra and Aqua, with a spatial resolution of 500 metres, and provides interannual global data on land cover types (from 2001 onwards), containing six classification systems, where the International Geosphere Biosphere Programme (IGBP) is used (Cai et al., 2014). The IGBP classifies land cover types into 17 categories, including 11 natural vegetation classifications, 3 land use and land mosaics, and 3 unvegetated land classifications. The product uses supervised classification in addition to additional post-processing of the data, i.e., some a priori knowledge and ancillary data are incorporated to improve the accuracy of the classification. This study uses the MOD12 Q1 dataset to classify global land cover. The following maps show the spatial distribution of masks globally (Figure R3) and for each of the 12 land cover types (Figure R4).

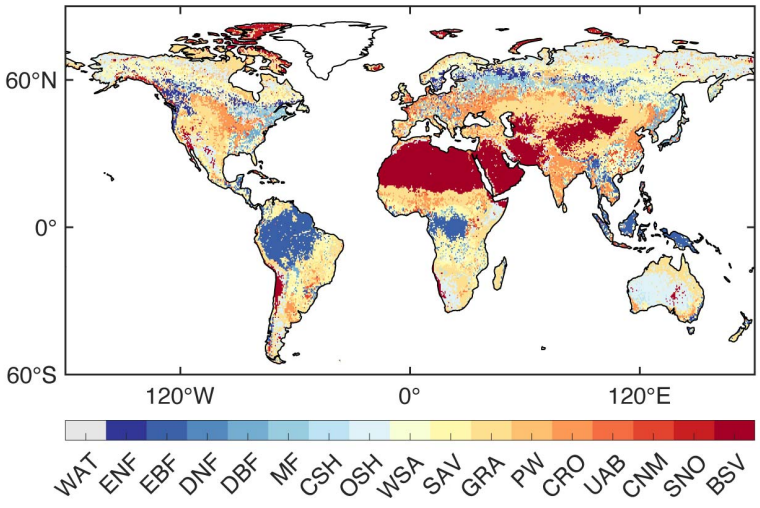

Figure R3 (Corresponding to Figure S3). Global spatial distribution of land cover types based on MOD12Q1.

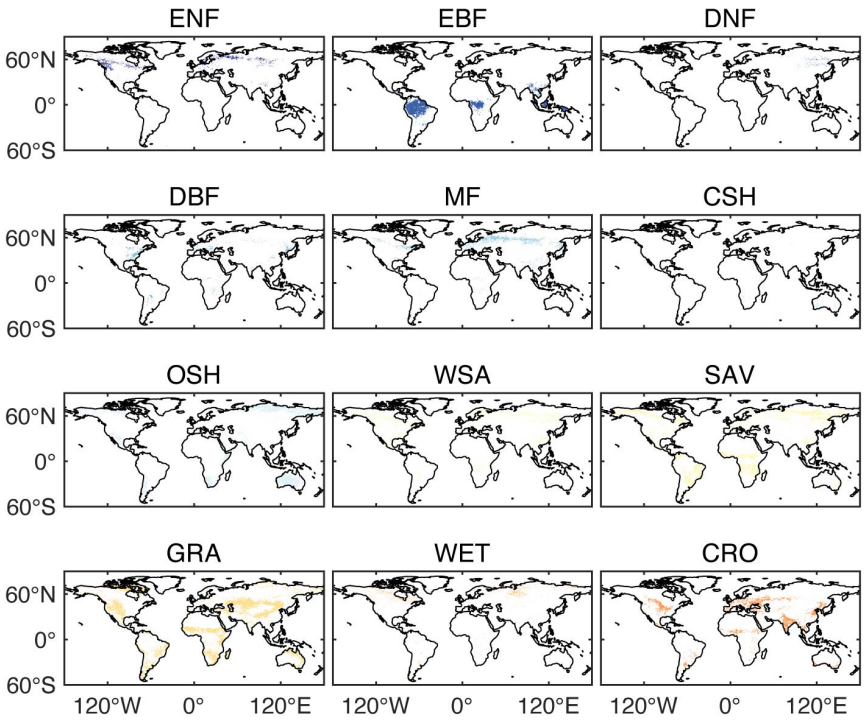

Figure R4. Global spatial distribution of 12 types of land cover types based on MOD12Q1.

On this basis, we also considered the comments of Reviewer #2. 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.

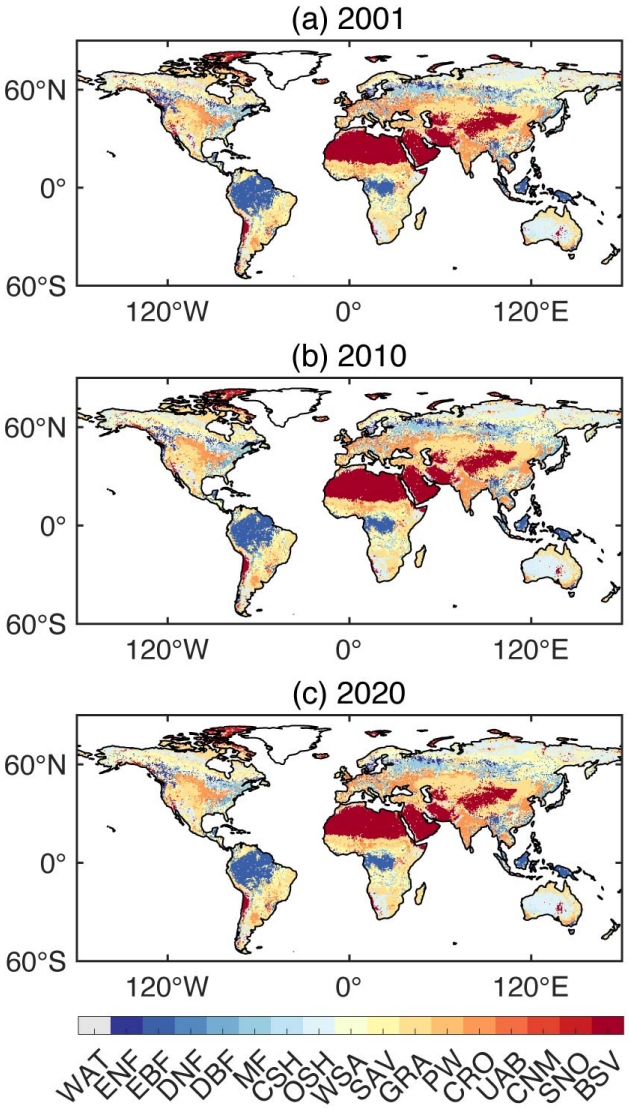

Figure R5 (Corresponding to Figure S33). Spatial distribution of land cover changes in three periods (2001, 2010, 2020) based on MOD12Q1.

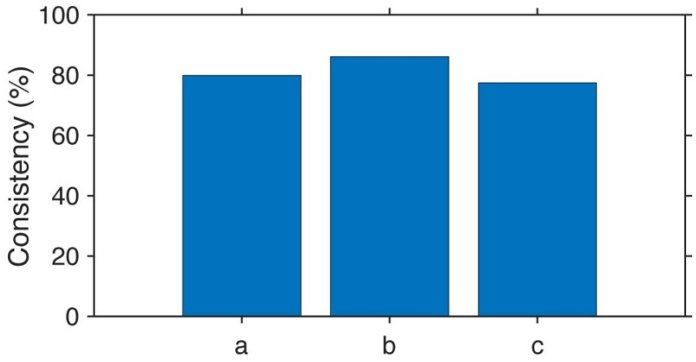

Figure R6 (Corresponding to Figure S34). Consistency between MOD12Q1 land cover types across years. Subfigures a-c show the level of consistency between land cover types for 2001 and 2010, 2010 and 2020, and 2001 and 2020, respectively. The

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.

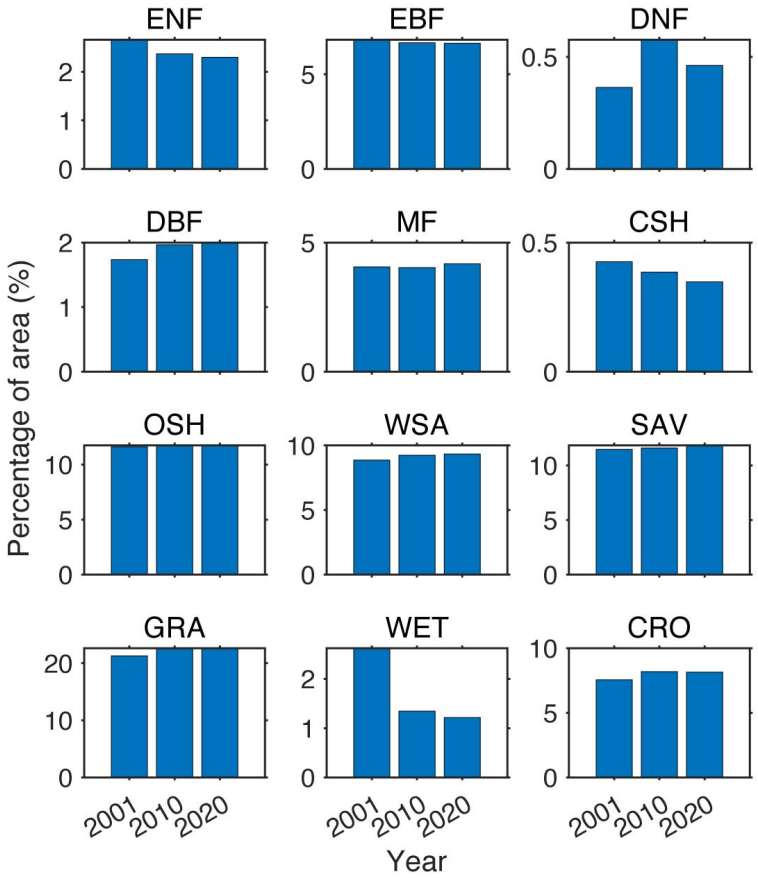

Figure R7 (Corresponding to Figure S35). Percentage of global land area covered by 12 vegetation types based on MOD12Q1. The bars in each subfigure represent the proportion of vegetation types in 2001, 2010 and 2020, respectively.

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."

10. In Figure 2, "the 12 vegetation cover types do not cover the entire study area. For areas not covered, an equal weighting approach was taken". Is this weight scheme reasonable?

Response: Thank you for your comment. Areas not covered by the ten vegetation cover types could not be analyzed for BMA as no site information was available. Here, we introduce a Bayesian-Three Cornered Hat (BTCH) method. This method is an advanced statistical approach for multi-source data fusion and error estimation, which is particularly applicable to the fields of remote sensing, Earth observation and model evaluation. Its core advantage lies in its ability to estimate the error characteristics (e.g., random error, systematic bias) and their relative weights of multiple data sources simultaneously without relying on real reference data. The computational process of BTCH method is detailed in Text S8.

We compared the results of ET fusion of uncovered areas of 10 vegetation types based on the BTCH method and the equal weighting method (Figure R8). The results showed that the correlation coefficients between the ET estimates based on the equal weighting method and the BTCH-based method were high in the area not covered by the 10 vegetation types, reaching more than 0.9 at most of the grid points (Figure R8). Moreover, only 11.6% of the global land area is not covered by 10 vegetation types, the percentage of uncovered areas is small and therefore introduces little error. This type of area is mainly found in North Africa, the Middle East and parts of Central Asia. In summary, in the areas not covered by the 10 vegetation types, it is reasonable for us to use an equal weighting approach to fuse all ET datasets.

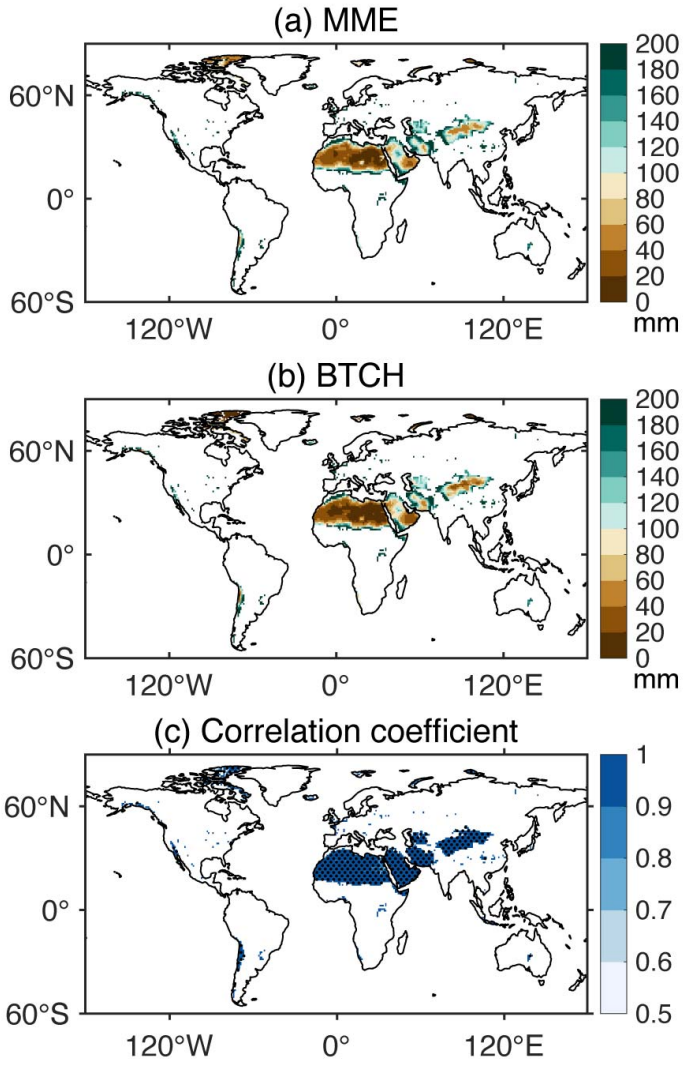

Figure R8 (Corresponding to Figure S36). Spatial distribution of multi-year average ET during 1980–2020 in the uncovered areas of 10 vegetation types. (a) the spatial distribution of ET based on the equal weight method, (b) the spatial distribution of ET based on the BTCH, (c) the correlation coefficient of ET between two methods, with the black dots representing that the ET of the grid point passed the significance test (p<0.05).

We have made the following additions in section 4.1 and Figure S36.

"In order to validate the reasonability of using an equal weighting scheme for the uncovered areas of 10 types of vegetation types, we compared the results of ET fusion of uncovered areas of 10 vegetation types based on the Bayesian-Three Cornered Hat (BTCH) method (Supplementary information Text S8) and the equal weighting

method (Fig. 36). The results showed that the correlation coefficients between the ET estimates based on the equal weighting method and the BTCH-based method were high in the area not covered by the 10 vegetation types, reaching more than 0.9 at most of the grid points (Fig. 36). Moreover, only 11.6% of the global land area is not covered by 10 vegetation types, the percentage of uncovered areas is small and therefore introduces little error. This type of area is mainly found in North Africa, the Middle East and parts of Central Asia. In summary, in the areas not covered by the 10 vegetation types, it is reasonable for us to use an equal weighting approach to fuse all ET datasets."

11. In Figure 4, 30 ET datasets were well evaluated, and Table 3 showed the guidelines for the use of ET datasets. So, in the BMA-ET fusion process, were all 30 ET datasets used for fusion, or only the recommended datasets used for fusion? If as the authors stated, the accuracies of RA and LSM are not good, why are they still used for fusion?

Response: Thank you for your comment. The BMA method will control the weights of different ET datasets. Bayesian model averaging (BMA) provides a methodology to explicitly handle conceptual model uncertainty in the interpretation and analysis of environmental systems. This method combines the predictive capabilities of multiple different models and jointly assesses their uncertainty. The probability density function (pdf) of the quantity of interest predicted by BMA is essentially a weighted average of individual pdf's predicted by a set of different models that are centered around their forecasts. The weights assigned to each of the models reflect their contribution to the forecast skill over the training period.

Based on the results of ET integrated evaluation, the performance of the datasets based on remote sensing inversion and machine learning reconstruction is relatively better. Therefore, we also selected eight ET datasets for the fusion study as a robustness test of the number of ET datasets involved in the fusion process. The eight ET datasets include four datasets based on remote sensing inversion (PML, GLEAM,

GLASS and PLSH datasets) and four datasets based on machine learning reconstruction (FLUXCOM-CRUNCEP_v8, FLUXCOM-GSWP3, FLUXCOM-WFDEI and MTE datasets). Only the better 8 ET data were selected for fusion, and there was little difference between the results of the two when all 30 ET data were involved in the fusion (Figure R9). The correlation coefficient between the 2 sets of ET fusion products exceeds 0.9 in most regions of the globe. This indicates that the BMA method gives higher weights to the better ET products. This also fully reflects the advantages of the BMA method. In order to prevent the effect of such a subjective behavior as selecting 8 out of 30 ET datasets, so this study still maintains the participation of all ET datasets.

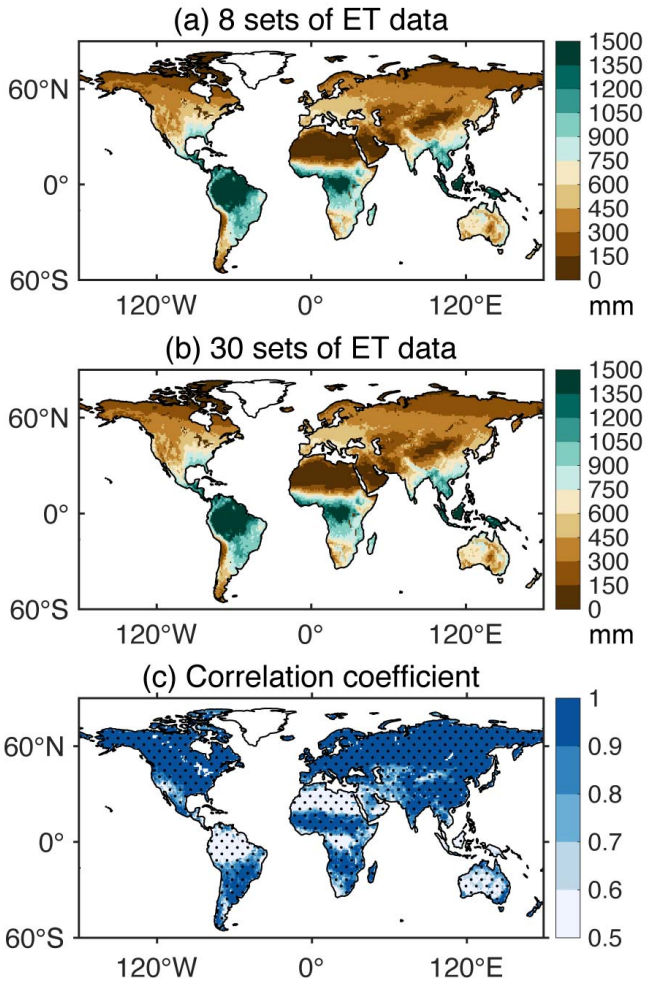

Figure R9. Spatial distribution of multi-year average ET during 1981–2018 over the globe. (a) the spatial distribution of ET fusion product from 8 sets of ET data, (b) the spatial distribution of ET fusion product from 30 sets of ET data, (c) the correlation

coefficient of two ET fusion products, with the black dots representing that the ET of the grid point passed the significance test (p<0.05).

12. In lines 237-238, the RS and ML ET datasets are recommended in the site scale validation results. Whereas, in lines 256-257, the ML ET datasets have greater TCH relative uncertainty. Do these two conclusions conflict? Please provide a detailed explanation.

Response: Thank you for your comment. The results show that the ML ET datasets have greater TCH relative uncertainty. This conclusion is for the basin scale. The basin-scale TCH analyses were performed based on the results of the basin-weighted average, rather than on a grid- by-grid basis within the basin. The results of basin averaging make it difficult to accurately characterize the true situation at the global grid scale. Therefore, we also used the TCH method to analyze the uncertainty over the globe on a grid-by-grid basis. For TCH relative uncertainty, machine learning–and remote sensing–based datasets have less relative uncertainty (Figure R10). The results of the global TCH analysis are consistent with the results of the site-based ET assessment, both of which concluded that the remote sensing and machine learning-based ET dataset has higher accuracy. Therefore, there is no conflict between the conclusions.

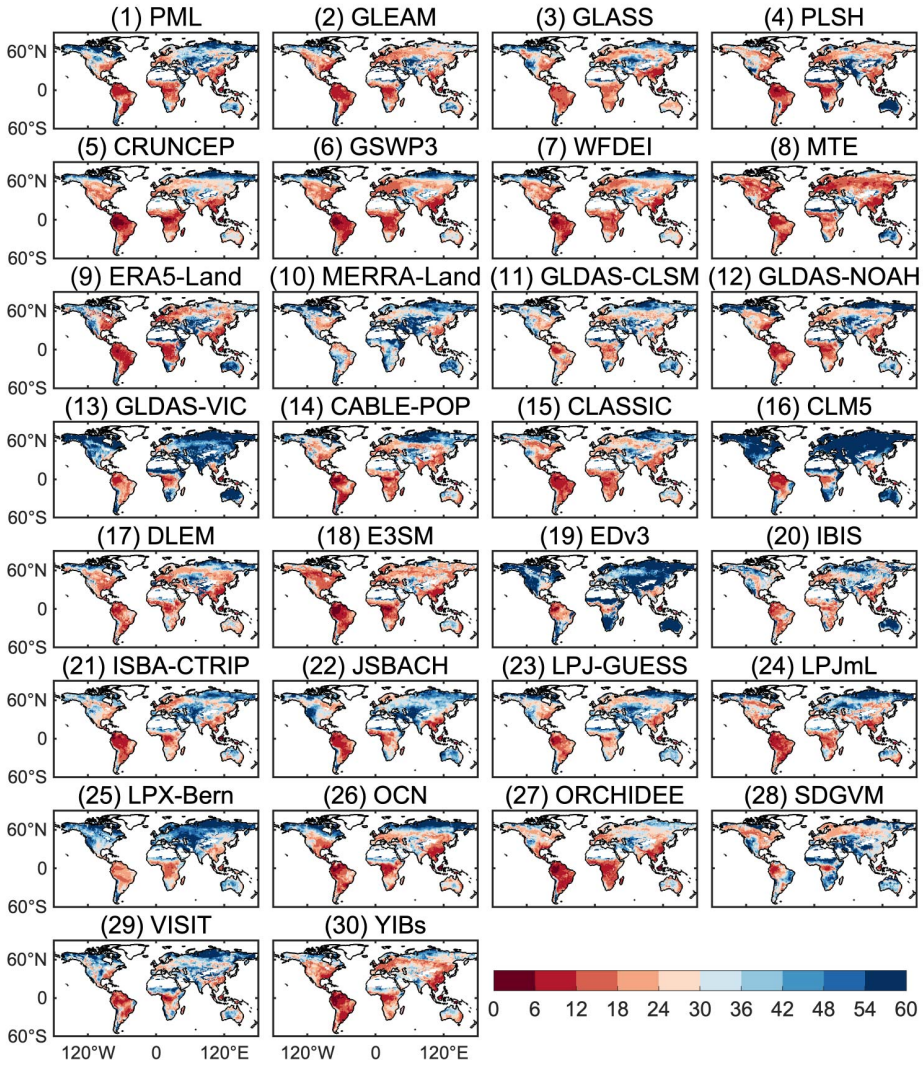

Figure R10 (Corresponding to Figure S24). Spatial distribution of relative uncertainty of TCH for each ET dataset on monthly scale for the common coverage years from 1982 to 2011 (unit: %).

13. In Figure 1, the common period of coverage for all ET datasets is 1982–2011. How did this study produce the BMA-ET dataset from 1980 to 2020?

Response: This issue is similar to the fifth comment you mentioned, so see the response to the fifth comment for more information. Since different ET datasets cover different years, for the common coverage years 1982–2011, the weights of each ET dataset under each vegetation type were obtained by performing BMA analyses based on 30 ET datasets, and these weights were applied to all the years 1982–2011. For the non-common coverage years 1980–1981 and 2012–2020, the weights for each year

were obtained by filtering all ET datasets covering that year for BMA analysis to obtain the corresponding weights.

We have made the following additions in section Text S7 and Figure S5.

"Text S7. Weighting scheme for ET dataset fusion

Analysis was performed based on the grid scale of each ET dataset. Among these datasets, 1982–2011 are the common coverage years for all ET datasets, and their weights are calculated as detailed in section 2.2 of the main text. As for the non-common coverage years 1980–1981 and 2012–2020, the weights for each year are obtained by filtering all ET datasets covering that year for BMA analyses to obtain the corresponding weights (see Fig. S5). For example, there are a total of 27 ET datasets covering the year 1980, and the weights obtained from the BMA analysis based on these 27 ET datasets are used as the weights of these 27 datasets in 1980; there are a total of 19 ET datasets covering the year 2019, and the weights obtained from the BMA analysis based on these 19 ET datasets are used as the weights of these 19 datasets in 2019."

14. In lines 355-356, the study recommended RS and ML based ET datasets (especially MTE and PML) based on the evaluation results. So why does the BMA-ET merge 30 ET datasets? Is it better to merge only MTE and PML?

Response: Thank you for your comment. This issue is similar to the eleventh comment you mentioned, so see the response to the eleventh comment for more information. The BMA method will control the weights of different ET datasets. Only the better 8 ET data were selected for fusion, and there was little difference between the results of the two when all 30 ET data were involved in the fusion. In order to prevent the effect of such a subjective behavior as selecting 8 out of 30 ET datasets, so this study still maintains the participation of all ET datasets. In addition, if only MTE and PML were fused, it would be difficult to fuse the two ET datasets to obtain ET data for a long time series (1980–2020) because the time periods covered by these two ET datasets are too short.

---------- 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

Cai, S., Liu, D., Sulla-Menashe, D., and Friedl, M. A.: Enhancing MODIS land cover product with a spatial–temporal modeling algorithm, Remote Sensing of Environment, 147, 243–255, https://doi.org/10.1016/j.rse.2014.03.012, 2014.

Castelvecchi, D.: Can we open the black box of AI?, Nature News, 538, 20, https://doi.org/10.1038/538020a, 2016.

Vrugt, J. A. and Robinson, B. A.: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, Water Resources Research, 43, https://doi.org/10.1029/2005WR004838, 2007.

Vrugt, J. A., Diks, C. G. H., and Clark, M. P.: Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling, Environ Fluid Mech, 8, 579–595, https://doi.org/10.1007/s10652-008-9106-3, 2008.

Zhang, H., Zhang, L., and Jiang, Y.: Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems, in: 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), 1–6, https://doi.org/10.1109/WCSP.2019.8927876, 2019.

Zhang, Q. and Zhu, S.: Visual interpretability for deep learning: a survey, Frontiers Inf Technol Electronic Eng, 19, 27–39, https://doi.org/10.1631/FITEE.1700808, 2018.