We are very grateful to Reviewer #1 (Dr. Marloes Mul) for the in-depth reading and the thorough review we received. We present below our detailed reply to the discussed points and further revision plan. The reviewer's comments appear in black and our responses appear in blue.

REVIEWER #1

I read the manuscript "How much water vapour does the Tibetan Plateau release into the atmosphere?" with great interest. The validation of many different ET products over these water towers of Asia has a lot of value. While the manuscript is generally well written and clear, I do have some specific comments and requests for clarification of the presented analyses.

Reply: We thank you for the review and the constructive feedback that will help us to further improve our work.

Regarding the validation:

- Provide clear explanation on the temporal scale the analyses were conducted (monthly?), this is not always clear

Reply: Both the validations based on eddy covariance observations and the basin-scale water balance method were conducted at monthly scale. We will make it clearer in the revised version.

- Provide clear explanation on the period used for the analyses (in some cases the overlap of the *in situ* data (either EC towers or water balance estimates) and products is rather short

Reply: It is true that the overlap period of the *in situ* data and products is short in some instances, and in some case such as the Namco site there is no overlap, since the in-situ measurements started in 2019 while some products did not extend beyond 2019. As regards the site-scale validation, we will add a table in the supplementary materials to include the information necessary for this purpose, i.e. the overlap period for validation, the number of observations, and values of the error metrics. As regards the basin-scale validation, the validation period will also be added in the main text.

Our approach was to utilize long time series data (as long as possible) for the inter-comparison and trend analyses. More precisely, for the inter-comparison analysis we used the overlap period of all products (2003~2013). The trend analysis was carried out for the available period for each dataset, being aware that the overlap period of all products was relatively short. We note that many satellite remote sensing ET datasets with high spatial resolution are estimated based on MODIS data, which started from 2000, while long-term ET datasets at comparable spatial resolution are still scarce.

- Basins used in the water balance estimation is not always clear, eg figure 1 doesn't show the Heihe basin (is this the Hexi corridor and is the entire basin included in the map/analyses?). In figure 1 what does the stripped area refer to? A table with information would be useful with some additional information on the data used from the studies by Ma and Zhang and Wang et al. Also the basins are referred to as the Yangtze/ Yellow river basin, but as far as I understand these only cover the upper part of the basin. Please provide some additional information on the extent of each of the basins analyses (eg provide name of the gauging station where the basin was delineated). Also in figure 3, there is a reference to TP, which basin/ area does this refer to (the entire TP area shown in figure 1 or the area of all the basins combined, which are two different areas)?

Reply: Sorry for the ambiguities in some of our illustrations and related information. We will also revise the figures and add more information accordingly. Overall, we used monthly ET_{wb} from five basins from previous studies (Ma and Zhang., 2022; Wang et al. 2021), including the headwaters of Yellow basin (UYE), headwaters of Yangtze basin (UYA), upper Heihe basin (UH), Inner Tibet Plateau (INTP) and Qaidam (QDM) basins. It is true that these only cover the upper part of the basin, and we will define explicitly the extent of these regions and present this information clearly. A new table will be added to provide additional information, i.e., the extent of the basins and the names of the gauging stations.

As regards Figure 3, we intended to use TP to represent the area of all the five basins combined. To avoid the potential misunderstanding, we will revise it to *5 basins* (the area of all the basins combined) in the revised version.

- Color scheme of figure 3 is not fully intuitive, for example the r2 is deep red for high (=good) values)

Reply: We will redraw the figure to make it more intuitive.

Figure 5: what do the different colors of the bars mean?

Reply: We intend to show the global satellite remote sensing-based ET dataset in dark blue and model-based ET dataset in light blue, and the regional ET dataset in red. We will add this description in Figure 5.

Trend analyses (figure 7):

- The calculation of the trends could be affected by an exceptional year with high or low ET at the beginning or end of the time series (since there is quite some yearly variation and the trends are often relatively minor). Could you say something about the significance of these trends as well? Also for the SynthesisET both the first two years and the last two years seem to be outliers and related to the "temporal inconsistencies" of the product. Was this data properly vetted before including in the analyses?

Reply: We fully agree with you that the trend could be affected by the exceptional years at the beginning or end of the time series. This is also why we choose a robust regression method to estimate the trend of ET, rather using simple linear regression, since the robust regression can reduce the impact of outliers. We will add the significance level of the trends in the figure and main text.

As regards the temporal inconsistencies of SynthesisET, we carefully checked it for several times and we are pretty sure about the existence of the temporal inconstancies. In fact, this issue was also noticed by the authors of the SynthesisET dataset, and they tried a different synthesis strategy d in a later regional study on the Northern China (Wang et al., 2021). This could also be seen from the temporal variation SynthesisET in Figure RC1-R1. Figure RC1-R1 will used to replace the Figure 7 in the manuscript.

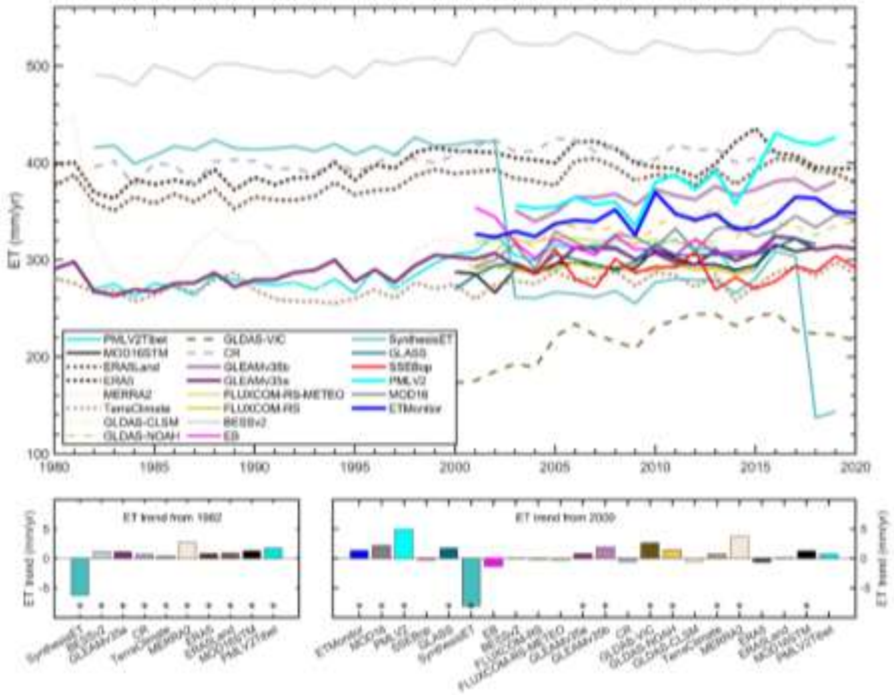
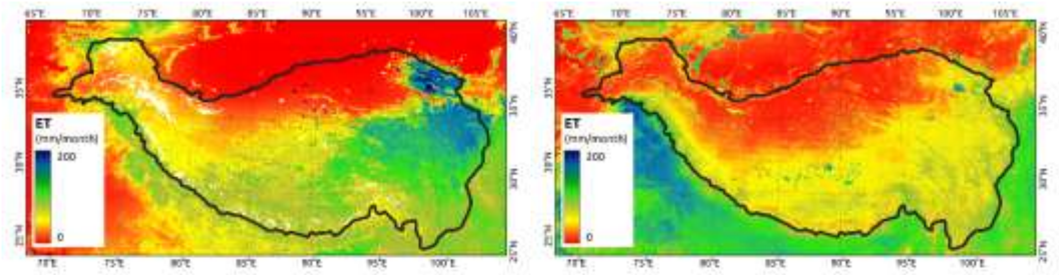


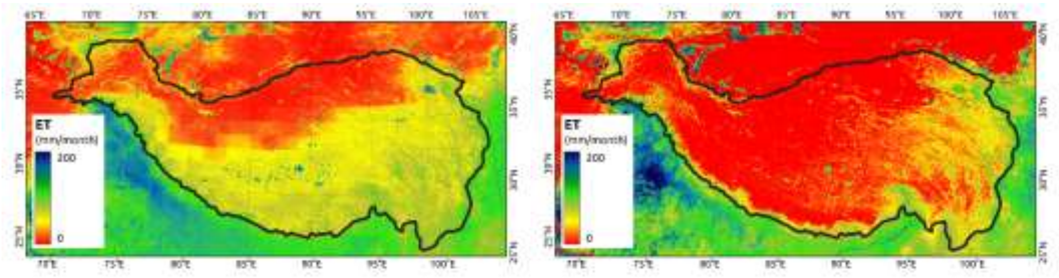
Figure RC1-R1: Yearly variation of ET in the TP by different products. The inset panel shows the annual ET trend by different products. *: trend with significance level ($p<0.05$). In the top panel, the reanalysis data is shown in dotted line, and the land surface model-based data is shown in dash line.

We also check the spatial variation of ET by SynthesisET (as shown in following Figure RC1-R2). Before 2000, SynthesisET showed quit high ET values (e.g., in the eastern TP). While after 2019, SynthesisET showed extremely low ET values in the eastern TP.



SynthesisET, July, 2000

SynthesisET, July, 2008



SynthesisET, July, 2013

SynthesisET, July, 2019

Figure RC1-R2: Example of spatial variation of ET by SynthesisET in July of different years.

Reference:

Wang, L.;Wu, B.; Elnashar, A.; Zeng, H.; Zhu,W.; Yan, N. Synthesizing a Regional Territorial Evapotranspiration Dataset for Northern China. Remote Sens. 2021, 13, 1076. https://doi.org/10.3390/rs13061076.

- Why are many of the products with longer time series (eg ERA5Land, SynthesisET, BESS, MERRA2) not presented with their full timeseries?

Reply: The ERA5Land ET shows very similar trend with ERA5. As regards SynthesisET, we already noticed its temporal inconsistence, thus we did not include it in the annual trend analysis. To reduce the concerns of reviewers, we will include all the long-term ET products in the revised version.

Analyses of "ET components"

- As mentioned by the authors these different sub-components of ET are not validated and with the wide range of values derived from the different products, what conclusions can really be drawn? This is especially a question for the open water ET (maps in figure 9 shows large areas evaporating from water surfaces) and sublimation (which is validated how?)

Reply: It is true that the evaluation of different ET components was still limited due to the scarcity of available data and a comprehensive evaluation based on more observations would help to further evaluate the ET components and improve the algorithm performance. This analysis on the ET components has not been fully investigated by previous publications. We intended to use it to explain the difference among ET products and to answer the question: which processes play a significant role in determining the total ET. We also noticed that previous studies mostly focus on total ET, e.g., magnitude, spatial variation, temporal trend, etc., while the ET components were not fully investigated. Meanwhile, many studies estimated based on big-leaf model, and a few studies estimate total ET based on the separate estimation of ET components. These components reflect the different water phase change processes that are regulated by different factors, e.g., transpiration is mainly controlled by the plant physiology through the regulation of stomata behavior, soil evaporation is determined by heat and mass transfer in the top soil with liquid water present at some depth below the surface t, the rainfall interception loss is mainly related to the canopy morphology and rainfall intensity and the sublimation is associated with higher enthalpy change than vaporization process and near surface air humidity and temperature. So, we believe this analysis on the ET components is helpful, because at least starts with treating correctly each water phase change.

It is important to note that reliable independent reference measurements on each component of total vapour flux are very scarce. The anonymous Referee #3 (RC3) suggest us to use the ensemble mean of the ET components by different products, which may be close to the truth. We will try to check if it works. We also notice that averaging properly would not provide good estimates, since the it applies only to random errors, not to the use of the wrong algorithm. According to the results in Section 3.2.3, the median values of the ratio of Es, Ec, and Ei to total ET is 50%, 30%, and 5%. A recent study shows the contributions of Es, Ec, and Ei to total ET are 68.21 %, 23.57 %, and 8.21 %, respectively in the Three Rivers Source of the Tibetan Plateau (Zhuang et al., 2024), which is actually quite close to our estimates. After the analysis in our study, we may generally conclude that soil evaporation (Es) contributes most to total ET in the whole TP, and further study should pay more attention to it.

Reference:

Zhuang, J., Li, Y, Bai, P, Chen, L, Guo, X., Xing, Y., Feng, A, Yu, W., Huang, M.: Changed evapotranspiration and its components induced by greening vegetation in the Three Rivers Source of the Tibetan Plateau. J. Hydrol., 633, 130970, https://doi.org/10.1016/j.jhydrol.2024.130970, 2024.

Analyses related to the “response to different environmental factors”

- The purpose of these analyses are not entirely clear to me. First, the analyses are done for the median value of the correlation, whereas it was already very clear that there is a large variance between the different products. Also several products utilize these input data (Rn, LAI, P) for estimating ET, how is this kind of dependency considered in the analyses? Do different types of models have stronger or weaker correlation with these environmental factors? And what does that mean for the interpretation of the analyses?

Reply: Thank you very much for the comments. Analyzing the impact of environmental factors on ET is helpful to reveal the governing factors and the mechanisms determining the variability of ET. Itis also helpful to analyze why and how the ET algorithms/product capture the ET variation caused by the environmental change. It is also true that different models have stronger or weaker correlation with these environmental factors, which indicate the observed response to forcing factors is algorithm dependent. Meanwhile, several products utilize these input data (Rn, LAI, P) for estimating ET, and these products may show higher correlation with these factors. Hence, we think both the algorithm itself and the input data can impact the response of estimated ET to environmental factors.

We also noticed that the current analysis is very limited and a more comprehensive analysis could be done to illustrate this issue better. Hence, we will remove it from current manuscript and prepare another paper on it for a more robust analysis.

- Did any of these factors also influence the partitioning of ET into ETc and ETs?

Reply: We did not mention this issue in the manuscript. But, we think the answer is yes. This is especially true for leaf area index. Higher leaf area index is generally associated with higher plant transpiration and interception loss. For example, a recent study shows that the vegetation greening (judged by increasing LAI by 0.009 $m^2/(m^2\ a)$ with $p < 0.05$) caused unequal different changes in ET and its components, i.e., 1.95 mm/a, −2.41 mm/a, 1.33 mm/a, and 3.03 mm/a for ET, Es, Ec, and Ei, respectively, in the Three Rivers Source of the Tibetan Plateau (Zhuang et al., 2024), which clearly indicates its influence on the ET partitioning.

Reference:

Zhuang, J., Li, Y., Bai, P., Chen, C., Guo, X., Xing Y., Feng A., Yu W., Huang, M.: Changed evapotranspiration and its components induced by greening vegetation in the Three Rivers Source of the Tibetan Plateau. Journal of Hydrology, 633, https://doi.org/10.1016/j.jhydrol.2024.130970, 2024.

Discussion:

- General reflection of the validation methods employed, doesn't really add much information. The incorporation of seasonal land cover conditions or lack thereof is only explained for 3 products, but then no reflection on how that has affected the results. Or how relevant negative latent heat fluxes are (does this happen often or only occasionally?). The reflection on the water balance estimations are also very general and could have been included in the introduction (there is no reflection based on this specific study). For example, the assumption of not incorporating meltwater could have been explained in the method but is not an outcome of this research.

Reply: We understand the reviewer' concern, and will revise the manuscript to focus more on the findings of the current study. We will further revise our discussion by focusing more on the topic and results of this study.

The in-situ observation by eddy covariance system is recognized as the standard method for monitoring energy and mass fluxes to validate high-resolution ET (Baldocchi, 2020). For example, Chen et al. (2024) validated several ET products with spatial resolution ranging from 1km to 50km by comparing ET estimates with eddy covariance observations at site scale. However, it is important to note that the tower-based eddy covariance observations have a very small footprint (approximately several hundred square meters depending on the weather conditions). Consequently, the direct comparison of site-scale observations with the coarse-resolution ET products (e.g., 25km) is problematic due to the severe problem of spatial mismatch of footprints. A more comprehensive validation approach, that considers both the in-situ measurements and basin-scale estimations, has been suggested to improve the reliability of estimated accuracy (Liu et al., 2023). Therefore, to increase the quality of our validation results, we also included validation based on basin-scale estimates of ET, which have a much larger footprint (roughly several hundred to kilo meters depending on meteorological conditions). Considering the relatively sparse distribution and small footprint of the flux-tower based eddy covariance observations, the water balance method is an useful validation method, especially of the coarse-resolution ET data products.

In this study, these two validation methods showed generally consistent results when validating the high-resolution ET. If judged by the KGE of site-scale validation, the accuracy of the high-resolution ET products can be ranked as: PMLV2 > ETMonitor > MOD16STM > GLASS > MOD16 > SynthesisET > SSEBop. If judged by the KGE of basin-scale validation, the accuracy of the high-resolution ET products can be list as: ETMonitor > PMLV2 > MOD16STM > SSEBop > GLASS > MOD16 > SynthesisET. Although both indicate that ETMonitor, PMLV2, and MOD16STM are most accurate and the rest four are less accurate among the high-resolution ET products, there was a difference in the ranking of ET products. This is probably related to the processes captured by these two validation methods. The eddy covariance observations captures the net water vapour flux integrated across different processes at certain point and during a certain period of time, i.e. plant transpiration in dense vegetation regions, snow sublimation in dry snow regions, evaporation of canopy-intercepted water when the canopy is wet due to intercepted rainfall, and the observed vaporization depends on the land site condition during the observation period, which may vary seasonally and yearly due to factors such as snow/ice occurrence, intercepted water and vegetation growth. The basin-scale water balance estimated *ETwb* is essentially the residual of observed liquid water fluxes, which is assumed to be the net water loss to the atmosphere at basin scale. Compared with the site-scale observations, the basin-scale *ETwb* can capture the impact of land cover change within a large catchment on the ET. For example, the mean water level of lakes in the TP increased by 0.20 m/yr from 2000 to 2009 and lake water mass increased significantly (Zhang et al. 2013), which surely caused higher ET in TP since open water evaporation is generally higher than other land cover types. However, most ET products (e.g., MOD16, PMLV2, etc.) assume constant land surface conditions throughout the year or multiple years, which indicates that they cannot capture the temporal changes of these vaporization process associated with changes in land cover. In contrast, ETMonitor adjusts the daily land cover based on seasonal land cover condition (water cover and snow/ice cover), which enables it to partly reflect the impact of seasonal and yearly extent of liquid and solid water on total ET (Zheng et al., 2022). This probably explains partly why ETMonitor performs slightly better than PMLV2 when validated by basin-scale water balance methods, while it is the opposite when validated with in-situ observations.

Reference:

Liu, H, Xin, X, Su, Z., Zeng, Y., Lian, T., Li, L., Shanshan S.: Hailong Zhang Intercomparison and evaluation of ten global ET products at site and basin scales. J. Hydrol., 617, 128887, https://doi.org/10.1016/j.jhydrol.2022.128887, 2023.
Zhang, G., Yao, T., Xie, H., Kang, S., and Lei, Y.: Increased mass over the Tibetan Plateau: From lakes or glaciers?, Geophys. Res. Lett., 40, https://doi.org/10.1002/grl.50462, 2013.
Zheng, C., Jia, L., and Hu, G.: Global land surface evapotranspiration monitoring by ETMonitor model driven by multi-source satellite earth observations, J. Hydrol., 613, 128444, https://doi.org/10.1016/j.jhydrol.2022.128444, 2022.
Baldocchi, D. D.: How eddy covariance flux measurements have contributed to our understanding of Global Change Biology, https://doi.org/10.1111/gcb.14807, 2020.
Chen, X. Yuan, L., Ma, Y., Chen, D., Su, Z., Cao., D.: A doubled increasing trend of evapotranspiration on the Tibetan Plateau. Sci. Bull., https://doi.org/10.1016/j.scib.2024.03.046, 2024.

- The discussion related to the different types of models comes a bit out of the blue, for example in table 2 the model type is not provided, which makes is difficult to validate a statement such as (first sentence) " PM-type model demonstrated superior accuracy compared to other models". Also ".. models that incorporate soil moisture to detect water stress…" can not be checked, which models do or do not incorporate soil moisture? Also to go in depth into the methodology of each product seems to go beyond the objective of this research, especially since it unclear why some models are singled out and others not (nor a statistical comparison between for example PM vs non-PM models is not done.

Reply: We will double check and revise the manuscript to make sure all the necessary information is included and the statements can be easily checked. We already stated in the manuscript that "Among the evaluated ET products, there are 14 products that primarily use remote sensing products, including 2 products (SSEBop and EB) based on land surface temperature (LST), 8 products

(ETMonitor, MOD16, MOD16-STM, PMLV2, PMLV2-Tibet, GLEAMv35a, GLEAMv35b, BESSv2) based on PM-types models (including Penman-Monteith equation, Priestley-Taylor equation, Shuttleworth-Wallace equation), 4 products (FLUXCOM-RS, FLUXCOM-RS-METEO, GLASS, SynthesisET) based on data-driven methods (machine learning method or ET products ensemble method)." To make the information more intuitive, we will move it to the Section 2.2.2. More information on whether soil moisture is considered in a given data product will be added in Table 2 by listing the main forcing data.

Our primary objective is to find out how accurate are the ET products in the TP, which is closely related to the algorithm applied in each product. Since we evaluate 22 products, there are 22 models to be discussed, which is actually too much and will make the manuscript unfocused. Therefore, we discussed the methodology of some representative ET products. The difference between the PM and non-PM model could be checked in Section 3.1.1, which showed that the best three products are all PM -type model-based products (ETMonitor, PMLV2, MOD16STM), while the LST-base (SSEBop) and data-driven products (GLASS and SynthesisET) had overall a low accuracy. We will present this statement more clearly in the revised version.

- The uncertainty of the SynthesisET product was already mentioned in the results section, is this really an important outcome of this research (important enough to single it out in the discussion?)

Reply: Thank you for the comments. In the results section, we evaluated its accuracy and compared with other products to identify a temporal inconsistence. In the discussion section, we try to explain the reason of its relatively poor performance, since we expected the fusion of different datasets should have improved the overall accuracy. We addressed the importance of the ensemble method in the discussion, which might be helpful to guide further studies.
