Reply to Referee's comments

Referee #2

Referee's comments are in black text

Authors's response are in blue text

1) General Comments: In this work, "Unveiling the Impact of Potential Evapotranspiration Method Selection on Trends in Hydrological Cycle Components Across Europe," the authors assess 12 potential evapotranspiration (PET) formulations across the European continent using regional hydrological modelling to quantify their impact on PET trends and their implications for the main hydrological cycle components: actual evapotranspiration (AET), total water storage (TWS), and runoff (Q). They conclude that the PET model selection conditions the simulated trends and influences the analyzed hydrological component. The paper reads well; I enjoyed it while reading it. Moreover, I think the study is relevant for the catchment hydrological community; the impact of PET formulations has usually been overcome in calibration frameworks, and assessing its actual impact at a continental scale is a valuable result.

   We deeply thank to Referee #2 for constructive feedback on this manuscript. The valuable provided comments will contribute significantly to enhancing the manuscript's quality. We are grateful for your time and effort invested in improving our study.

   However, I have some concerns, especially regarding the methodological approach. On the one hand, I missed information about the model framework, for instance, what baseline calibration you used and how the main analyzed hydrological cycle components are linked to the model. This is key to understanding the impact of your analysis. On the other hand, the authors talked about trends, but no proper statistical trend analysis has been performed.

   Details regarding the model setup and its performance are thoroughly addressed in Referee #1 comments under Section 3) Methods and data, specifically in subsections (e) and (f). Additionally, we

conducted an analysis considering significant trends, as discussed in Referee #1 comment on subsection (h) of Section 3) Methods and data.

Specific comments

2) Introduction

a) In paragraph one (lines 24-33), it would be nice to briefly mention that other concepts like reference evapotranspiration are widely used when computing AET.

We agree with your comment. We have addressed the discussion on reference evapotranspiration in Referee #1's comment under subsection (a) of the Introduction section. The text that we will include in the revised manuscript is as follows: *Potential evapotranspiration (PET) is the potential to evaporate water from the land surface to the atmosphere without Any limitation to water availability. Although the concept has been in use for centuries, Thornthwaite (1948) was the first to formally introduce the term "potential evapotranspiration" in the scientific literature. A related but distinct concept is "reference crop evapotranspiration", which is sometimes used interchangeably with PET. However, these terms differ in their conceptual basis and applications. Reference crop evapotranspiration specifically estimates the water requirements of a standardized reference crop under ideal conditions, whereas potential evapotranspiration provides a broader representation of water and energy exchange processes over diverse landscapes and large regions (Xiang et al., 2020).*

b) In paragraph two (lines 24-33), I would also include some sentences explaining why there have been more than 50 models for computing PET. Are they physically based formulations? Are they empirical and therefore linked to where they were initially formulated?

We would like to clarify that, based on recent literature, more than 100 empirical PET formulations have been developed. We will correct this information in the revised manuscript. The majority of these are temperature-based methods (40+), followed by radiation-based methods (30+), while combination-based methods are the least common (10+). Many of these empirical

methods were initially developed and tested for particular regional scales or climatic conditions. For instance, the Thornthwaite method is most suitable for humid climates, while the Hargreaves-Samani method is particularly effective in arid and semi-arid regions. Similarly, the Hamon method is suitable for all climates. We will correct and reframe lines 24-33 in the revised manuscript.

c) In the third paragraph (lines 41-47). Since your study assesses the impact of PET selection in other water cycle components, I would include more context and references about the connection between different components.

We agree with your comment. We will further extend this paragraph in revised manuscript.

3) Methods and data

a) Could you elaborate a bit about the quality of your data? Later in the discussion, you mentioned that their uncertainties were important to your results.

We agree with the Referee's comment, which has also been highlighted by the other two Referees. PET estimation and hydrological modeling are highly dependent on input data quality. The EM-Earth dataset provides high-quality precipitation and temperature data and has been shown to perform well over Europe (Tang et al., 2022). It has undergone climatology-based bias correction and accounts for precipitation undercatch. However, since EM-Earth does not include all necessary variables for PET estimation, we utilize ERA5-Land as a complementary dataset. ERA5-Land has been demonstrated to perform better than other reanalysis datasets, including ERA5 and ERA-Interim (Muñoz-Sabater et al., 2021). Several recent global studies follow a similar strategy, combining precipitation and temperature from EM-Earth with radiation, wind speed, and other meteorological variables from ERA5-Land (Tang et al., 2023; Yin et al., 2024).

b) How were the criteria used for choosing the 533 catchments? You mentioned that you try to cover all European climates, but is that the only reason? Why is there no catchment in Italy or Greece? Why are there these big differences between catchment sizes?

We acknowledge your comments. To select the catchments, different filters were applied. First, the catchment area should be greater than 500 km^2 to ensure that only sufficiently large basins are considered, given the spatial resolution of the meteorological dataset. Second, the observed discharge data for each catchment should be available for more than 10 years. Finally, the catchment should maintain a closed water balance based on Budyko space. It is assessed using the Evaporative Index ((P-Q)/P); we consider the catchment which have Evaporative Index of less than one. Catchment fulfill all these three conditions, then we consider that catchment. We will add this to the revised manuscript.

c) Why do the authors select these 12 specific PET models? Please add in Appendix A1 reference to each one of the chosen formulations.

We selected widely used PET methods from different categories to ensure broad applicability across hydrology, agriculture, and climate science. Additionally, we included a PET formulation that accounts for stomatal responses, which has gained increasing attention due to its ability to capture the impact of stomatal regulation on evapotranspiration. We will add the references in the Appendix A1 in the revised manuscript.

d) The authors mention that "the basins were not calibrated for each PET method to access their response in hydrological cycle components." I understand that this is a hypothesis of your study, but in any case, I assume they must be a baseline calibration. How does the model perform in this baseline calibration? Which PET is considered in this baseline calibration? Which parameter set was used in this reference calibration? Which the target variable that the model was calibrated for? I think that, in general, a deeper description of the model might help the reader to understand the implications of selecting one or other PET. Especially how the parameterization of PET-AET-soil water balance interaction is solved.

Thank you for raising this question. We have discussed the modeling details, including the model setup and its performance, in Referee #1 comments, Section 3) Methods and Data, specifically in subsections (e) and (f). The discussion is as follows:

For each basin, we performed 12 model runs, with each run corresponding to one PET method. Therefore, for the 553 catchments, the total number of model runs is 6 636 (553 $\times$ 12). The mHM model was run at a daily time step with a spatial resolution of 0.125° $\times$ 0.125° grid resolution. We did not perform any model calibration in our study. We used the default model's parameterization because we wanted to mimic how large-scale/global hydrologic models performed, as if they would be employed across continents or global scale. The basin-wise setup used here, enabled us to estimate corresponding river discharge, and quantify all components of the water balance equation. The default parameterization of mHM has been shown to perform well in previous studies (Kumar et al., 2013; Rakovec et al., 2016). Furthermore, it has been demonstrated as one of the best-performing configurations compared to other large-scale hydrological models (Samaniego et al., 2019). For instance, Samaniego et al. (2019) compared the performance of mHM with other hydrological models across 357 catchments. Their results showed that the median Kling–Gupta efficiency (KGE) for mHM was approximately 0.6 across these catchments. To address reviewer's concern regarding the model's performance, we conducted an evaluation of its performance against discharge across the basins, as presented in Figure 1. Overall, the model performed well, with median KGE values ranging from 0.6 to 0.75 for most PET methods. However, the Blaney-Criddle method showed a median KGE slightly higher than 0.3, which was lower compared to other methods.

The default parameterization came from the model developers, and it was originally established over a diverse set of German basins, in the pioneering work of Samaniego et al. (2010). Since then, mHM has become a well-established model that has been extensively evaluated across various basins and hydrological variables (Rakovec et al., 2016; Samaniego et al., 2019; Boeing

et al., 2024). For example, Rakovec et al. (2016) analyzed the model's performance across 400 European catchments. Their evaluation compared mHM's discharge simulations using 36 different parameter sets and found that the model's performance was consistent regardless of parameterization. Introducing new model setups or performing additional calibration and comparative analyses is beyond the scope of this study. We agree that the calibration aspect is important and offers interesting insights. However, we prefer to explore it in our future research. Additionally, we will discuss the potential effects of model calibration in the discussion section to provide further context on this limitation in our manuscript.

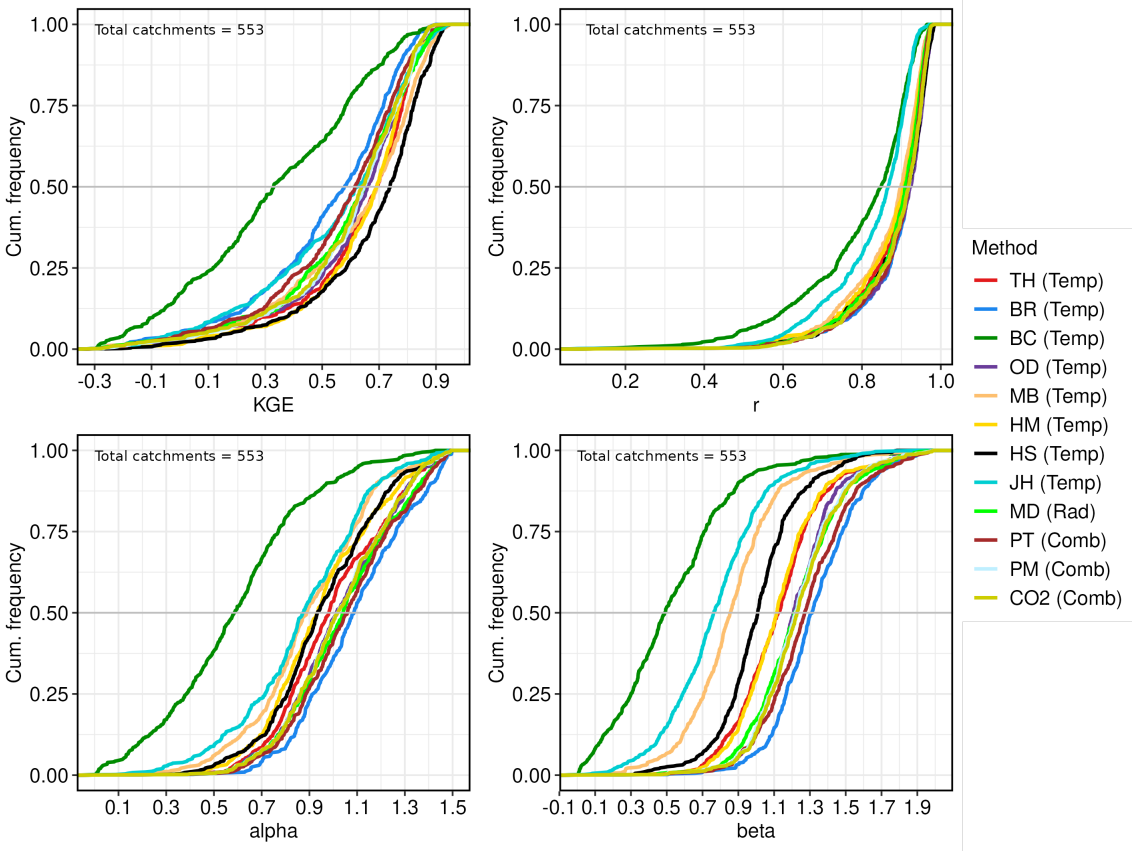

Figure 1. Evaluation of the hydrological model (mHM) performance using simulated monthly streamflow across 553 catchments, forced with EM-Earth meteorological data. The figure presents the cumulative frequency distributions of the Kling–Gupta efficiency (KGE) and its three components: correlation (r), variability ratio (alpha), and bias ratio (beta), providing insights into the model's performance across different PET methods.

e) Regarding the trend analysis, to my knowledge, Sen's slope method is a non-parametric test to compute the magnitude and direction of linear change on a time series, as the authors state in lines 146-150. To be able to talk about a trend, it needs to be significant. To assess that, other statistical tests should be performed. The Mann-Kendall test and its variations are the most commonly used for these purposes. For instance, one paper you cited, Anabalón and Sharma, 2017, used a test of the Kendall family (Seasonal Kendall trend test) to determine the trends and the Sen's slope test to determine their magnitudes. If only Sen's slope is computed, the authors might talk about the evolution of the slope but not about trends.

These are reasonable concerns, also raised by other two Reviewers, as the inclusion of weak trends can indeed introduce noise to the analysis. considering all trends allows for an assessment of the spatial consistency of directional changes, which is a key observation in our study. If these weaker trends were purely random noise, their distribution would be approximately symmetric, with equal numbers exhibiting positive and negative changes. In contrast, the fact that the weaker trends predominantly align in the same direction suggests that, while they may not meet conventional significance thresholds (e.g., $p = 0.05$), they are not statistically irrelevant. The overwhelming consistency in their direction suggests a potential underlying signal rather mere stochastic variability. Overreliance on statistical significance can lead to rejecting meaningful patterns simply because they do not meet an arbitrary threshold, due to low variability rather than the absence of a real effect.

We also respectfully disagree that in order to talk about a trend it needs to be significant. This is a common misconception in time series analysis, which has risen a lot of criticism in many scientific disciplines. We think the best example is the milestone Editorial in the American Statistican by Ronald L. Wasserstein and Lazar (2019). Among many suggestions about the correct use of statistical significance they state that "no single index should substitute for

scientific reasoning" and caution against the rigid use of p-values as an absolute determinant of scientific conclusions.

Still, we recognize the importance of showing how trend filtering, i.e., statistical significance testing, affects our results. Therefore, we will maintain our original analysis, which includes all trends, while complementing it with the analysis of significant trends in the supplementary material (Figure 2 & Figure 3 shows DCI considering significant trend with p value 0.05). Additionally, we will articulate more clearly the reasons for our decision and describe the impact of trend significance in the revised manuscript.

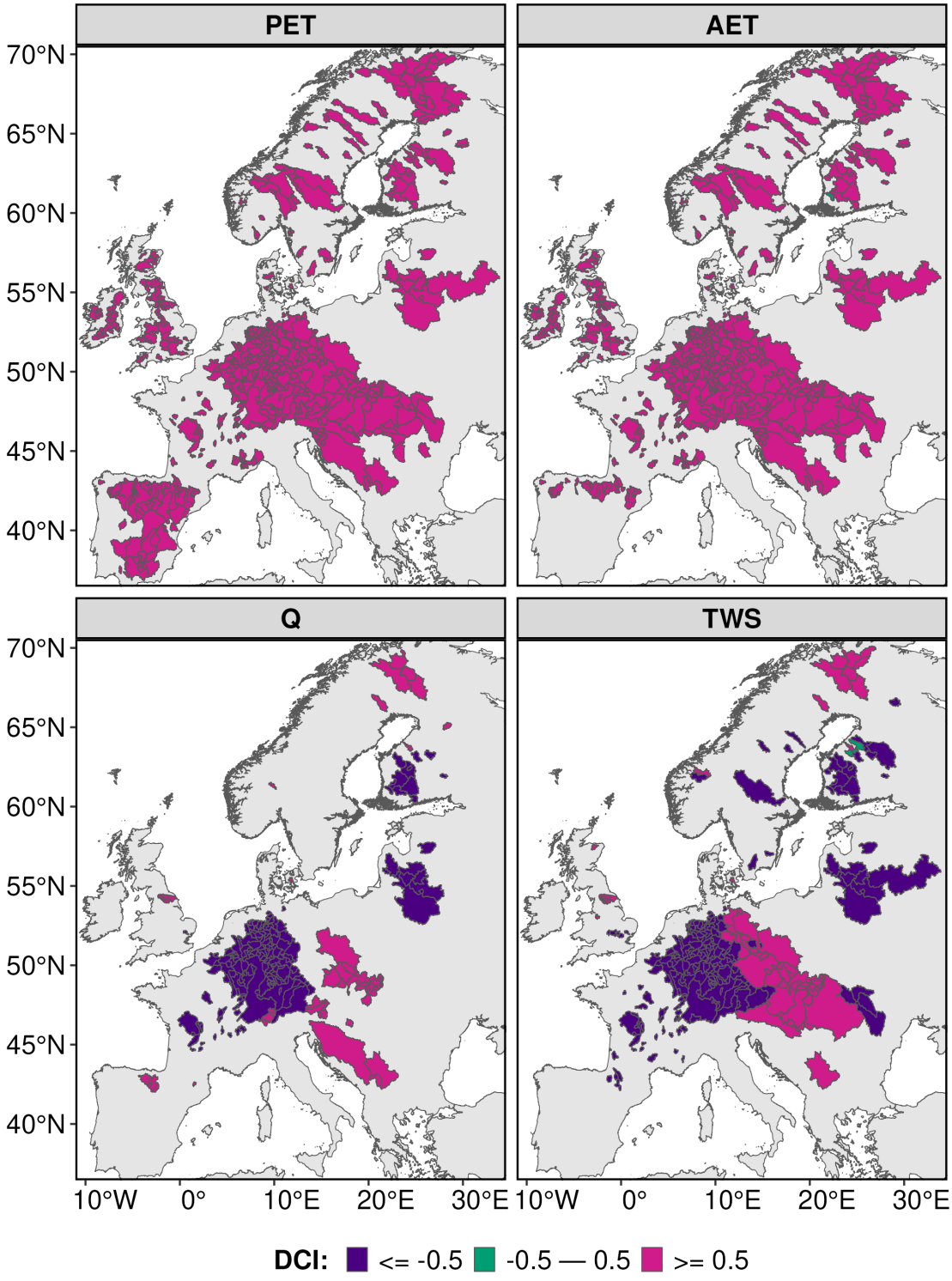

Figure 2. Spatial distribution of annual scale data concurrence index (DCI) for PET, AET, Q, and TWS by considering only significant trend at 95% significance level. PET represents potential evapotranspiration, AET represents actual evapotranspiration, Q represents runoff at the outlet of the catchment and TWS represents total water storage.

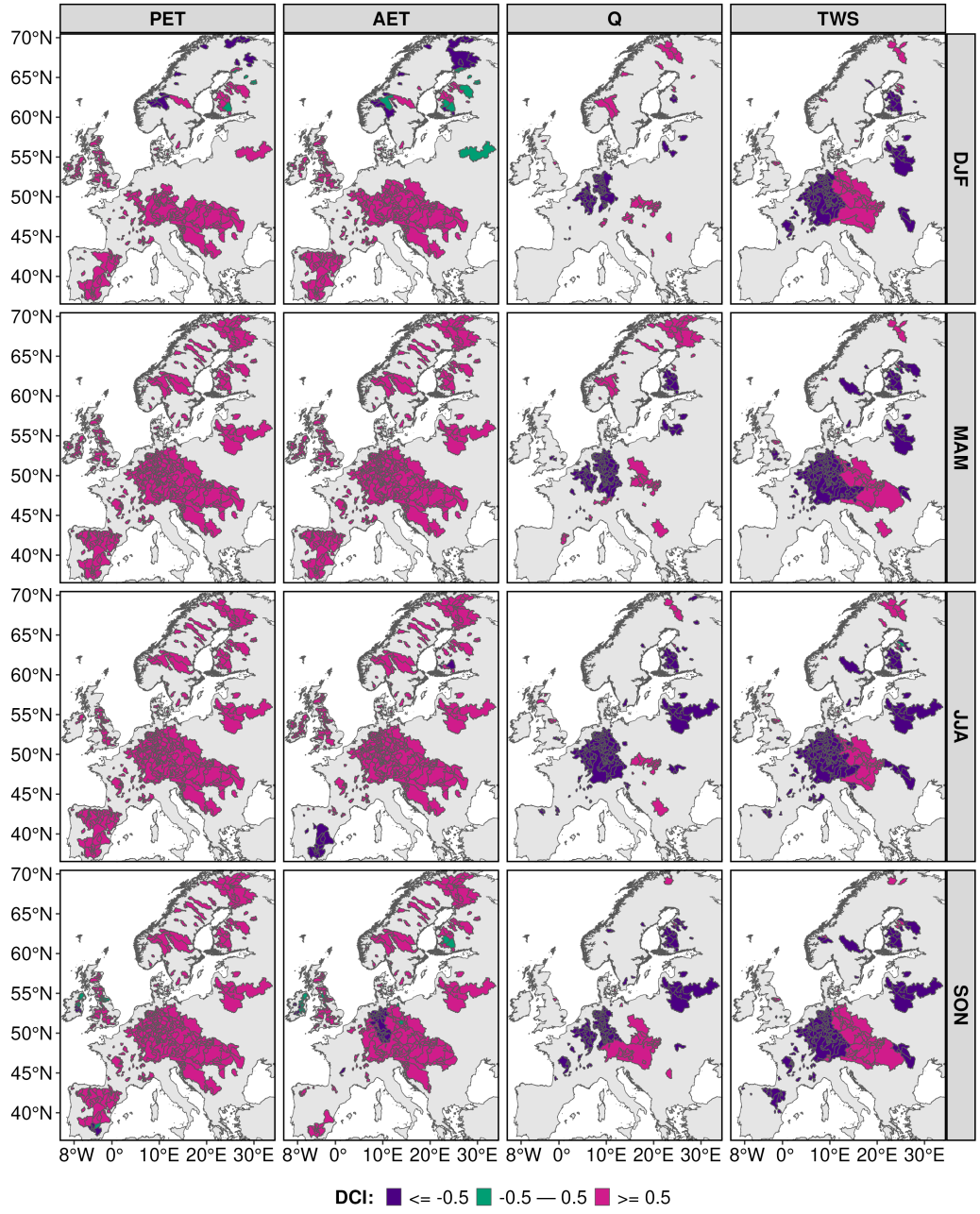

Figure 3. Spatial distribution of seasonal scale data concurrence index (DCI) for PET, AET, Q, and TWS by considering only significant trend at 95 % significance level. PET represents potential evapotranspiration, AET represents actual evapotranspiration, Q represents runoff at the outlet of the catchment and TWS represents total water storage.

4) Results

a) As I stated before, the authors should not talk about trends but rather the evolution of the slope. So, sections 3.1, 3.2, and 3.3 might be rewritten in this sense or carried out the trend analysis and talk about trends.

In our previous response, we acknowledged that the obtained results are meaningful even in the absence of statistical significance. Thus, we will modify these sections to preserve the original findings and integrate additional results corresponding to statistically significant trends.

b) Besides talking about trends, I think it would be nice for the reader to have some information about the actual values, at least for PET. I would include 12 maps, one per PET model selected with actual PET values. It would help the reader to spot differences.

We agree with your comments. We will add the figure in the supplementary material.

5) Discussion

a) There are topics I would highlight in the discussion that it did not. For instance, are the different sizes of the catchments conditioning your results? Another issue I would appreciate including is a discussion about why specific models produce different results when they are within the same category. That is, why are there differences between temperature models? Is it linked to their specific formulation?

Thank you for your constructive comments. Differences in PET models within the same category arise due to variations in their mathematical formulations and underlying assumptions. For instance, in temperature-based methods, Thornthwaite incorporates only the average temperature and was originally developed for humid climates, whereas the Hargreaves-Samani method considers the diurnal temperature range along with the average temperature and was initially formulated for arid and semi-arid climates. Similarly, among combination-type methods, the Penman-Monteith model explicitly accounts for wind speed and vapor pressure deficit, while the Priestley-Taylor method simplifies this by using an empirical coefficient.

Furthermore, we will include a discussion on catchment size conditioning in our revised manuscript.

6) Summary and conclusions

a) Conclusion 7 is obvious. I would remove it.

We agree with your comment; however, Referee #3 recommended incorporating a discussion on precipitation in the main text of the manuscript. Thus, we consider it appropriate to retain this point in the conclusion with necessary modifications, even though it appears obvious.

7) Technical corrections

a) Please homogenize “combination type” vs. “combinational type.”

We will update it in the revised manuscript.

b) Figure 1b: it is unclear what each dot represents. I assume they are catchments, but it seems to me less than 533.

We agree with your comments. It seems there may have been some misunderstanding regarding Figure 1b. To clarify, Figure 1b represents only the test catchments, which are specifically highlighted with thick black borders in Figure 1a. To improve its clarity, the figure (Figure 4) has been further revised based on the comments provided by Referee#3.

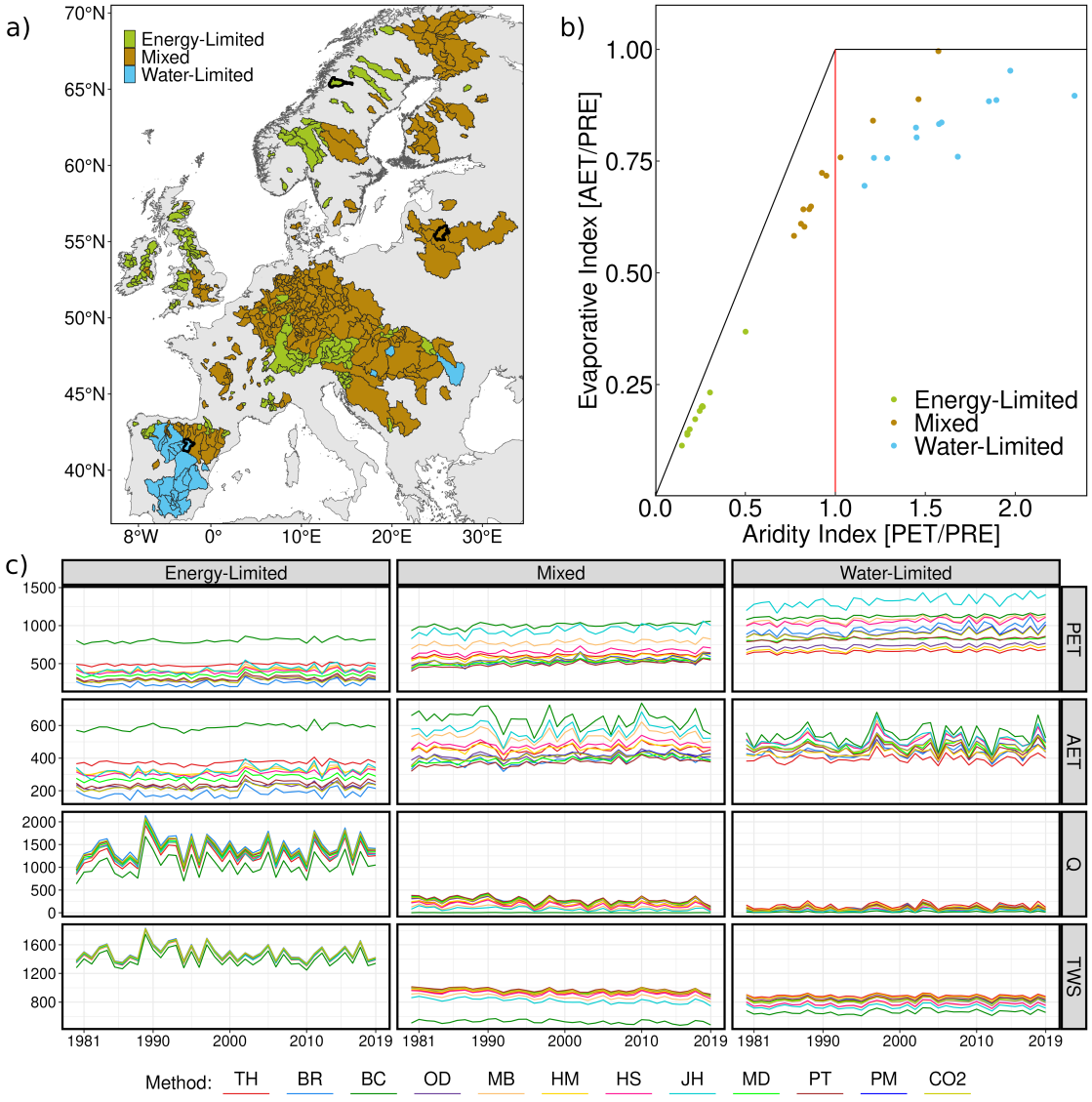

Figure 4. Catchment classification to energy-limited, mixed, and water-limited categories. a) Catchment locations; black borders indicate a representative catchment of each category. b) Classification example within the Budyko space for the representative catchments. c) Annual time series of simulated hydrological components from the mesoscale hydrological model for each representative catchment and PET estimation method (TH: Thornthwaite, BR: Bair-Robertson, BC: Blaney-Criddle, OD: Oudin, MB: McGuinness-Borden, HM: Hamon, HS: Hargreaves-Samani, JH: Jensen-Haise, MD: Milly-Dunne, PT: Priestley-Taylor, PM: Penman-Monteith, CO_2: Modified Penman-Monteith accounts CO_2.). All units are in mm year^{-1}.

c) Line 141: “For the each” should be “for each.”

We will correct it in the revised manuscript.

d) In Figures 2 and 3, I recommend using the same range on the y-axis for each variable to see the differences. In addition, I would add the number of catchments in each category, that is, 189, 330, and 34 for energy-limited, mixed, and water-limited, respectively.

Thank you for highlighting this aspect. A uniform y-axis range across catchment categories would indeed facilitate direct inter-catchment category comparisons. However, our primary goal is to analyze and compare the changes in PET methods within each catchment category individually. Adjusting the y-axis range for each category ensures that trends for PET methods remain distinguishable. Using the same y-axis range would cause trends in energy-limited catchments, which are generally higher, to dominate the visual scale. This would obscure the subtle variations among PET methods in water-limited catchments. Consequently, the comparison within water-limited categories would become less clear and less informative.

e) Figure 6, not all categories sum 553. Please revise. In addition, I think the bar representation with different numbers of catchments and models in each is not intuitive. Maybe the same structure but in a table format for the upper part, in which models are rows, would be more intuitive.

We are agree with your comments. We have addressed the issues related to catchment counts and improvements in figure interaction in Referee #1's comment under Section 4) Results, subsection (g). The discussion is as follows:

In Figure 6, we presented only ten possible hydrological cycle combinations, whereas there are actually 14 possible combinations, as shown in Figure 5. The excluded combinations were not considered because they were represented by only a very small number of catchments. The remaining two catchments from the Thornthwaite method belonged to the (PRE-, AET+, Q+, TWS-) combination. Including these catchments brings the total count to 553 (240 + 156 + 59 + 70 + 24 + 2 + 2). This ensures that all catchments are accounted for in the analysis.

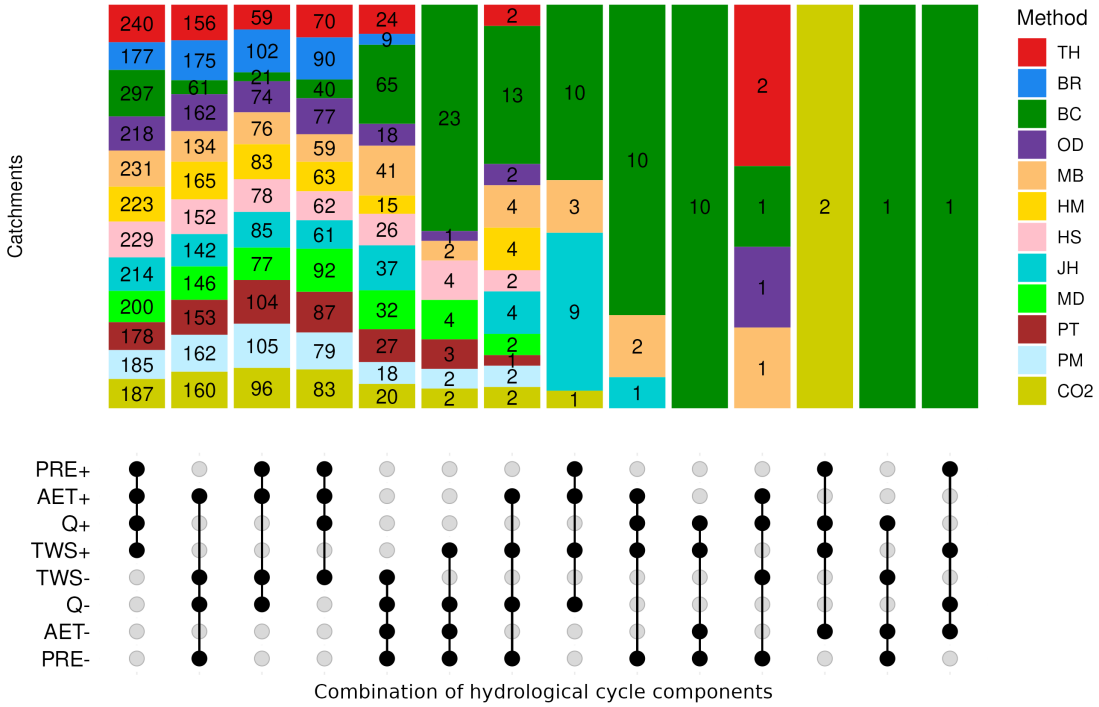

Figure 5. Fourteen distinct combinations of hydrological cycle components and their respective influence of PET methods on an annual scale. PRE+, AET+, Q+, and TWS+ represent an increasing trend for PRE, AET, Q, and TWS respectively. Similarly, PRE-, AET-, Q- and TWS- represent a decreasing trend. Where PRE is precipitation, AET is actual evapotranspiration, Q is runoff and TWS is total water storage. Abbreviations used for different PET methods are TH: Thornthwaite, BR: Bair-Robertson, BC: Blaney-Criddle, OD: Oudin, MB: McGuinness-Borden, HM: Hamon, HS: Hargreaves-Samani, JH: Jensen-Haise, MD: Milly-Dunne, PT: Priestley-Taylor, PM: Penman-Monteith, CO_2: Modified Penman-Monteith accounts CO_2.

f) Please check the name of the PET model throughout the manuscript; there are some inconsistencies.

We will remove these inconsistencies in the revised manuscript.

References

Boeing, F., Wagener, T., Marx, A., Rakovec, O., Kumar, R., Samaniego, L., and Attinger, S.: Increasing influence of evapotranspiration on prolonged water storage recovery in Germany, Environmental Research Letters, 19, 024 047, https://doi.org/10.1088/1748-9326/ad24ce, 2024.

Kumar, R., Livneh, B., and Samaniego, L.: Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme, Water Resources Research, 49, 5700–5714, 2013.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., Attinger, S., Schäfer, D., Schrön, M., and Samaniego, L.: Multiscale and Multivariate Evaluation of Water Fluxes and States over European River Basins, Journal of Hydrometeorology, 17, 287–307, https://doi.org/10.1175/JHM-D-15-0054.1, 2016.

Ronald L. Wasserstein, A. L. S. and Lazar, N. A.: Moving to a World Beyond "p < 0.05", The American Statistician, 73, 1–19, https://doi.org/10.1080/00031305.2019.1583913, 2019.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, https://doi.org/10.1029/2008WR007327, 2010.

Samaniego, L., Thober, S., Wanders, N., Pan, M., Rakovec, O., Sheffield, J., Wood, E. F., Prudhomme, C., Rees, G., Houghton-Carr, H., Fry, M., Smith, K., Watts, G., Hisdal, H., Estrela, T., Buontempo, C., Marx, A., and Kumar, R.: Hydrological Forecasts and Projections for Improved Decision-Making in the Water Sector in Europe, Bulletin of the American Meteorological Society, 100, 2451–2472, https://doi.org/10.1175/BAMS-D-17-0274.1, 2019.

Tang, G., Clark, M. P., and Papalexiou, S. M.: EM-Earth: The Ensemble Meteorological Dataset for Planet Earth, Bulletin of the American Meteorological Society, 103, E996–E1018, https://doi.org/10.1175/BAMS-D-21-0106.1, 2022.

Tang, G., Clark, M. P., Knoben, W. J. M., Liu, H., Gharari, S., Arnal, L., Beck, H. E., Wood, A. W., Newman, A. J., and Papalexiou, S. M.: The Impact of Meteorological Forcing Uncertainty on Hydrological Modeling: A Global Analysis of Cryosphere Basins, Water Resources Research, 59, e2022WR033 767, https://doi.org/10.1029/2022WR033767, 2023.

Thornthwaite, C. W.: An Approach toward a Rational Classification of Climate, Geographical Review, 38, 55, https://doi.org/10.2307/210739, 1948.

Xiang, K., Li, Y., Horton, R., and Feng, H.: Similarity and difference of potential evapotranspiration and reference crop evapotranspiration – a review, Agricultural Water Management, 232, 106 043, https://doi.org/10.1016/j.agwat.2020.106043, 2020.

Yin, Z., Lin, P., Riggs, R., Allen, G. H., Lei, X., Zheng, Z., and Cai, S.: A synthesis of Global Streamflow Characteristics, Hydrometeorology, and Catchment Attributes (GSHA) for large sample river-centric studies, Earth System Science Data, 16, 1559–1587, https://doi.org/10.5194/essd-16-1559-2024, 2024.