Reply to Referee's comments

Referee #1: Clerc-Schwarzenbach, Franziska

Referee's comments are in black text

Authors's response are in blue text

1) General Comments: Thank you for giving me the opportunity to review the manuscript "Unveiling the impact of potential evapotranspiration method selection on trends in hydrological cycle components across Europe" by Vishal Thakur and co-authors. I enjoyed reading this manuscript and consider the findings to be valuable and highly relevant: Even though potential evapotranspiration is a crucial component of the hydrological cycle, factors like data availability or convenience often play a role in the selection of a potential evapotranspiration formulation. There is no uniform way of dealing with potential evapotranspiration in hydrology, and often, different concepts (such as reference evapotranspiration and potential evapotranspiration) are used interchangeably. Therefore, studies like the one by Vishal Thakur et al., analysing the effects of different formulations of potential evapotranspiration are of great value to shed light on this often-neglected topic. The manuscript presents an analysis of the influence of different potential evapotranspiration formulations on different simulated components of the hydrological cycle. To do so, the authors make use of a large-scale modelling approach and analyse the modelling results with effective methods. They come to the important conclusion that the choice of a potential evapotranspiration method has an influence on the results when studying the hydrological cycle and its components.

   We would like to deeply thank you for your constructive feedback, which plays a key role in improving the quality of our manuscript. The depth and thoughtfulness of your review have greatly aided us in presenting our findings more effectively and aligning them with broader scientific discussions. We deeply appreciate your time, effort, and dedication in offering such a thorough and thoughtful review, which has significantly enhances the quality of our manuscript.

My main concern is that the majority of the chosen formulations (or methods) are temperature-based. Since temperatures were rising in the studied period between 1980 and 2019, the potential evapotranspiration methods based on temperature do show a positive trend. So far, it is not clear if temperature-based methods are still reliable under the conditions of a warming climate (see for example studies on the so-called "pan evaporation paradox": Li et al. (2013, 10.1002/wrcr.20202); Wang et al. (2017,10.1002/wat2.1207)). With eight out of twelve methods being temperature-based, a possible overestimation may strongly influence the results of this study.

We agree that our study includes more temperature-based PET methods compared to radiation-based and combination methods. This imbalance in the number of methods could potentially influence the Data Concurrence Index (DCI) analysis. To address this issue, we randomly selected four temperature-based methods, resulting in 70 combinations when paired with the other PET methods (radiation and combination methods). Upon analyzing these 70 combinations, PET consistently exhibited strong positive DCI (DCI $\geq$ 0.5) across all 553 catchments. For AET, the median number of catchments showing strong positive DCI was 535 $\pm$ 10, while disagreement (DCI between -0.5 and 0.5) was observed in 17 $\pm$ 10 catchments. For Q, 269 $\pm$ 3 catchments showed strong negative agreement (DCI $\leq$ -0.5), 28 $\pm$ 3 catchments exhibited disagreement, and 256 $\pm$ 4 catchments displayed strong positive agreement. Similarly, for TWS, 368 $\pm$ 6 catchments showed strong negative agreement, 53 $\pm$ 7 exhibited disagreement, and 132 $\pm$ 4 demonstrated strong positive agreement. The analysis of these 70 different combinations (four temperature-based methods and four complex methods) suggests that only a small number of catchments are affected in the DCI analysis. Out of 70 two combinations are shown in Figure 1 and Figure 2. However, the trend comparisons at annual and seasonal scales (Sections 3.1 and 3.2) remain unaffected by the varying number of PET method categories, as trends were evaluated for individual methods rather than aggregated categories. We will incorporate this information into the supplementary material in the revised manuscript.

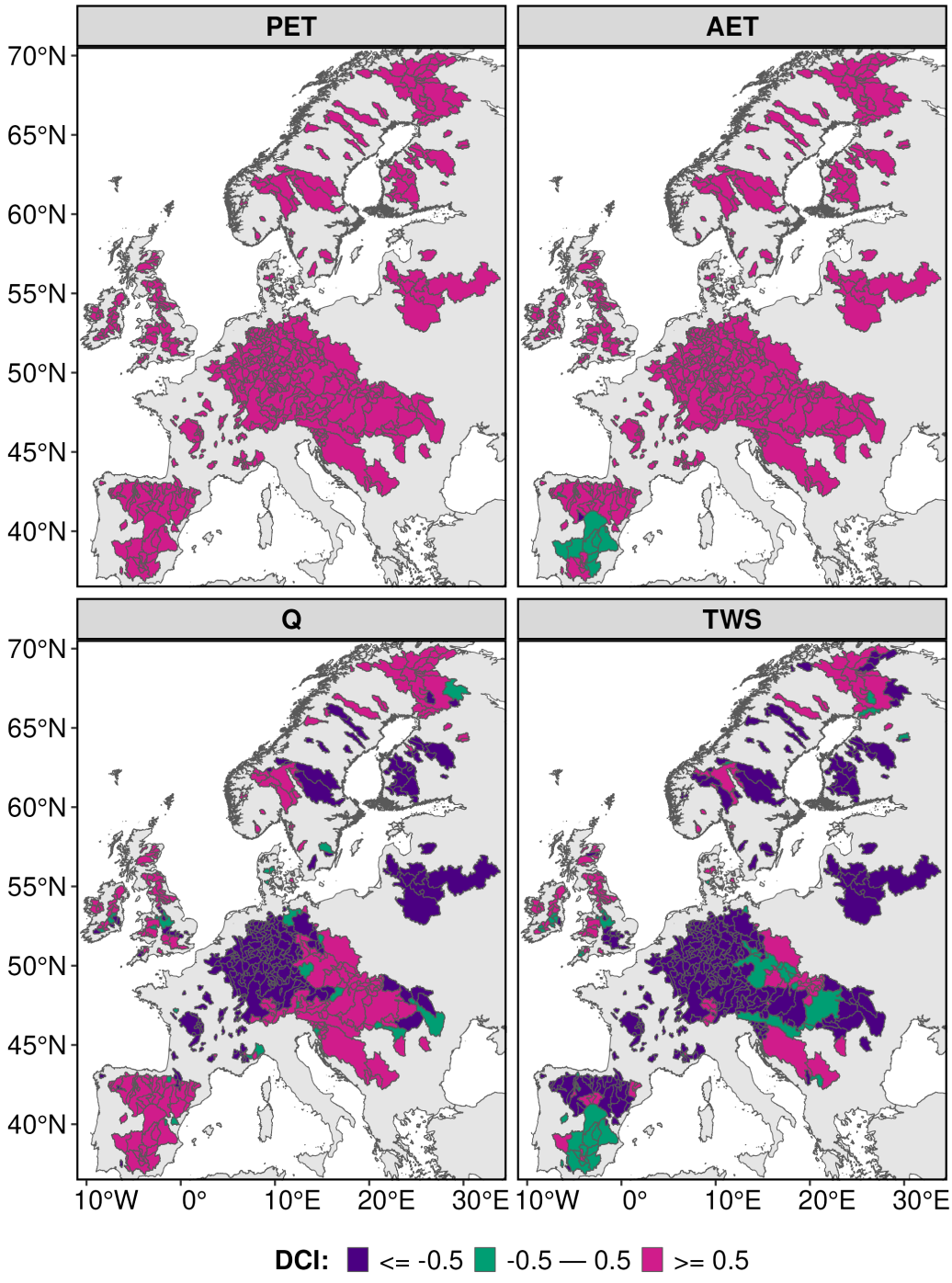

Figure 1. Spatial distribution of annual scale data concurrence index (DCI) for PET, AET, Q, and TWS. Considering eight PET methods namely Blaney-Criddle, Bair-Robertson, Hamon, Hargreaves-Samani, Milly-Dunne, Priestley-Taylor, Penman-Monteith and Modified Penman-Monteith accounts CO2.

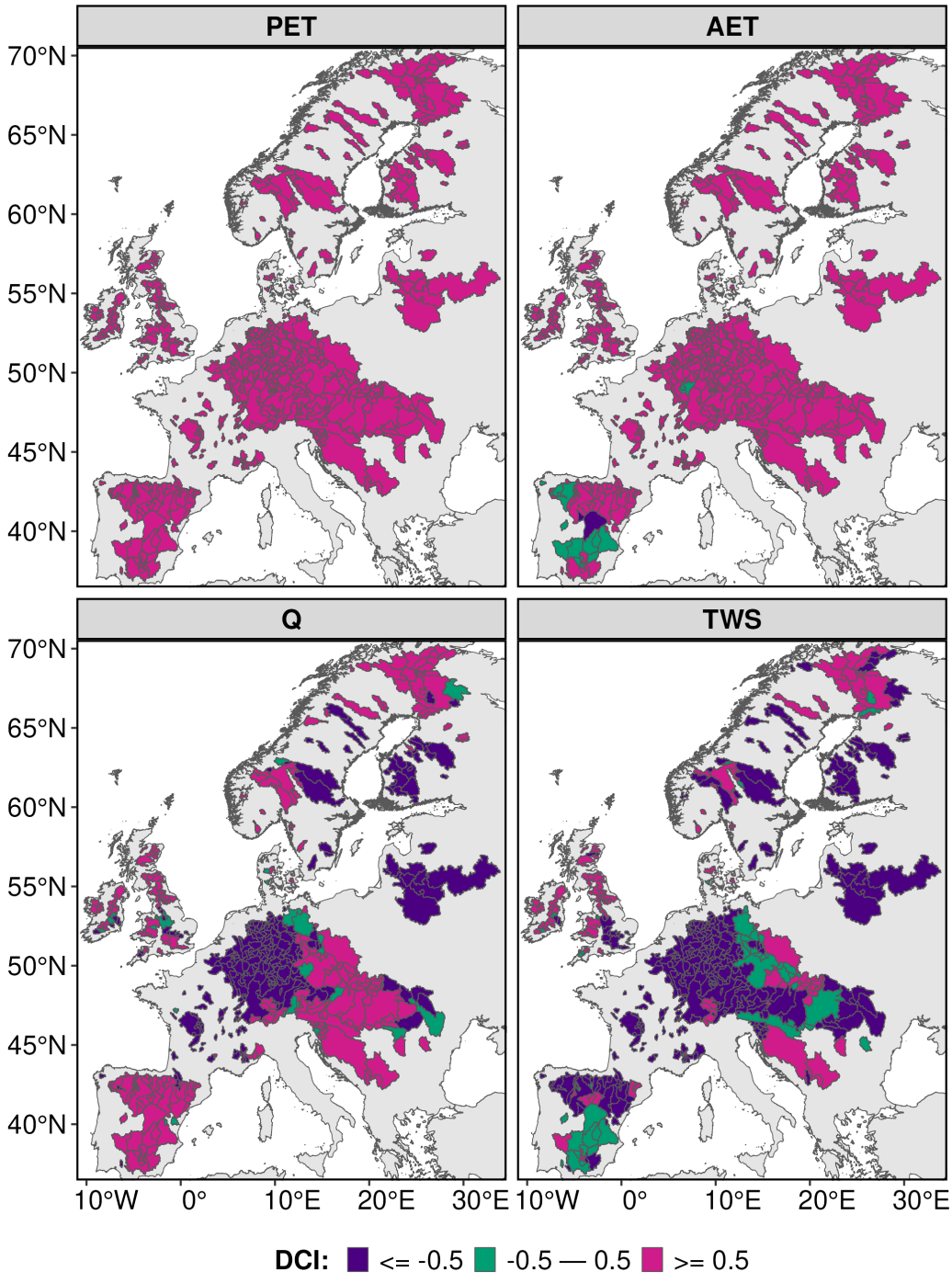

Figure 2. Spatial distribution of annual scale data concurrence index (DCI) for PET, AET, Q, and TWS. Considering eight PET methods namely Jensen-Haise, McGuinness–Bordne, Oudin, Thornthwaite, Milly-Dunne, Priestley-Taylor, Penman-Monteith and Modified Penman-Monteith accounts CO2.

Specific comments and suggestions that will hopefully help to improve the different parts of the manuscript are listed below.

2) Introduction

a) While Thornthwaite (1948) was the first to introduce the term “evapotranspiration”, according to Miralles et al. (2020, 10.1029/2020WR028055), the concept itself was already used earlier. This should be specified. Furthermore, as the different concepts regarding evaporation (including “reference crop evapotranspiration”) are often used interchangeably in hydrology, I would suggest to add a statement about this problem, e.g., referring to the Miralles et al. paper mentioned above as well as the already cited paper by Xiang et al. (2020). In addition, I would argue that there should be some indication if you consider evapotranspiration to include interception or not in your study. See also the commentary by Savenije (2004, 10.1002/hyp.5563).

Thank you very much for your fruitful comments. We have elaborated development of potential evapotranspiration and added a small introduction about reference evapotranspiration. Modified sentences are 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).

mHM model incorporates the interception process, where a portion of actual evapotranspiration (AET) results from interception evaporation. This process is estimated as a fraction of potential

evapotranspiration (PET) using a power function derived from Deardorff (1978) and Liang et al. (1994). Furthermore, if evaporation exceeds the intercepted water, the interception storage is completely depleted.

b) In the first sentence of the paragraph starting at line 41, PET is stated to directly influence AET. In the second sentence, an alternative to this direct influence is presented. Therefore, the first sentence should be adjusted so that it becomes clear that this is just one possibility. Furthermore, depending on how interception is dealt with, this may be a necessary component to add in the list of fluxes that make up AET (in the second sentence of this paragraph).

We agree with your comment. We will update line 41 and provide a detailed description of the components contributing to AET.

3) Methods and data

a) In the beginning of section 2.2, you give the study by Hersbach et al. (2020) as a reference for ERA5-Land. I would argue that the suitable reference there is the paper by Muñoz-Sabater et al. (2021), that you cite later in the manuscript. Furthermore, in line 97, there is a reference missing for EM-Earth and SC-Earth, or it is not clear to me if the Tang et al. (2022) reference belongs to this. If the EM-Earth data are based on ERA5 data (and not on ERA5-Land data), I would suggest to include the reference to the Hersbach et al. (2020) paper there.

We will correct reference typos; the reference (Muñoz-Sabater et al.; 2021) to ERA5-Land will be corrected in the revised manuscript. The EM-Earth dataset is distinct and primarily generated by integrating multiple data sources, with SC-Earth and ERA5 being the two major contributors. To improve clarity, a reference will be added to line 97. *The EM-Earth dataset (Tang et al., 2022) is derived from observed station data SC-Earth (Tang et al., 2021) and ERA5 data (Hersbach et al., 2020).*

b) I assume the AET, Q, and TWS data in Figure 1c to be simulated values. In my opinion, this needs to be mentioned in the figure caption as it only becomes clear after reading further.

The caption for Figure 1(c) has been revised for improved clarity as follows: *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) In line 122, you state that some of the temperature-based methods also use some extraterrestrial radiation term calculated based on latitude. Potentially, this could already be mentioned earlier, when the definition of the temperature-based category is given (i.e., it could become clearer that also radiation terms can be required and a method is still considered to be temperature-based).

For better clarity entire paragraph (Line 117 - 129) was rewritten. *We incorporate 12 PET methods at a daily scale from all three categories of estimation: temperature-based, radiation-based, and combination-based methods (Table 2). Temperature-based methods require temperature data, which can include average temperature, minimum temperature, or maximum temperature. Additionally, PET methods that incorporate extraterrestrial radiation are also considered under this category. Combination-based methods require a larger number of variables compared to temperature- and radiation-based methods to estimate various physical terms, such as wind speed and surface pressure (Table 2). Most temperature-based methods use only daily average temperature (Thornthwaite, Oudin, Hamon, Jensen-Haise, McGuinness-Bordne, and Blaney-Criddle), while Baier-Robertson employs both minimum and maximum daily temperatures. Some of these methods also include an extraterrestrial radiation term in their formulation. However, as this radiation term is calculated based on latitudinal information, only temperature data is required for PET calculation. We utilize only one radiation-based method, Milly-Dunne PET, which requires only net radiation data to estimate PET. The combination-based methods, such as Penman-Monteith and Priestley-Taylor, have a stronger*

physical basis. In our analysis, all these physical terms are estimated following Allen (1998). Additionally, within the combination category, we employ the modified Penman-Monteith (CO2) method, which accounts for temporal variations in changing carbon dioxide concentrations. Formulation details, including mathematical equations and associated constants for each PET method, are provided in Table A1.

d) When you list the terms that are required for the Penman-Monteith and the Priestley-Taylor method, I would suggest to list all the terms (as it's only two equations), instead of giving some examples and concluding with "etc.". Alternatively, you could give one or two examples and then place a reference to the table where you list all the input data required for each method.
We have incorporated two examples and provided a reference to the relevant table in the manuscript. The modified version of the paragraph has been addressed in subsection(c) of section 3) Methods and data of response letter.

e) For me, the description of the modelling part (paragraph starting on line 139) is hard to follow. Therefore, I ask you to give more information about this part of your study: You do one model run per catchment and PET method (553 x 12 = 6636). Is there any model calibration? If yes, please elaborate on that: What objective function(s) did you use? What was the spatial resolution? If no, where do you take the model settings from? Are these the same for all catchments? How do you make sure that the parameterization that you use matches your catchments? Did you think about equifinality and how other possible parameterizations could influence your results? How well did the model perform for the different catchments? In the very end of the discussion, the reader learns that a default parameterization was used. Please add this to the methods part and make sure that the questions listed above become clear.
We agree with your comments. We address them 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 x 12). The mHM model was run at a daily time step

with a spatial resolution of 0.125° x 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 3. 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. We will add its details in the revised manuscript.

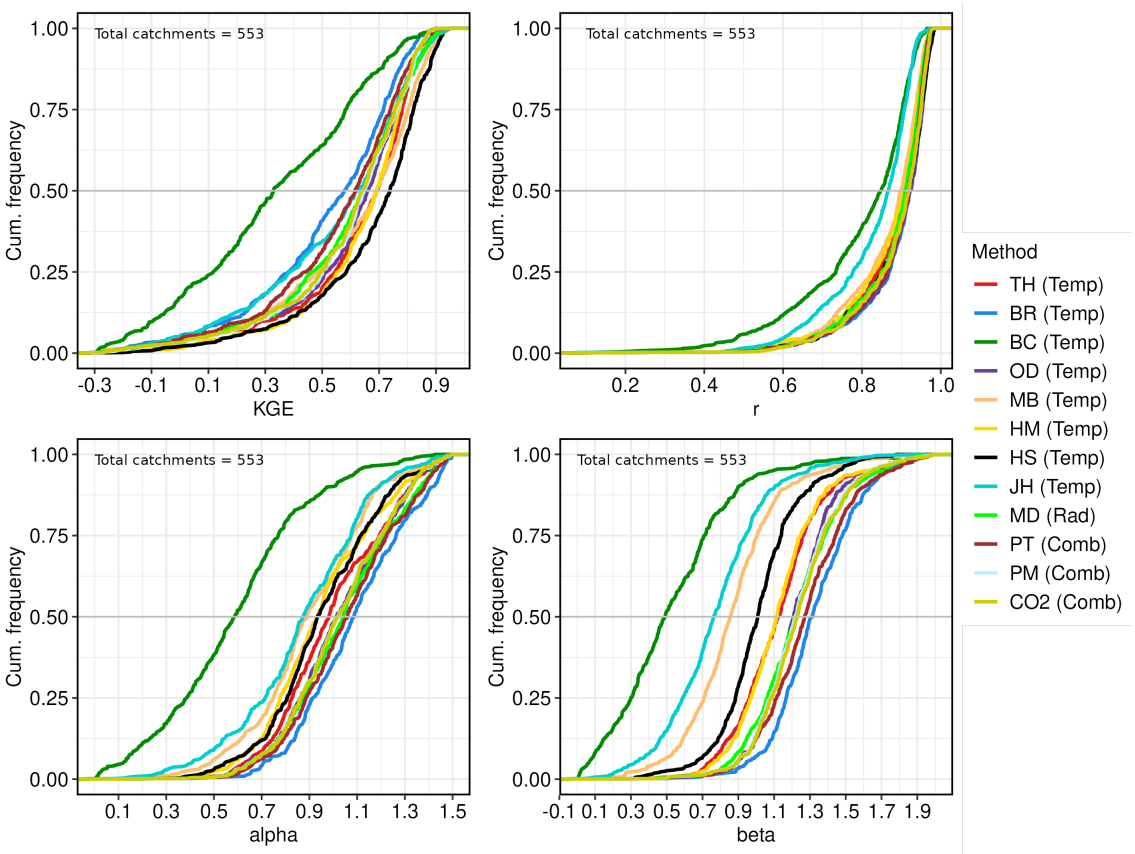

Figure 3. Evaluation of the hydrological model (mHM) performance using simulated monthly streamflow across 553 catchments, forced with EM-Earth meteorological data. The figure presents the cumulative frequency distributions of the Kling–Gupta efficiency (KGE) and its three components: correlation (r), variability ratio (alpha), and bias ratio (beta), providing insights into the model's performance across different PET methods.

f) Related to the comment above, where does this default parameterization come from? Was there a method for potential evapotranspiration involved when this default parameterization was obtained? If so, may this have varying effects on the results based on this (or similar) methods (that can potentially profit from compensating effects) and the results based on different methods (that cannot profit from any compensating effects)? Potentially and if possible (may be limited due to data availability constraints) it would be interesting to compare the current results to the results of a calibrated model (i.e., calibrated for each of the different methods) – you also include a similar remark in the discussion. With additional calibrations that make use of the different

methods, it could become clearer if the (assumed) use of one method for the default parameterization affects the results. If additional calibrated model runs are not possible, I would suggest you to include some text on the potential effects of the default parameterization in the discussion.

We acknowledge your comments. 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.

g) I suggest you to clarify in lines 158 and 159 that you compare winter data to winter data, spring data to spring data, etc. so that it is immediately clear that you do not compare all four seasons with each other.

Following your comments, Lines 158–1599 were further elaborated as follows: *The slopes of each PET method were analyzed exclusively within seasons (for eaxmple winter season compared with winter season), without cross-seasonal comparisons.*

h) In section 2.3.4, you describe the modification of the DCI by including also non significant trends to be able to include all trend estimates. Are you sure that you are not including too much noise

by also counting very weak positive and negative trends? I suggest to at least show that your findings do not change if only reasonably significant trends are considered.

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 4 & Figure 5 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.

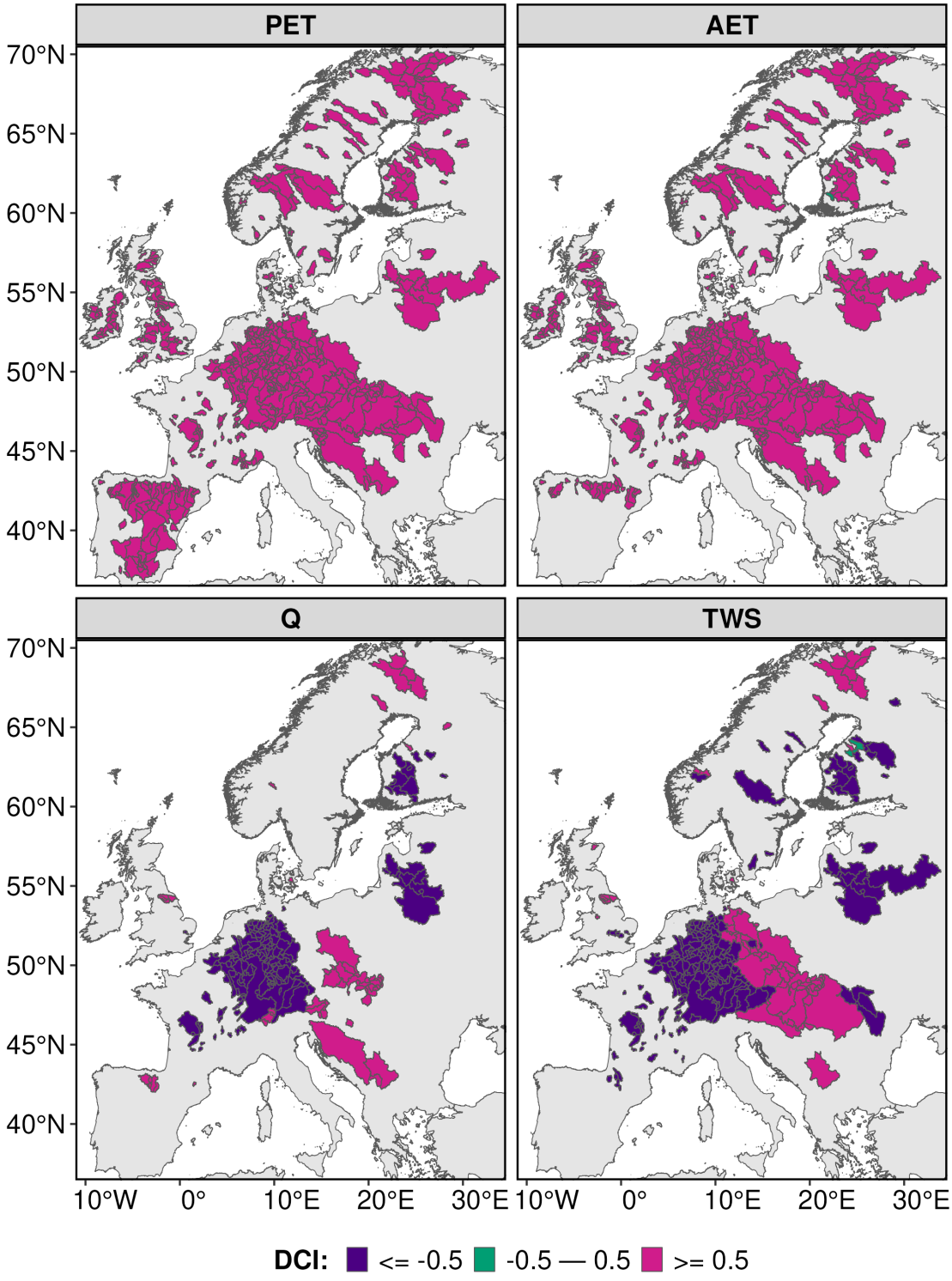

Figure 4. Spatial distribution of annual scale data concurrence index (DCI) for PET, AET, Q, and TWS by considering only significant trend at 95% significance level. PET represents potential evapotranspiration, AET represents actual evapotranspiration, Q represents runoff at the outlet of the catchment and TWS represents total water storage.

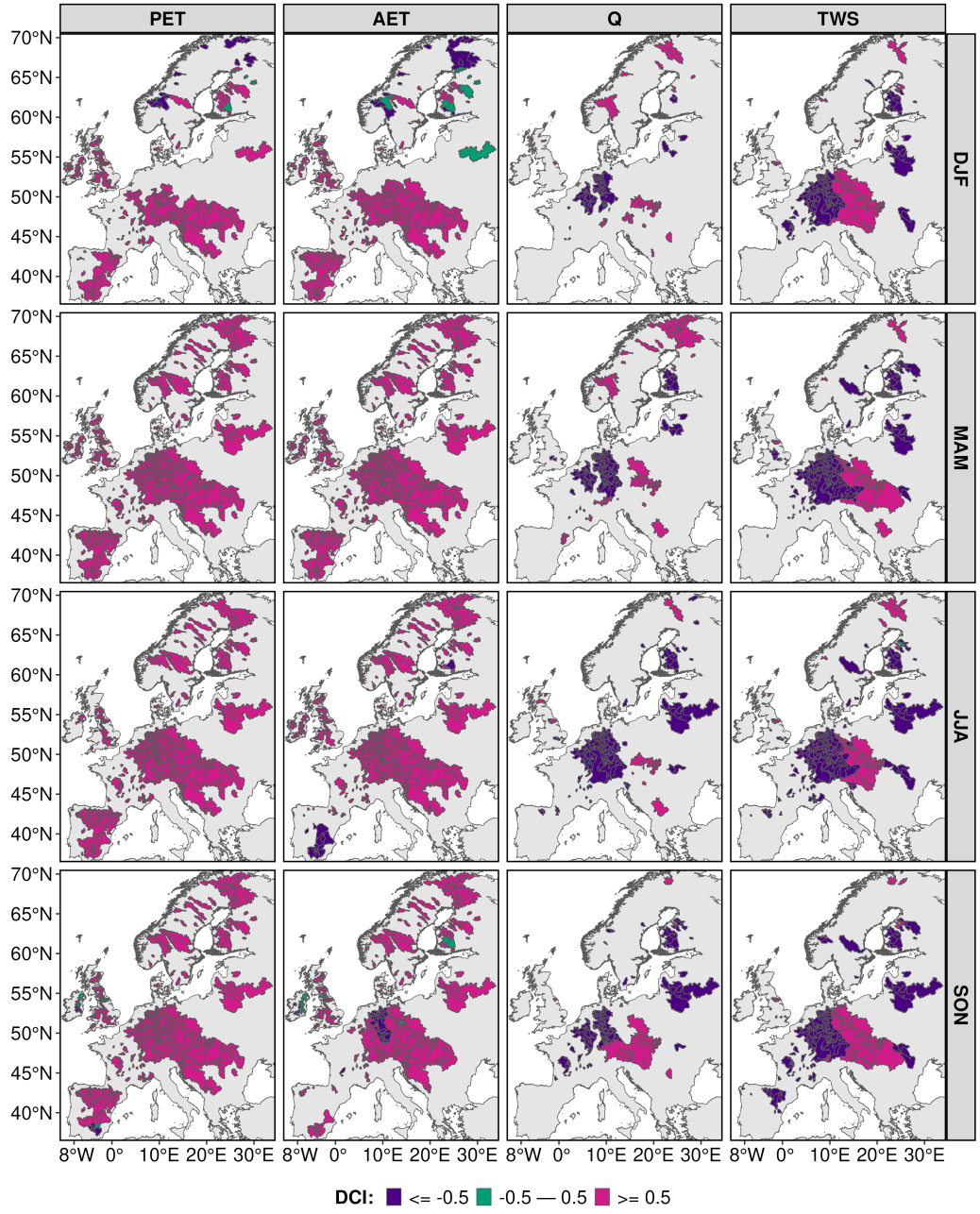

Figure 5. Spatial distribution of seasonal scale data concurrence index (DCI) for PET, AET, Q, and TWS by considering only significant trend at 95 % significance level. PET represents potential evapotranspiration, AET represents actual evapotranspiration, Q represents runoff at the outlet of the catchment and TWS represents total water storage.

4) Results

a) In the first paragraph of section 3.1, you describe the different trends in PET for the three categories of catchments. As PET describes a potential, i.e., the maximum ET that could be achieved if there was no water-limitation, the trend in PET should not depend on if a catchment is waterlimited or energy limited. Thus, I suggest to formulate this paragraph differently, such that it is clear that the PET trend is influenced by other factors, but not by the availability of water.
We will revise this paragraph by linking the factors influencing PET and relating them to the context of catchment types. We will update this paragraph and incorporate into the revised manuscript.

b) In the last paragraph of section 3.2, you suddenly write about December and not about the winter season (line 234). If you did monthly analyses, this needs to be stated clearly in the methods section. Otherwise, December may be the wrong term here.
Yes, you are right. We made a typing mistake in this line. The paragraph was intended for seasonal comparison. Instead of "December," it should refer to the "winter season." The corrected sentence for lines 234–235 is as follows: *In contrast, for energy-limited catchments, the winter and summer seasons are the main contributors, with their impacts varying depending on the selected PET method (Figure S13)*

c) As already mentioned in the general comments: With the rising temperatures, all the temperature-based methods tend to show positive trends. Since most PET methods included are temperature-based, they have the largest influence on the DCI. It would be good to test if this unequal distribution of the different types influences your results.
We agree that our study has an unequal distribution of PET methods across different categories. To address this, we conducted an analysis using an equal distribution of PET methods, selecting four temperature-based methods and four complex methods. A detailed discussion of the observed result is provided in 1) General Comments.

d) To make the part about the combinations of hydrological cycle component changes (section 3.4) easier to understand also before studying Figure 6, I suggest you to add an example of a "possible combination of hydrological cycle component changes" after you introduce this term on lines 269 and 270. Similarly, it may be good to inform the reader what you mean with "the first five hydrological cycle combinations" (lines 270 and 271).

We agree with your comments. For better clarity example of combination hydrological cycle will be added. *Here, we demonstrate the overall impact of PET methods on possible combinations of hydrological cycle component changes. For instance, one hydrological cycle combination involves an increase in all components (PRE+, AET+, Q+, TWS+). Conversely, another combination represents decreasing all components (PRE-, AET-, Q-, TWS-).*

For a better understanding of lines 270 and 271 additional text will be added to revised text. *These five combinations are as follows: (1) PRE+, AET+, Q+, TWS+; (2) PRE-, AET+, Q-, TWS-; (3) PRE+, AET+, Q-, TWS-; (4) PRE+, AET+, Q+, TWS-; and (5) PRE-, AET-, Q-, TWS-.*

e) In lines 272 and 273 you state that the temperature-based methods account for more catchments with positive trends across all hydrological cycle components than combinational methods (and you put this in comparison to the Blaney-Criddle method leading to positive trends in all components for most catchments). This statement is not clear: Do you want to say that the temperature-based methods lead to the "all positive" combination in more catchments than the combinational methods? If so, this is not true for the Baier-Robertson method, if I interpret Figure 6 correctly. Please reformulate this statement and reconsider if this should be formulated as a comparison to the first part of the sentence.

We will revise the sentence to improve its clarity.

f) The sentence regarding the combinations with AET+ and TWS- (lines 280 and 281) should come before the statement about the last five combinations to be consistent with the order in Figure 6.

We will move the sentence to the appropriate place in revised manuscript.

g) I think that Figure 6 should be improved so that it can be grasped more quickly and easily:

We acknowledge your comments regarding Figure 6. We used a specific R package to generate this figure, which has certain limitations. However, we will make every effort to implement the suggested modifications to improve its clarity and interactivity. Additionally, we will include explanatory text to help readers better interpret the figure. In the worst-case scenario, if the necessary improvements cannot be made, we may consider removing the figure from the manuscript.

- I suggest you to put the lower part (in the following called "combinations") to the top and the upper part (in the following called "catchment count") to the bottom. The reader first wants to know what combination we are looking at, and then wants to look at the results for this combination. I see that you used the "combinations" part as an "axis label", but this is not so clear when looking at the plot in the beginning as it looks more like two different parts.

  We will make the changes accordingly in the revised manuscript.

- For the "combinations", consider displaying them differently. For example, a simple table-like graphic in which you could even work with additional colours would in my opinion be easier to interpret than the current way to display the combinations, see below. If you decide to do so, do not forget to change the caption.

  We agree with your comment. As an alternative for better clarity, we will incorporate color in the combinations (for example, blue for a positive trend and brown for a negative trend).

- For the "catchment count", I think it would improve the readability of the figure if not all bars were the same height, i.e., if the height of the bar would decrease from left to right, proportionally to the higher number of catchments that are contained in the bars on the left than in the ones on the right. This way, it would become clearer which combinations occur how often. For that, you could consider using horizontal instead of vertical bars to gain more

space (this would also allow you to use the “combinations” part as “axis labels”, but then of the vertical axis).

Thank you very much for your constructive suggestion. We will adjust the height of the bars accordingly. However, doing so may make it challenging to display the catchment numbers for combinations with a lower catchment count.

- Please double-check the catchment count for each method. To my understanding, the number of catchments per method contained in this plot should always sum up to 553. However, for example for the Thornthwaite method, the number of catchments only sums up to 240+156+59+70+24+2=551.

  We agree with your comment. In Figure 6, we presented only ten possible hydrological cycle combinations, whereas there are actually 14 possible combinations, as shown in Figure 6. 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.

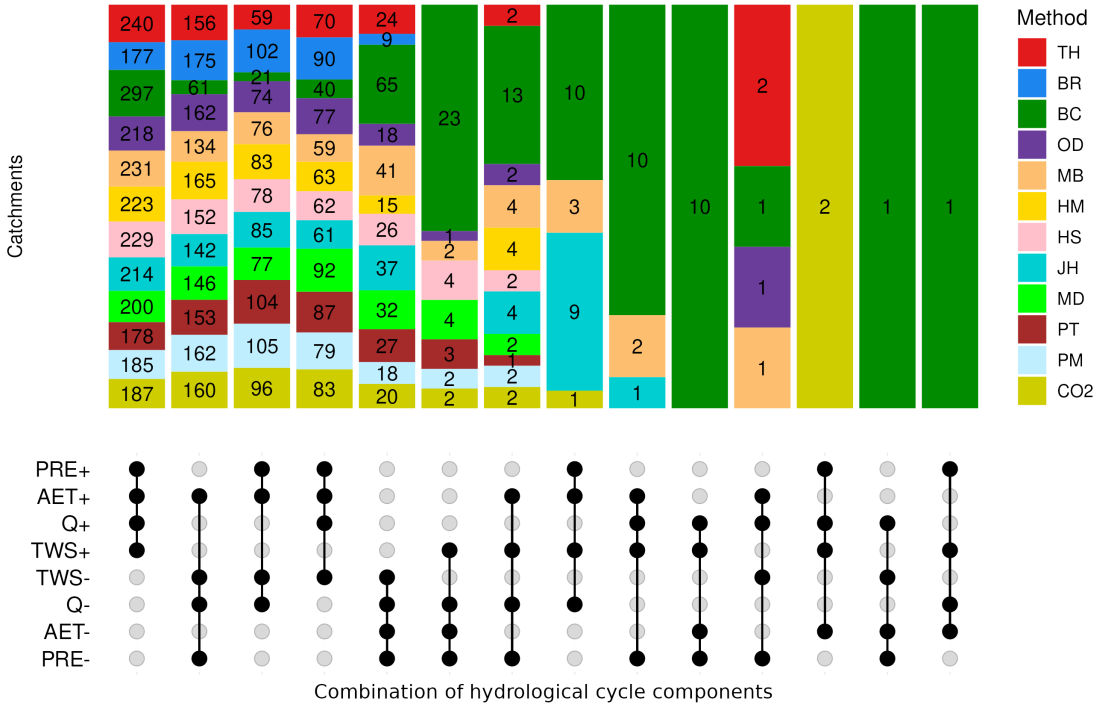

Figure 6. Fourteen distinct combinations of hydrological cycle components and their respective influence of PET methods on an annual scale. PRE+, AET+, Q+, and TWS+ represent an increasing trend for PRE, AET, Q, and TWS respectively. Similarly, PRE-, AET-, Q- and TWS- represent a decreasing trend. Where 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.

- As the Thornthwaite method and the Blaney-Criddle method are displayed in neighbouring fields for the seventh combination, please consider not using red and green for these two methods as they are hard to distinguish for colour-blind people. Alternatively, please use patterns in addition, to make it possible to distinguish the two colours.
- We will update the colors in the revised manuscript.

5) Discussion

a) In line 302, you compare your results to the ones in the study by Hanselmann et al. (2024).Please note that while the authors of this study are affiliated in Poland, the research has been conducted for Spitsbergen.

We will update the discussion with a suitable reference.

b) In the paragraph starting on line 336, you discuss your results as well as the results of the study by Teuling et al. (2019). Based on the Teuling et al. paper as well as based on the PET methods that you considered in your study, I assume that you mean the Penman-Monteith method in this paragraph when you write about the Penman method. Please double-check and correct. The same issue occurs again in line 384.

Yes you are right. We will correct the typo penman to Penman-Monteith.

c) Later in the same paragraph, where you discuss a possible drying hydrological cycle, I think that a discussion of the study by Milly & Dunne (2016) that you refer to elsewhere in your manuscript is lacking as they studied this topic.

We agree with your comments. Our intention is to focus on a specific combination of hydrological cycle components. To enhance clarity, a minor modification will be made to lines 340–345.

Similarly, many catchments demonstrate positive changes in AET and negative changes in Q, TWS, and PRE (drying condition of the hydrological cycle). The number of catchments demonstrating these trends varies depending on the choice of PET method. Massari et al. (2022) found that over Europe runoff deficit are more pronounced in water-limited regions due to increased AET, whereas energy-limited catchments exhibit smaller deficits. During these drying conditions, AET is further influenced by reductions in TWS (Massari et al., 2022).

d) In line 368, you write about methods that consistently overestimate PET. It is unclear to me how you define overestimation here: Do you just consider the methods leading to the highest estimates to be overestimating PET? If so, I would not consider it to be surprising that the catchments shift to the energy-limited category (as all the higher estimates are excluded). If all the low estimates

would be excluded, the catchments would probably shift to the water-limited category. Please elaborate more on what you mean there.

In line 368, methods that consistently overestimate PET refer to those producing higher estimates compared to other PET methods. Specifically, the Jensen-Haise and Blaney-Criddle methods show higher PET values than other methods. Table S1 in the manuscript illustrates catchment shifts when different PET methods with higher estimates are excluded.

e) When you discuss the limitations of your study, you state that on the one hand, the PET methods were found to be sensitive to the input data, and on the other hand, that hydrological models are sensitive to the precipitation input. Based on this, I think it would be good to write something about the data quality of the input data that you used (as you basically state that your results are highly sensitive to these data).

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).

6) Summary and conclusions

a) In the very last part of the summary, you recommend an ensemble of PET formulations instead of one single formulation. While agreeing with this statement, I think that it should be supported by the study and not only occur in the end without being mentioned before. Thus, I suggest you to

use the ensemble of the different methods as a thirteenth option of the PET calculation in your manuscript.

We appreciate your suggestion. However, incorporating the 13th option would introduce additional considerations, such as creating an ensemble for each category separately or forming an ensemble across different categories. Given these complexities, we have decided to remove this point from the summary related to the ensemble option in the manuscript. This adjustment will help streamline the focus of the paper, ensuring clarity and consistency in our analysis while avoiding potential ambiguities.

7) Appendix

a) Please add the references from where you obtained the formulations for the different methods to Table A1.

We will add the references to Table A1 as well. Previously, references for the PET methods were included in Table 2. Adding references to Table A1 will provide a comprehensive citation for all the formulations, ensuring clarity for the readers.

8) Technical corrections

a) The words of the title should not be capitalized.

We will update the title in the revised manuscript.

b) In the abstract, the third category is called “combination type”, later it is called “combinational type”. Please improve for consistency.

We will homogenize it throughout the manuscript.

c) Please note that “ERA5-Land” is written with a capital L (not “ERA5-land”). This is not consistently correct throughout the manuscript.

We will correct it in the revised manuscript.

d) In line 75, you specify two of the components again after the abbreviations have often been used in the preceding introduction. Either, leave this away, i.e., just give the abbreviations, or give the full names of the components plus the abbreviations for all three components consistently. Furthermore, I think the sentence starting at the end of this line is redundant (stating the same as the preceding sentence in other words), please double-check and correct.

We will correct it in the revised manuscript.

e) On line 117 there is a full stop after "estimation", but then the sentence seems to continue. Please double-check and correct.

We will correct it in the revised manuscript.

f) For the McGuiness-Bordne method, first occurring on line 119, you use different ways of spelling. To my knowledge, "McGuiness-Bordne" is the correct spelling, please double-check and adjust accordingly (do not forget figures and captions). Similarly, you don't spell "BaierRobertson" and "Milly-Dunne" consistently (you sometimes write "Bair-Robertson", and in Table A1 "Milley-Dunne"). Please make sure that you use the correct spelling throughout your manuscript, including also the supplementary material (occurrence of "Milley and Dunne" in text S2). For all methods consisting of two last names, decide if you want to use a hyphen, a space, or an "and" to connect them.

We will correct typos throughout the manuscript.

g) Double-check the references given in the tables: For example, in Table 2, the reference for the Hargreaves-Samani method is in a different style than the other references.

We will update the reference.

h) It is mathematically problematic to use several letters for the same variable in an equation, e.g., "ND" could be interpreted as "N times D". Please consider reformulating.

We agree with your comment; we will correct it in the revised manuscript.

i) The first sentence of the results (line 175) is a repetition of the methods, please consider deleting.

We will delete it.

j) In the paragraph starting on line 184, it is not fully clear that the trends that are described are AET trends. I suggest you to formulate this clearer, for example by rewriting the second sentence to "... all PET methods lead to a positive AET trend in terms of median values." Similarly, in the following paragraphs describing the trends in Q and TWS (and in the results section in general), make sure that it is always clear that it isn't the PET methods that have trends but the PET methods that induce trends on the different components (if I understand your statements correctly).

We acknowledge your comment; we will elaborate this text with better clarity in the revised manuscript.

k) In line 206, you state: "... but the overall pattern of PET methods matches well with PET in all categories." This statement is unclear, did you mean a good match of PET with AET?

Our intention was to convey that PET trend patterns align with AET trend patterns. For example, if a particular PET method shows a higher trend in PET, the same method will also exhibit a higher trend in AET, though with a lower magnitude. For better clarity, we will revise this statement in manuscript.

l) In the caption of figure 3, you use mm seas-1 year-1 as a trend unit. I assume that 1 mm seas-1 year-1 in AET means that each year, 1 mm more goes to AET during the season of interest. As this seems to be an unusual unit (at least for me), I suggest you to explain this in the methods section of your manuscript already.

Thank you for your comment. For example, if the trend in the summer season is 1 mm seas^{-1} year^{-1}. it means that each year, an additional 1 mm will be added to the summer season. Sometimes, it is also represented as mm/season per year. We will clarify this in the methods section of the manuscript.

m) In line 219, the sentence starting with "AET in the summer season" is unclearly formulated.

For better clarity in the sentence, it will be formulated as follows: *In the summer season, energy-limited and mixed catchments generally exhibit an overall positive trend in AET.*

n) The references in lines 309, 310, 313 are not formatted correctly

We will correct the reference in the revised manuscript

o) In the supplementary material (Text S1), make sure that each excerpt starts on a new line and that there is always a space between the colon after the study name and the excerpt itself. Furthermore, look for typos in the copied excerpts (e.g., in the second excerpt from the Anabalón & Sharma study, you are missing a lot of spaces). For the Shi et al. (2023) paper, it would be good to indicate which paper you mean (a or b).

We will update in the revised manuscript.

p) Please include the methods' abbreviations in the caption of Table S1.

We will add abbreviations in Table S1.

q) I assume that Figure S1 and Figure S4 show the winter season, please double-check and correct the captions.

Thank you for pointing out this typo. We will correct it in the revised manuscript.

r) In Figure S4, some of the numbers in the plot are not readable. Please see suggestions to improve Figure 6 of the manuscript (that also apply for all the other figures of this type). This should solve the problem. However, make sure that the different numbers are not written on top of each other.

We will correct it in revised manuscript.

s) For Figures S10 and S11, the correct axis labels would be "Trend in . . . [. . .]" as the data are showing the trend and not the value of a certain component.

We will correct it in revised manuscript.

9) Individual typos

We appreciate for providing these typos; we will remove all the typos 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.

Deardorff, J. W.: Efficient prediction of ground surface temperature and moisture, with inclusion of a layer of vegetation, Journal of Geophysical Research: Oceans, 83, 1889–1903, https://doi.org/10.1029/JC083iC04p01889, 1978.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

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.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, Journal of Geophysical Research: Atmospheres, 99, 14 415–14 428, https://doi.org/10.1029/94JD00483, 1994.

Massari, C., Avanzi, F., Bruno, G., Gabellani, S., Penna, D., and Camici, S.: Evaporation enhancement drives the European water-budget deficit during multi-year droughts, Hydrology and Earth System Sciences, 26, 1527–1543, https://doi.org/10.5194/hess-26-1527-2022, 2022.

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.: SC-Earth: A Station-Based Serially Complete Earth Dataset from 1950 to 2019, Journal of Climate, 34, 6493–6511, https://doi.org/10.1175/JCLI-D-21-0067.1, 2021.

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.