Reply to Referee's comments

Referee #3

Referee's comments are in black text

Authors's response are in blue text

1) Overall evaluation

The authors presented a study on the impact of 12 PET formulations on the trend of a set of components of the hydrological cycle in 553 catchments across Europe. They used a large-scale rainfall-runoff model to simulate actual evapotranspiration (AET), total water storage (TWS) and runoff (Q) multiple times by varying the PET forcing according to the 12 selected methods. Then, they analysed the annual and seasonal trend of PET, AET, TWS and Q obtained thought the different PET methods. They concluded that the choice of PET formulation influences the components of the hydrological cycle.

The work has a strong potential and the issue is of great interest in the field of catchment hydrology. In addition, this experiment could help fill a gap in the literature, which currently lacks a clear understanding of the effects of different PET formulations on rainfall-runoff modelling. However, I have few major concerns, especially about the methodological approach, which I think should be addressed in order to enhance the reliability of the results, facilitate and improve their interpretation, and meet the standards required for publication in HESS.

We sincerely thank Referee #3 for constructive feedback, which helps to improve the quality of our manuscript. We sincerely appreciate the time and effort invested in providing such a thorough and insightful review.

Most of my concerns were already highlighted in detail by the other two referees. Therefore, I would focus exclusively on the most critical issues, which need significant improvements.

2) General comments

a) Modelling framework and model accuracy

A more detailed description of the modelling framework is certainly needed in order to better understand the experiment and its results. Please provide information about model spatial and temporal resolution, model calibration (or previously calibrated model settings) including objective function(s), calibration/validation period, input data used, etc. If a default parameterisation is used, as stated in the very last part of the manuscript, I believe the authors should elaborate about it and its impact on the outcomes of the analysis (i.e. can it be reliable?). In general, I suggest providing a brief overview about model performances against observed streamflow (which I suppose were used somehow for model parameterisation and/or to evaluate the default parameterisation) across the study catchments. I am aware that's definitely not the focus of the study but, since the entire analysis is based on a set of model outputs (streamflow included), I believe it is important to verify (and show) model accuracy in order to consolidate the interpretation of the results and draw solid conclusions. In fact, even if on one hand good model accuracy in reproducing streamflow does not guarantee a faithful reproduction of other hydrological components, on the other hand I would tend not to rely on the state variables of a poorly performing model. Maybe you can mention about model performance in the text and report the details in the Supplement.

Finally, I agree with referee Franziska Clerc-Schwarzenbach that, if a method for potential evapotranspiration was involved in the model parameterisation, authors should provide details about it and comment about the potential effect it could have on the outcomes of the experiment.

We thank reviewer for providing constructive comments. The detailed model setup and and its performance is discussed Referee #1 comment, in Section 3) Methods and Data, 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 (553×12). 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.

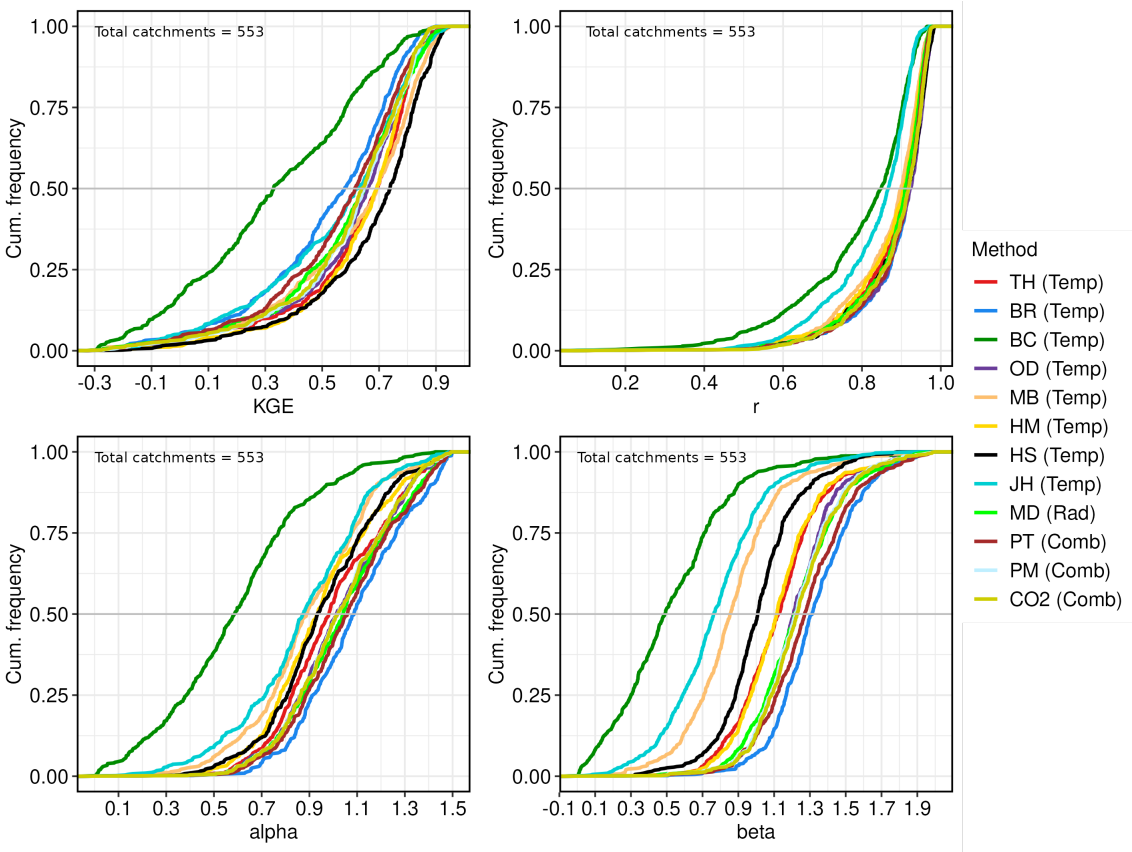

Figure 1. Evaluation of the hydrological model (mHM) performance using simulated monthly streamflow across 553 catchments, forced with EM-Earth meteorological data. The figure presents the cumulative frequency distributions of the Kling–Gupta efficiency (KGE) and its three components: correlation (r),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

b) Trend analysis

First of all, I am sorry to say that the trend analysis is lacking. In particular, authors computed and took into account exclusively the non-parametric Sen's slope test, which estimates the magnitude of the trend of a time series but does not ensure its statistical significance. To affirm that a signal has a trend, it must be statistically significant. Therefore, I ask to the authors to complete the trend

analysis by associating a significance test (e.g. Mann-Kendall) to each trend magnitude (Sen's slope) and, consequently, change all the results and their interpretation accordingly.

In addition, I suggest excluding (maybe adopting a threshold) very week positive/negative trends when computing DCI, which may include a lot of noise and mask some aspects of your results.

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.

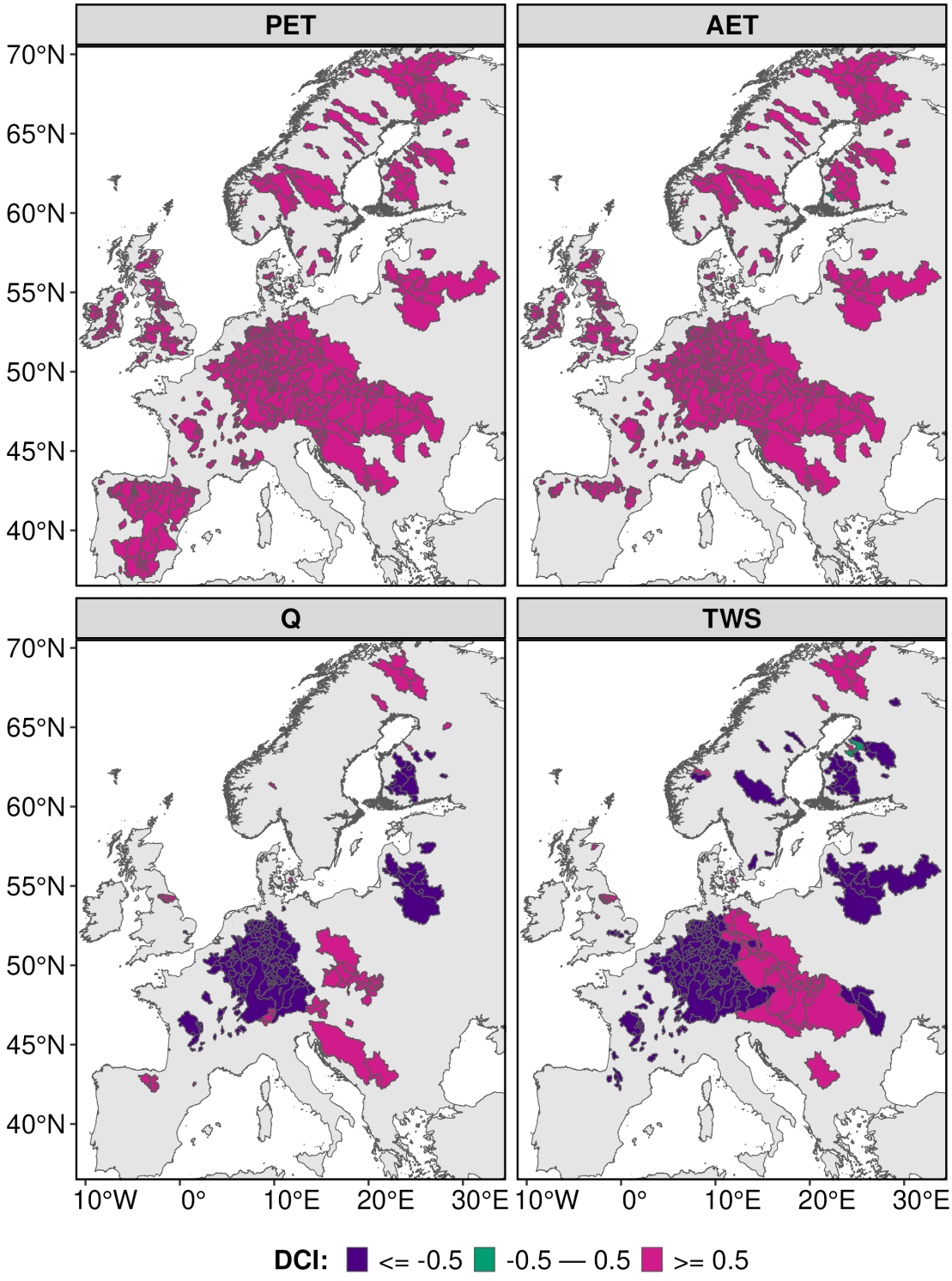

Figure 2. Spatial distribution of annual scale data concurrence index (DCI) for PET, AET, Q, and TWS by considering only significant trend at 95% significance level. PET represents potential evapotranspiration, AET represents actual evapotranspiration, Q represents runoff at the outlet of the catchment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

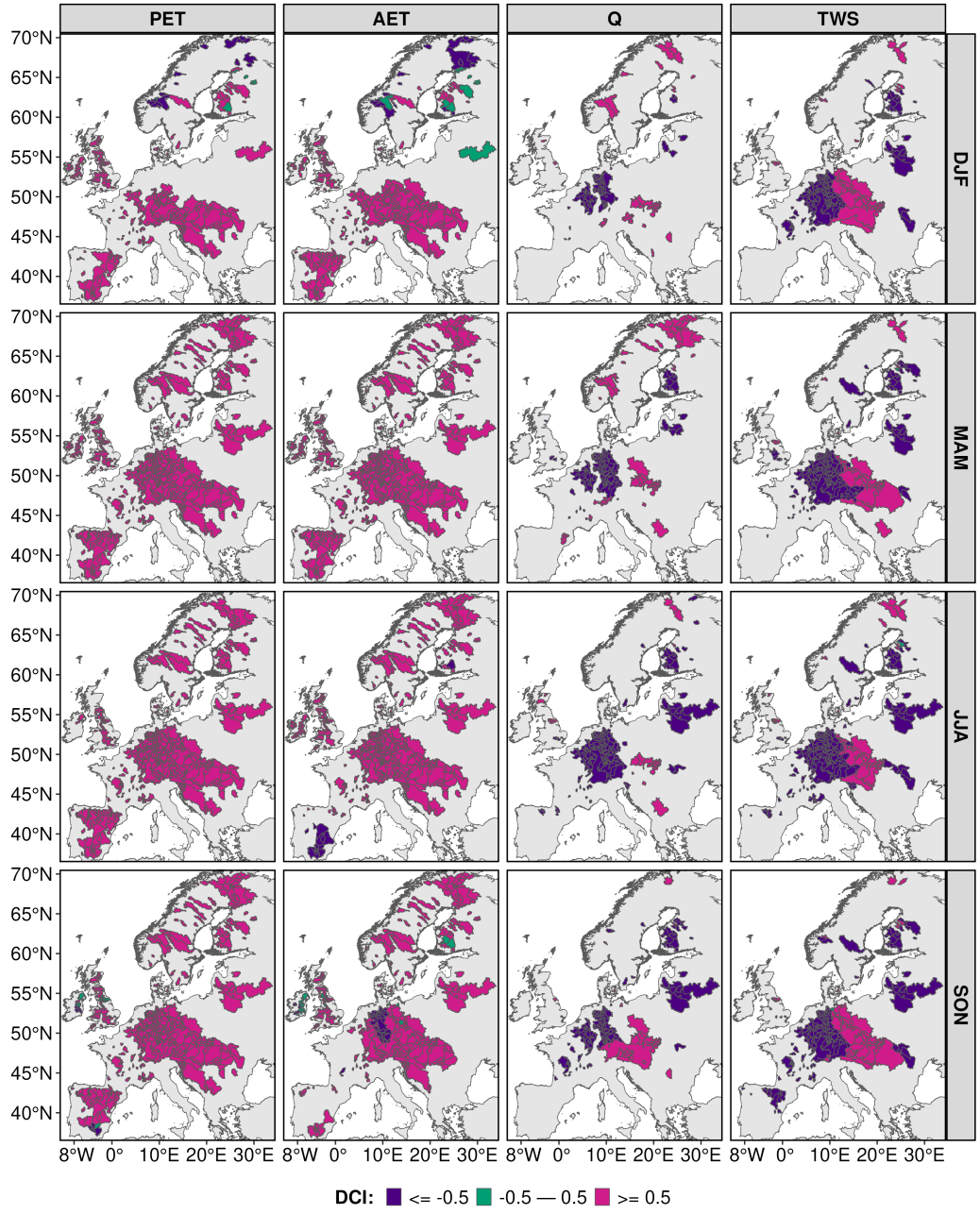

Figure 3. Spatial distribution of seasonal scale data concurrence index (DCI) for PET, AET, Q, and TWS by considering only significant trend at 95 % significance level. PET represents potential evapotranspiration, AET represents actual evapotranspiration, Q represents runoff at the outlet of the catchment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

c) Results and discussion

I personally find some parts of the results section very hard to follow. In particular, please consider reviewing the text on seasonal trends (Section 3.2) and on combination of hydrological cycle components (Section 3.4).

We will revise it with better clarity in our manuscript.

In addition, when commenting DCI outcomes in Figure 4 and 5, authors refer to Northern/central/Southern Europe to develop the description. It would be useful to be more specific, because sometimes the text is misleading. For instance, at line 256 you state "... Q shows a strong decreasing trend for all PET methods in most central European catchments" but if I look at figure 5, central-Eastern DCI for Q are mostly negative. Is eastern Europe not included in "central"? If so, comment also about Eastern Europe. Again, Great Britain is considered Northern or central Europe.

We agree with you. we will clarify it further in our revised manuscript.

Figure 6 is not intuitive and difficult to interpret (and must be revised since some of the PET methods don't sum 553?). I strongly agree with the suggestions of Franziska Clerc-Schwarzenbach and Anonmymous referee #2. Also, the figure format and meaning should be explained in detail in the text before commenting it. Moreover, I suggest adding maps of the catchments coloured accordingly to the obtained combinations (or at least some of them), in order to be able to locate basins in space.

We will update the figure as we described in Referee #1 comments section 4) Results and subsection g).

The trends of AET, TWS and Q are strongly influenced not only by the PET method but also by PRE trends. Even if it is obvious, I would report PRE trends (and their significance) in the results (and not only in the Supplement) and use it to justify the trend direction of the other components.

Thank you very much for your comment, we will add it in the result section. We will update it in revised manuscript

Finally, the discussion about the obtained combinations of hydrological cycle components is poor. I believe it should be extended.

Thank you for your comment. A similar point was also raised by Referee #2. We will further elaborate on the discussion in this paragraph in our revised manuscript.

3) Additional minor comments

a) Figure 1b: I would specify in the text (not only in the caption) that the example refers to the catchments with bolder black contours in panel a. In addition, I would avoid interpolating the points: please use just dots of different colour.

Thank you very much for your comment. The interpolation lines in Figure 1b have been removed and are now represented as colored points for each catchment category (Figure 4). We will revise the text as suggested by the Referee in the revised manuscript.

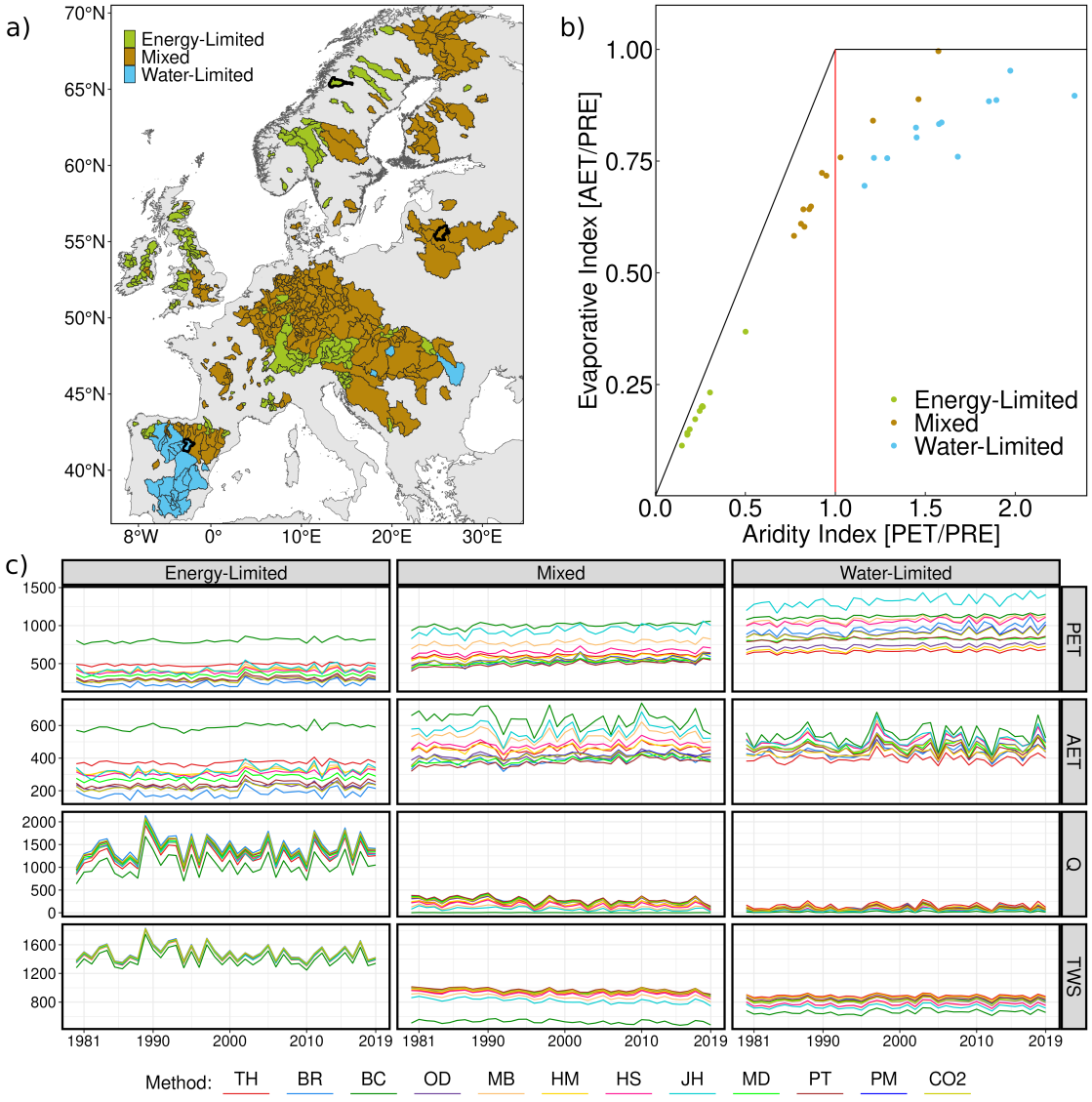

Figure 4. Catchment classification to energy-limited, mixed, and water-limited categories. a) Catchment locations; black borders indicate a representative catchment of each category. b) Classification example within the Budyko space for the representative catchments. c) Annual time series of simulated 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}.

b) line 95: Please give some information about time coverage of the datasets, which I guess can justify your following choice regarding the simulation period.

Thank you for your comment. We have included the record length along with the temporal and spatial information for each dataset in Table 1 of the manuscript.

c) lines 103-104: I perfectly understand this choice, since ERA5-Land precipitation and temperature are known to be often not accurate, leading to a degradation of model performances. However, since one may wonder why not all variables from ERA5-Land are used, I would refer to recent studies highlighting such issues (e.g Clerc-Schwarzenbach et al. 2024, Tarek et al. 2020)

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).

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.

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.