Dear Ryan Teuling

Thank you for providing us with the review from the reviewer training.

Dear Reviewer from the peer review training

Thank you for choosing our manuscript and for your comments on it. Please find our replies to your comments below. We used *blue italic font* to distinguish the comments from our replies.

Best regards,

Franziska Clerc-Schwarzenbach & Thiago do Nascimento

Main comments

Regarding the methods, the authors have included all catchments in their studies, regardless of being impacted by human activities or not (L94-97). This can be questionable, as only the climate forcings are used as the hydrological model inputs. This modeling approach may only be applicable to the natural sites without human intervention. Among the 3423 catchments, these include much noise in the modeling results. Importantly, this approach makes it hard to differentiate if one type of meteorological data is better than the other one because of natural condition or human intervention or the quality (or station density) of the meteorological data itself.

We thank the reviewer for raising this important point. After discussion, we decided to incorporate some information on lakes and reservoirs available in EStreams into our filtering procedure. Specifically, we now will retain only catchments that meet the following criteria:

- Number of lakes upstream < 5
- Normalized upstream capacity < 0.2

The normalized upstream capacity was computed following Salwey et al. (2023), and the threshold value was defined based on their findings. We acknowledge that this filter may exclude some catchments that are not substantially affected in their water balance by regulation; however, we chose to adopt a more conservative (stricter) filtering approach, aiming to exclude all heavily regulated catchments.

Furthermore, we opted to consider only the number of lakes and the normalized upstream capacity because this information is readily available in EStreams, ensuring a consistent and fair filtering process across countries, even where anthropogenic impacts are not explicitly indicated in their CAMELS-like datasets.

We will include this information in the cascade of exclusion criteria in section 2.1.

Salwey, S., Coxon, G., Pianosi, F., Singer, M. B., & Hutton, C. (2023). National-scale detection of reservoir impacts through hydrological signatures. Water Resources Research, 59, e2022WR033893. https://doi.org/10.1029/2022WR033893

What about other possible governing factors, such as climate types, topography, land use and land cover, and geology? These are all not addressed or analyzed by coming to the conclusion due to spatial resolution and station density. Simply saying one is better than the other without analyzing the possible governing factors could limit the applicability and generalization of the research outputs. Therefore, more analysis on process-based understanding and transferable knowledge is needed to make robust conclusions supported by the evidence.

Climate, topography, land use, land cover, and geology all remain constant, independent of the meteorological forcings used to calibrate a model (i.e., in the different scenarios, the only factor that we changed are the meteorological forcings). Thus, we would argue that the catchment characteristics are not relevant for the evaluation of the hydrological efficacy of the meteorological forcings when used for a hydrological model calibration.

The authors adopted the potential evapotranspiration data derived from different approaches: it is calculated with the simplest approach (only temperature based) in E-OBS, but with different varieties of methods in the CAMELS. If the authors want to do a comparison, it should be "apple" to "apple". It is recommended that the potential evapotranspiration should be calculated with the same methods for both types of datasets.

For this study, we used different existing datasets. These datasets are openly available and contain data that are ready to be used. Having said that, it is most likely that a user of the dataset will make use of those potential evapotranspiration data that are provided in the dataset of interest. As the different datasets (i.e., the different CAMELS and CAMELS-like datasets as well as the EStreams dataset) were created by different teams and for different

regions, it is in the nature of the subject that different approaches to calculate potential evapotranspiration were used.

We argue that a comparison of the results (in this case: the potential evapotranspiration data) is especially relevant when different approaches (in this case: different equations to calculate potential evapotranspiration) were used with the same goal. After all, it is in the interest of the modeler to know how different the input data is depending on which dataset they choose to gather data for a certain catchment. Thus, we consider it valuable to not change the data that were made available in the different datasets, but to work with what is provided (and is thus used by the community).

Regarding the results, it would be more useful to state the governing factors (climatology, topography, land use, etc.) why E-OBS has over- or underestimations compared with CAMELS, besides simply stating which countries or regions have higher or lower meteorological values. More exactly, why one dataset is better than the other one in some countries yes while some countries not?

We thank the reviewer for making us aware that this occurs to be incomplete. We checked for correlations of our results with other catchment characteristics and did not find anything besides what we stated. We will include a sentence in the results section making this clear. Furthermore, as suggested by Reviewer 2, we will include a discussion on potential reasons for the differences in the precipitation data.

Another key aspect is that the authors calibrate the models individually with different climate datasets. Therefore, not only the climate data are different, but the model parameters are different. Therefore, the model performance lower or higher is not only due to climate data quality but also the model parameters.

The 'optimal' parametrization of a bucket-type hydrological model may differ depending on the meteorological input data (for example, if the model tries to compensate for a bias in the data). We do not see any possibility of making a fair comparison of the model performances without informing the model with the different meteorological input data that it has to deal with then. Furthermore, note that for each model performance value, we used ten independently optimized parametrizations to avoid a strong dependence from one parameterization.

Specific comments

L10: Maybe mentioned the spatial resolution of the meteorological data from the E-OBS?

While this is surely important information, we do not think that it should be part of the abstract for which the length is limited. All information on the spatial resolution of the E-OBS data is given in section 2.2. Note that in the next version, we will use the E-OBS data at a resolution of 0.1° as these is being made available in EStreams in the meantime.

L16: Model performance is SLIGHTLY lower when E-OBS data are used compared with CAMELS data: is this difference statistically significant?

We thank the reviewer for this question. To evaluate whether the difference in model performance between E-OBS and CAMELS forcing data is statistically significant, we conducted a Wilcoxon signed-rank test (Wilcoxon, 1945) on the paired KGE values. The test indicates that the median KGE for CAMELS (0.883) is slightly higher than for E-OBS (0.867), and that this difference is statistically significant (Wilcoxon signed-rank test: $W = 3.59 \times 10^{6}$, $p < 10^{-29}$). Although the difference is statistically significant due to the large sample size, the effect size is small, suggesting that the practical difference in model performance is minor. We will add this information to the manuscript. (Note however that the numbers may change due to the change in spatial resolution of the E-OBS forcing data).

Wilcoxon, F.: Individual comparisons by ranking methods, Biometrics Bull., 1, 80–83, 1945.

L48-53: the authors actually come to the same conclusion as the referred literature, and mentioned the same thing in the abstract. So what is the added value of evaluating E-OBS vs. CAMELS? Just because of a larger scale of detailed dataset?

As stated in L53-54: “Yet, evaluations of the E-OBS data for a larger extent, and specifically for hydrological modelling, remain unexplored.” – So far, there have been no tests of the E-OBS data in a hydrological model, and especially not in a comparison to alternative data. For a hydrological modeler working on large-sample hydrology in Europe, this study will support an informed decision for (or against) a certain dataset.

L84: Why exclude the catchments with area more than 2000 km2? What is the impact or relation between the catchment area and the meteorological data?

We thank the reviewer for this question, pointing out that a statement on the motivation for this decision is missing. For a large catchment system (arbitrary threshold of 2000 km^2), a bucket-type hydrological model may not be the most suitable choice. Therefore, we excluded catchments larger than that from this study. We will add a sentence clarifying this in the cascade of criteria in section 2.1.

L123-132: Why are the annual differences of precipitation and evapotranspiration between the datasets compared but not the seasonal differences? While for temperature, you compared the daily differences?

We are aware that the comparisons we made only provide a limited picture of the differences in the meteorological input data. However, as this is not a study purely on data comparison, we decided to include one measure per variable. For temperature, it is not possible to calculate an annual sum that can be compared. Therefore, the mean daily difference (which is the same as when the annual mean temperature is calculated first, and the mean difference is calculated then) is given in that case.

Figure 4: What are the reasons for the different model performance among the countries? What are the governing factors? Simply stating the KGE is higher here or lower there without providing further reasons sounds not helpful.

We thank the reviewer for their comments. However, note that in the section “Results” we only describe the results of the study, without interpretation. The reasons for the higher or lower model performances – as far as they could be identified – are given in the section “Discussion”. Motivated by a comment of Reviewer 2, we will avoid comparing the model performances (i.e., the KGE values) for different regions between each other.

Figure 6: The important thing is not the exact number of catchments in a country where E-OBS dataset is better or worse than the CAMELS datasets, but why E-OBS is better/worse than the CAMELS in these catchments?

We thank the reviewer for pointing this out. We think that it can be helpful to know which meteorological input data lead to a more successful streamflow simulation, since the model performance can be interpreted as an aggregated measure for hydrological efficacy (and thus gives an indication of which data may be of a higher quality). Regarding the reasons for the lower or higher model performances, these are discussed in the section "Discussion", where the modelling results are interpreted.

L275-278: Why is the model performance lower in Great Britain which shows opposite behavior? Please explain.

In scenario III, we use the (higher) potential evapotranspiration data from EStreams, while in scenario II, we use the (lower) potential evapotranspiration data from CAMELS (in this case, CAMELS-GB). In section 4.3, we explain why for the karstic catchments in Great Britain, the higher potential evapotranspiration data were beneficial. We will add a sentence at lines 275-278 indicating that for these catchments, potential evapotranspiration was more important than elsewhere to make the link to the explanation provided later.

Figure 7: Simply stating the station density plays the key role seems not convincing, as the author stated that other factors may also play a role. It would be more interesting to analyze other factors as well? Are the relationships between the station numbers and the KGE statistically significant?

We have computed the correlations between KGE and catchment descriptors, and this statement is based on that. Specifically, only the correlation with the number of precipitation (and temperature) stations achieved a statistically significant coefficient.

We agree that there is room for improvement in the text and results, and we will add the table with the correlations in the appendix. We will also improve the discussion and also make some improvements in the figure, such as including the Spearman ranking coefficient (relationships) and p-values (significance) for each subplot (country).

Figure 8: What about a trend assessment on the data? Is there a significant relationship between model performance and aridity index?

We believe that this is a fair assessment, and agree that it is currently lacking. We will compute the correlation between the two variables (Spearman ranking coefficient and p-value) and consequently plot the LOWESS (locally weighted) smooth line for the trend assessment for each subplot in Figure 8.

L366-369: it is too assertive and not supported by evidence. It is a very simple method to calculate the potential evapotranspiration which does not consider solar radiation impact. It is also too assertive to say different calculation approaches of potential evapotranspiration will not change the results.

Note that the evidence that the Hargreaves equation is reliable (e.g., in Central Europe) does not origin from the current study, but from the study by Pimentel et al. (2023), cited in this sentence. Furthermore, the choice for the Hargreaves equation was not made for this study, but when the EStreams dataset was published. The statement that the different potential evapotranspiration data did not affect the model performance results strongly is supported by Figure A6.

To address this comment, we will change L366-369 to:

"Furthermore, the differences in E_{pot} data did not affect model performance results strongly (as can be seen in Figure A6), so it can be expected that the use of a different equation would not change the findings of this study."

Additionally, we will add two other references supporting the reliability of the Hargreaves, as suggested by Reviewer 2.

Weiland, F., Tisseuil, C., Dürr, H., Vrac, M., & van Beek, L. (2012). *Selecting the optimal method to calculate daily global reference potential evaporation from CFSR reanalysis data for application in a hydrological model study*. Hydrology and Earth System Sciences, 16, 983–1000.

Bangi, S. C., & Soraganvi, V. S. (2023). *A modified temperature-based model for estimation of potential evapotranspiration over Ghataprabha river basin, South India*. Spatial Information Research, 31, 583–595.