Dear Reviewer 3

Many thanks for reviewing our manuscript. Your remarks will be very helpful in further improving the study. Please find below our replies to the review comments and how we will implement them in the revised version of the paper. We used *blue italic font* to distinguish the comments from our replies.

Best regards,

Franziska Clerc-Schwarzenbach & Thiago do Nascimento

Major remarks

First, your paper would deserve a better title. As I understood, you ask a much more general and (from my point of view) interesting question: how can we use a classical precipitation-runoff model such as HBV in order to compare the quality of precipitation data. I believe you should put this point at the forefront of your paper. You should discuss the "good sense" (almost philosophical) hypothesis of your approach: even if your hydrological model is imperfect, the difference of efficiency when calibrating the model with different forcings cannot be due to some random factor. Better performances cannot be due to chance. You could perhaps look at this chapter of the famous Ray Linsley (1982) who discussed the topic, I only remember this short citation "if the data are too poor for the use of a good simulation model they are also inadequate for any other model", but there must be some other interesting citations there.

We thank the reviewer for bringing this up and for the valuable suggestion for literature. We think that the paper will benefit from putting more emphasis on how we use the hydrological model and why we think that it is a meaningful way to do so. We will thus include this in the introduction. In addition, we will change the title to put a spotlight on the methodology already there (see also comment by Reviewer 2).

Second, I believe that it is worth comparing the CAMELS outputs with the E-obs outputs, introducing a further class of inputs (ERA-5) makes things more complex. I would simply have discarded the LAMAH dataset, stating that you aim at comparing the "best ground-based estimate" of the CAMELS datasets with the E-obs… it is definitely not a big surprise that ERA-5 estimates are not good… and it makes your paper unnecessarily more complex. You do not have to show us everything you have done, if you have pushed open at a few open doors in the course of your research (what we all do…) you do not need to tell us about it.

We thank the reviewer for encouraging us to exclude the Austrian catchments from the analysis. In fact, much of the presentation of the results and the discussion would be simpler if the special case of LamaH-CE was not included in the study. Thus, we decided to include a short note on what makes the LamaH-CE dataset different from the others and why we thus did not use it for this study in section 2.1 and exclude these catchments from further analyses.

Third, I was wondering whether it would have been interesting to restrict the dataset to the less-regulated (reservoir-impacted) catchments. I know for example that there are quite a few regulated catchments in the Swiss CAMELS dataset. It will not change the results, but a focus on the less regulated catchments could perhaps show even clearer differences.

We thank the reviewer for raising this issue. Based on this recommendation as well as the recommendation by Reviewer 2, we decided to exclude catchments with a normalized upstream capacity larger than 0.2 (Salwey et al., 2023) as well as catchments with 5 or more lakes upstreams. We use these criteria as they are available in the EStreams dataset and can thus be used for all catchments of the study consistently.

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

Minor comments

I believe that before mentioning (l.36) that "the inclusion of an increasing number of catchments in one dataset almost always goes hand in hand with difficulties in providing high-quality forcing data" you should underline that large samples also come with their load of problematic discharge stations. In my experience of building a CAMELS dataset, a large part of the effort was absorbed by scrutinizing collectively the time series, the locations, etc. And because E-streams did not make any sorting, there must be along with the hydrometric stations a few (or more) non sense stations (probably a few buoys in France...) or at least stations which measure a level that cannot be related to any significant hydrological flux.

We thank the reviewer for addressing this potential issue. We will underline along line 36 that larger large-sample datasets are expected to be more prone to wrong streamflow data than smaller large-sample datasets that were sorted and filtered by hand. However, since we only

use catchments that are included in one of the CAMELS datasets (where we assume that filtering took place in all cases), we believe that only meaningful stations were included in our study. Furthermore, since we used the streamflow data from the CAMELS datasets in all scenarios (see next comment), we do not expect any issues regarding wrong or inconsistent streamflow stations.

Did you check that the discharge data were exactly the same in E-stream and CAMELS?

For streamflow, we used the data provided in the different CAMELS datasets for all scenarios, for two reasons: a) to make sure that differences in model performances were due to the meteorological input data and not affected by potential differences in the streamflow data, and b) since EStreams only provides information on how to get to the streamflow data of the different stations, but not streamflow data directly. We will add this information in section 2.4 and are thankful for the remark on this that made us aware that this information is currently missing.

In 3.3.1 (Number of E-OBS precipitation stations): I believe you should mention that the number of E-Obs stations is correlated with the size of the catchments... and as you (and all the conceptual modelers) know, the largest catchments get the best KGE criteria.

We thank the reviewer for raising this point. We will include a remark on this dependency, as well as further information regarding the relationship between catchment areas and numbers of E-OBS stations at this place in the manuscript.