Dear Reviewer 1, dear Alex

Thank you very much for your encouraging and detailed review of our manuscript. Your comments will be very helpful for further improvements. 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. Of course, we will also implement the technical corrections. Thank you for spotting them.

Best wishes,

Franziska Clerc-Schwarzenbach & Thiago do Nascimento

L11:
"limitations of data quality" -> maybe indicate that data quality is expected to vary in space? (e.g. "limitations and regional variations of data quality")

We thank the reviewer for the feedback. We will modify the text to: "limitations and regional variations of data quality."

L36-L39:
The number of catchments is not directly the problem, the mixture of different regions / countries is the challenge, as meteorological data is often available on a national level (e.g. provided by national meteorological organizations)

We thank the reviewer for this input. We will adjust the introduction to make sure that this is explicitly stated and to avoid that the number of catchments is stated to be the problem.

L40, L45, L55, L72, L417…:
I know what you want to say here, but I don't like the word "standardization" in this context, as it usually refers to something else when it comes to data processing, and e.g. ERA5 or E-OBS are just datasets on a larger scale with different sources and processing methods, they do not "standardize" smaller datasets.

We thank the reviewer for raising this point. After careful consideration, we will change the wording at all instances, to make sure that we point to the consistency of the data over large spatial extents without using the potentially misleading word "standardization".

L91-L93:
Why did you apply these criteria?

Catchments with an average streamflow above 10 mm/day were excluded since such values exceed typical ranges reported in LSH datasets (normally < 5 mm/day) by far, and may indicate data inconsistencies (e.g., overestimated streamflow or underestimated area), or glacier-dominated hydrology.

Catchments with runoff ratios above 1.1 were removed because natural runoff rarely exceeds precipitation by large margins, and such instances could indicate data errors or strong human influence.

We will add our reasonings for these constraints in section 2.1.

L111-L112:
The resolution of 0.25° of E-OBS is very coarse, I know that e.g. the precipitation data for CAMELS-DE has a resolution of 1x1 km, this could be an additional source for limitations of Estreams data, also for comparisons in this study. Maybe you could think about an update of Estreams in the future? I think this could be worth it (not part of this study).

We thank the reviewer for the feedback and input. In the meantime, EStreams is being updated with the E-OBS data with a 0.1° resolution. We will thus rerun all the model runs and redo the analysis with the data with a 0.1° resolution. Preliminary results show that this will not change our results strongly, still we think it is fair to include the data with the highest resolution available. However, we will include in section 2.2 that the different spatial resolutions of the CAMELS and the E-OBS datasets are expected to lead to different performances.

L116-L117:
I think the main thing here is that the quality and uncertainty of E-OBS data have a larger (regional) spread, some regions will have very good quality data (where station measurements are available), other regions with less station measurements will have worse data quality. Even if the data comes from the same source (E-OBS), quality and uncertainty varies regionally. I think this is a major challenge in LSH and people need to be aware of this.

We thank the reviewer for these thoughts and the valuable discussion. We will include these differences already here and will stress that using the same dataset in two different regions does not necessarily imply the same data quality for both regions. We would like also to point out that this is further discussed in section 4.2.

L149:
Maybe add a small explanation on why you designed the scenarios this way, and which questions you aim to answer with the different scenarios (I and II are quite clear, but why did you do III-V?)

We thank the reviewer for pointing out that this is unclear. We also believe that this point touches some of the discussion in the comment from the EGU peer review training (CC1).

Scenarios III-V were chosen to disentangle the impact of each forcing time series from E-OBS in the results driven by scenario II. In other words:

- Scenario III was chosen to evaluate the impact of precipitation.
- Scenario IV was chosen to evaluate the impact of E_{pot}, and consequently to evaluate whether using a different E_{pot} formulation would change our main results. (See also comment about the possibility of using different E_{pot} formulations by the reviewer from the peer review training, i.e., CC1.)
- Scenario V was chosen to evaluate the impact of temperature.

We will update section 2.4 with explanations on the motivation behind each of the scenarios, by extending the statement in the very beginning of the section.

L152:
This could also go into limitations, but the catchment shapes are also not identical between EStreams and the CAMELS datasets, which results in different areas for which the meteorological data was "cut out" and aggregated, which can also lead to differences.

We thank the reviewer for raising this point. We will include a remark on this issue in section 2.2 to make sure that readers (and, more importantly, users of the datasets) are aware of this.

L312:
I think it is hard to see any patterns in this figure with the mixture of scenario I and II with circles and triangles. I am not sure on how to improve the figure, but you could calculate a regression line and also report the p-values? This could also be used to back up your statement in L310-311

We thank the reviewer for suggesting these helpful improvements to the figure. We will modify the figure by including the correlation between the two variables (Spearman ranking coefficient and p-value) and additionally plot the lowess (locally weighted) smooth line for the trend assessment for each subplot.

L324-L325:
results in Austria are bad as ERA5 data is used, not a "local" dataset, maybe add this here?

We thank the reviewer for calling our attention to the missing remark on the special issue of Austria at this instance. As suggested by Reviewer 3, we will exclude the results for Austria from the study to avoid including a dataset that fundamentally differs from the others (i.e., that is not a national dataset). We will note however that the LamaH-CE dataset is different to make sure that this point still comes across and users are aware of it.

L344-L353:
Here you have the paragraph about limitations of ERA5 data in Austria, but I think it does not really fit in the paragraph ("Evaluation of the E-OBS data in comparison to the E-OBS station density"). Maybe it could fit better in Section 4.1? I think the point about ERA5 data used in Austria is very important in this study, as this is a fundamental difference to the other CAMELS datasets, where local, highest-quality data is used, in Austria it is quite the opposite. You should make this point very clear, also in the beginning, as you do not test whether "local" CAMELS data is better than E-OBS data in the case of Austria.

We thank the reviewer for this comment. As in the new version, we will only add a statement about Austria and why we did not use it in the study, we think that this problem is solved (while still making sure that the message comes across).

It would also be interesting to see how the different CAMELS precipitation data was collected / processed (maybe not so easy to find out). I only know about CAMELS-DE, but

HYRAS is also based on interpolated station data (I guess mostly the same stations as used for E-OBS), which would explain the relative similarities, but it is still interesting to see that there are differences (maybe due to different interpolation / processing methods or the coarser resolution of E-OBS)

We believe that this is a valuable remark, and thank the reviewer for it. Following the suggestion of Reviewer 2, we will include information on the origins of the data in Table 2. In addition, we will add a section in the discussion in order to discuss the potential implications of the different data origins.

L398-L401:
For smaller catchments, having E-OBS data from the 0.1° version could also help (again, maybe this is worth an update for EStreams, which of course is not part of this study, just a general suggestion)

We thank the reviewer for sharing this thought. In the meantime, EStreams is being made available with forcing data from E-OBS at a 0.1° spatial resolution. Thus, we will rerun all the model runs and redo the analysis with the highest resolution data.

L402…:
You could add to the conclusion that local datasets are usually the best, but using E-OBS data and EStreams offer a great harmonized data source for LSH studies covering all of Europe, especially as an alternative to ERA5 which has shown limitations in Austria. Maybe extend a little bit on this and how E-OBS could be an alternative to ERA5 which was mostly the standard before.

We thank the reviewer for the suggestion to enrich the conclusions. We will modify the last paragraph of the conclusions to something in those lines:

“Overall, while local or nationally curated datasets often yield the best model performances due to their finer spatial resolution and denser station networks, our results suggest that the meteorological forcing E-OBS data in EStreams represents a valuable and harmonized alternative for pan-European studies. The advantage of E-OBS lies in its observational basis, consistent methodology, and coverage across all of Europe, making it especially useful when national datasets are unavailable or inconsistent. As such, E-OBS and EStreams provide a practical foundation for expanding large-sample hydrology beyond

national boundaries while maintaining sufficient data quality for robust model applications.”