Dear Reviewer 2

Thank you very much for your encouraging and detailed review of our manuscript. Your comments 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. Of course, we will also implement the technical corrections. Thank you for making us aware of them.

Best regards,

Franziska Clerc-Schwarzenbach & Thiago do Nascimento

Major comments

1. Framing of "quality assessment"
The title of the paper indicates that a direct evaluation of the quality of the E-OBS meteorological forcing data is performed, e.g. through comparison to meteorological station observations. However, the aim of the paper is to compare E-OBS forcing data in EStreams to meteorological forcings from regional datasets, which themselves may still contain errors and biases. As stated in the introduction, rather than a quality assessment, the study evaluates the "overall hydrological efficacy of the meteorological forcing data". Clarifying this distinction in the title would better align expectations with the actual scope of the work.

We thank the reviewer for pointing out this weakness about our manuscript title and for the suggestions for improvement. We agree that the current use of the term "quality" in the title might be misleading, and following insights also brought by Reviewer 3, we will change the title accordingly.

2. Use of KGE as the sole performance metric
The performance of the hydrological model is evaluated with the Kling–Gupta Efficiency (KGE). While KGE is widely used, it depends strongly on flow variability and can therefore mislead when used as the only performance indicator in a large-sample study such as this one. Because important conclusions are drawn from spatial differences in KGE (e.g. higher model performance in wetter catchments), I recommend including at least one complementary error-based metric (e.g. RMSE, NRMSE, or percent bias) to better distinguish between variability effects and true model accuracy. See e.g. Williams, 2025.

We thank the reviewer for their insights and the valuable hint to current literature. We will keep the KGE as a performance measure to be able to compare the different scenarios (and because people are used to interpreting it). We will though take care not to draw conclusions based on solely the comparison of the KGE *between* catchments, to avoid the dependency on flow variability and catchment area, for example. We will additionally use the percent bias (PBIAS) as a complementary error-based metric to improve the discussion of variability effects.

3. Influence of basin regulation and anthropogenic impacts
The paper briefly acknowledges human influence in some catchments (e.g. dams, diversions), but this is not reflected in the data selection criteria. The EStreams dataset includes information about dams and total upstream reservoir volume. Additionally, for some of the regional datasets used (e.g. BULL in Spain and LamaH-CE in Austria), further information on the degree of human impact is also available. I strongly recommend using this metadata to exclude regulated or heavily influenced basins in Section 2.1 ("Subset of catchments") wherever possible. Even if human influence affects hydrological model simulations across all forcing products (scenarios) similarly, removing impacted basins would increase the robustness and interpretability of the comparison. For example, in the Discussion (lines 333–336), the manuscript notes that the low model performance in Spain may be linked to human influence. Excluding impacted basins where metadata is available (such as in the BULL dataset) would help clarify whether regulation is indeed a substantial contributor to the lower model performances in those regions.

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

4. Differences between forcing products
I miss a discussion on the different types of the local datasets (e.g. observation-based, reanalysis-based, ..). These types have different strengths and limitations depending on factors such as terrain complexity and station density, which may contribute to regional performance differences. Also, it would be interesting to know how much overlap there is in the source data between the EStreams and CAMELS forcings. Further, some regional forcing datasets have considerably higher spatial resolution (e.g. 1 5 km) than E-OBS (0.25°), yet the implications of these differences are not discussed. A short discussion of whether resolution differences contribute to the observed spatial patterns would strengthen the interpretation.

We thank the reviewer for bringing up this very important point. Following this recommendation and also the suggestion of Reviewer 1, we will add information on the different origins of forcing data in Table 2 and include a section where we discuss the potential implications of these differences in the discussion.

Minor comments

A major finding is that the mean annual precipitation sums in the E-OBS data are lower than in the regional datasets. The potential reasons for this are not discussed anywhere. Possible reasons for this could be discussed in the Discussion section (4.1).

We will add potential reasons for the clearly lower precipitation data in E-OBS to section 4.1.

Line 91-93: Please explain why gauges with average streamflow above 10 mm/d were omitted, as well as gauges with runoff ratio above 1.1.

Catchments with an average streamflow above 10 mm/day were excluded, as such values significantly exceed typical ranges reported in LSH datasets, 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.

Also, in accordance with Reviewer 1, we will add the motivations behind these exclusion criteria in the cascade given in section 2.1.

Also, a map that shows which of the candidate basins were eliminated in which filter step would be helpful (visualize section 2.1). It would show why a certain EStreams basin is not included in the final analysis. This should be easy to make, but as-is, the description of the selection filter is not all that informative. The map could go in the appendix or supplemental material.

We thank the reviewer for this suggestion. We will add a figure with these maps in the appendix.

Line 108: Please remove the text in line 108 beginning with "EStreams is a ready-to-use product…" through to "…for the evaluation of the E-OBS meteorological data.", as it does not add information beyond what has already been stated.

We will remove the sentence as suggested.

The following sentence "Note that there is also a version of E-OBS at a resolution of 0.1° available, but not represented in EStreams." can instead be moved to directly follow the sentence that starts with "In EStreams…" in line 103. In addition, it would be helpful to briefly explain why the 0.1° E-OBS version was not used in EStreams, as this is not addressed in the original EStreams paper.

We thank the reviewer for raising this issue. In the meantime, EStreams is being updated with the E-OBS data at a 0.1° resolution. We will thus rerun the model runs and redo the analyses for the new data obtained from EStreams. Therefore, we believe that this problem is solved. Preliminary results indicate that there will only be slight differences to the current results.

Line 120: I suggest adding a column to Table 2 to specify what type of dataset in each case (e.g. observation-based, reanalysis-based, ..)

We thank the reviewer for the suggestion and will modify Table 2 accordingly.

Line 142: Specify the spatial resolution of the DEM

The resolution used was 30 m. We will add this information to L142.

Line 262: The sentence "For the catchments in the center of Austria, the CAMELS data sometimes led to better model performances than the E-OBS data, while the opposite was the case in most other catchments (see above)." Can be removed.

We will remove the sentence as suggested by the reviewer, as the Austrian catchments will no longer be part of the study: Based on a suggestion by Reviewer 3, we decided to exclude the LamaH-CE dataset because it is substantially different from the other CAMELS datasets. We will state the reason for this exclusion in section 2.1.

Line 294: Add (Fig. 7) to the end of the sentence.

We will do this.

Line 298: Figure A8 is not discussed anywhere in the manuscript. Consider removing it or adding a brief interpretation of its relevance.

We will remove it.

Line 361: consider replacing the word "striking" (e.g. with "considerable")

We will replace the word accordingly.

Line 364: This text is confusing: "The Epot calculations for each catchment in EStreams with the Hargreaves equation (do Nascimento et al., 2024) thus also affected the resulting Epot data that we used to represent the E-OBS Epot. However, the Hargreaves equation was found to be reliable, among other regions especially in Central Europe (Pimentel et al., 2023) and this choice can therefore be supported" I suggest replacing it with something like: "Epot calculated with the Hargreaves equation, as in Estreams, has been found to be reliable e.g. in Central Europe (Pimental et al., 2023)."

We thank the reviewer for the suggestion. We will modify the sentence to something along the lines:

"E_{pot} calculated with the Hargreaves equation, as in EStreams, has been found to be a reliable method in various hydrological modelling applications, including in Central Europe (Pimental et al., 2023), and other regions (Weiland et al., 2012; Bangi and Soraganvi, 2023)."

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.

Additionally, since you state "among other regions", consider citing other references that support the use of the Hargreaves equation in other regions.

Please see the reply to the comment above.

Line 383: Consider stating earlier in the manuscript (e.g. in Sect. 2: Data and Methods) that the methodology of this study is based on the study by Clerc-Schwarzenbach et al. (2024).

We will make this sentence into Section 2.