Manuscript review: "Evaluating the quality of the E-OBS meteorological forcing data in EStreams for large-sample hydrology studies in Europe"

This paper compares the E-OBS meteorological forcing data used in the pan-European EStreams dataset to the meteorological forcing data of nine regional datasets. As such, it provides useful insights into the suitability of EStreams for large-sample hydrology applications. The results show that precipitation in EStreams is generally lower than in the regional datasets, while temperature and potential evapotranspiration are higher. Hydrological model performance is typically slightly lower with EStreams than with the regional forcing datasets. The paper is clearly written, and the dataset comparison is valuable to the LSH community. However, some aspects related to the methodology, interpretation and presentation can be improved (see comments below).

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.

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.

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.

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.

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).

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.

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.

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.

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.

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, ..)

Line 142: Specify the spatial resolution of the DEM

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.

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

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

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

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)."

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

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).

Technical corrections

Line 47: Change "arised" to "arose"

Line 348: If using "previous studies," cite more than one source or change to singular.

References

Williams, G. P.: Friends don’t let friends use Nash-Sutcliffe Efficiency (NSE) or KGE for hydrologic model accuracy evaluation: A rant with data and suggestions for better practice, Environmental Modelling & Software, 194, 106665, https://doi.org/10.1016/J.ENVSOFT.2025.106665, 2025.