General comments

I would like to thank the authors for their work on extending the CAMELS-DE dataset. The addition of hourly data for 1,611 catchments, meteorological forecasts and hydrological model benchmarks represents a substantial contribution with clear value to the hydrological community. These additions are well suited to the CAMELS family, and the manuscript is a good fit for ESSD.

The manuscript is well written, and the dataset is well organised. My inspection confirmed the availability of all data products, although some aspects of their documentation would benefit from clarification.

My main recommendations concern the length and focus of the manuscript, the interpretation of the hydrological benchmarks, and consistency between the manuscript and files that are made available.

Overall, I consider this an extremely valuable contribution. Addressing the annexed comments would strengthen the confidence in the dataset and help users understand its capabilities and limitations.

Major comments

Manuscript length

I appreciate the detailed work and thought the authors have put into describing two key additions of the dataset: forecasts and benchmarks. However, I feel that the level of detail makes the paper overly long and includes specialised information that may not be relevant to all users of the dataset.

Sections 5 and 7 each read almost as standalone pieces. While I agree that this information should be available, I believe the technical details and tests used to produce the dataset would fit better in supplementary material or a separate paper.

Suggestions for Section 5

Sections 5.2 and 5.3 could be simplified to provide clearer guidance on usage of the deterministic model and the ensemble forecasts. The additional information and tests could be included in the supplementary material, as they are useful for those intending to use the data.

Suggestions for Section 7

Section 7 reads almost as a substantial paper in its own right, with separate methodology and results sections. I suggest a similar approach here: the authors must clarify how they envisage users using these benchmarks as information. As discussed in the manuscript, many of the poor NSE values are associated with human influence or Karst systems.

Could these results be cross-checked against humaninfluence_attributes, hydrogeology_attributes, or similar information?

Benchmarking

NSE is necessary to describe the overall quality of the time series, but it should not be the only metric used to assess time series whose main purpose is the study of high flows. I propose using Bias-FHV, as used in Large Sample Hydrological studies (e.g. Kratzert, 2019), alongside NSE to assess high flows specifically. I believe the combination of these two metrics is particularly useful for identifying artefacts in station records.

A limitation of this approach should also be acknowledged: the metrics depend on the training, validation and testing periods selected by the authors. I have inspected the 10 stations with the highest Bias-FHV in the dataset, which also have poor NSE. Nine of them seem to occur for a clear reason and might have produced different results if the split was different. DE110080, DE110670, DEA13080, DED12230 abrupt shift in the timeseries somewhere around 2015 (around 2018/19 for 12230), results are, potentially, at least partially explained by discontinuity between the training and testing data; DEA10540 and DEA12970 small testing period sample size could explain results; DEG10090, DEG10860, DE112550 little validation period.

Introduction

The introduction connects the novelty of sub-daily resolution and forecasts with the need for these additions. However, it does not make the same connection for benchmarking. Adding a subsection explaining the value of benchmarks as a consistent baseline would help readers understand this aspect of the paper.

Minor comments

L43 – The phrase “with a few exceptions discussed below” is unnecessary, as the word “predominantly” already makes it clear that there are exceptions.

L66 – “should also be noted that uncertainty is typically higher and data quality is often worse in comparison to daily data” I disagree with the wording. The data are noisier, rather than necessarily of poorer quality.

L93 – This paragraph summarises the introduction and states what has been done in the paper. It would benefit from a clearer opening, such as “In this paper, we…”, followed by the main contributions.

L136 – Split the paragraph here, as the discussion shifts from Section 4 to Section 5.

L181 – Why are the panels in Figure 3 ordered differently from the bullet points? Either reorder the bullet points so that Figure 3a is cited first, followed by Figures 3b and 3c, or change the order of the panels.

L192 – It would be useful to report the number of catchments with complete data coverage (100%, with no gaps).

L209–210 – Similarly, why is Section 4.2 described before Section 4.1? Could the order be changed here?

L211–213 – This text is unnecessary here and could be removed.

L243 / Figure 4 – The number of stations drops by one-third between 2021 and 2025. It would be useful to explain this, if possible, as it appears to reflect when the data were collected or requested rather than the closure of gauging stations.

L410–425 – This level of scientific interpretation seems excessive for a data paper. I suggest either simplifying the text by removing explanations of the underlying causes or adding references if this interpretation is considered essential to the analysis.

L414 – Please provide a reference to support "likely related to an underestimation of convective heavy rainfall events".

L515 – "1 October to 30 September of the following year": Please also define the water year at the hourly time step. If it starts at 00:00, state explicitly that it runs from 00:00 on 1 October to 23:00 on 30 September of the following year.

Dataset comments

The metadata columns and number of stations all check out; I found no issues.

Please ensure that the names in Table 3 are consistent with those in the files. Some differ: climatic versus climate; hydrologic versus hydrology; simulation_benchmark versus Benchmark Simulations; and topographic versus Location and topography.

The paper states that unavailable values are represented by NaN, but they are actually represented by blank entries. Either convention is acceptable, but the description and files need to be consistent.

I confirm that hourly data are available for 1,611 stations; I found no issues.

Please clarify that the simulated hydrological time series are provided in a separate file and folder, rather than in the time series files. Table 1 gives the impression that the simulations are also available in the CSV files.

The description at L320 and the column "perc_nan_precipitation_original" are confusing. The column contains 0 when the original data are available; when the value is greater than 0, only gap-filled precipitation is available. This needs clarification, as the paper gives the impression that 100% indicates full coverage.

The format used for the benchmark simulations is not particularly user-friendly. I understand the need to save storage space, but if the intention is for users to work with

these simulations, they could be provided in CSV format. If they are included only for verification and FAIRness, the current format is acceptable.

The catchment boundaries and gauging station files can be read without issues

I also confirm that ICON-D2 data are available for 1,611 catchments at a three-hourly resolution and cover three years and ten months, as stated in the paper. The rainfall time series I inspected also appear appropriate.