Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-71-RC2, 2021

Comment on essd-2021-71

Anonymous Referee #2

Referee comment on "CCAM: China Catchment Attributes and Meteorology dataset" by Zhen Hao et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-71-RC2, 2021

Large-sample hydrology is growing in popularity, with more and more hydrological data sets being released and made publicly available. Hydrological data from China are particularly rare, so this dataset is a welcome and timely addition. I commend the authors for the time they invested in the production and documentation of this dataset.

The dataset draws a lot of inspiration from the CAMELS datasets. The paper title sounds like the title of a CAMELS dataset and many of its components, including the variable names, the reference datasets, the layout of the maps, are to a great extent the same, which facilitates comparisons with the existing CAMELS datasets. The data set does seem to have a name or acronym, but it contains a dataset called Normal-Camels-YR, which encompasses normalised (and standardised) streamflow time series from 102 gauges in the Yellow River basin - it is important to stress that although attributes and forcing attributes are provided for thousands of catchments, streamflow timeseries are only provided for these 102 catchments.

In the abstract, the authors state their "dataset provides numerous opportunities for comparative hydrological research, such as examining the difference in hydrological behaviours across different catchments and building general rainfall-runoff modelling frameworks for many catchments instead of limited to a few". My concern is that the scope of this dataset (in its current form) might be more limited, because of the following restrictions imposed to the streamflow data:

- the streamflow records are standardised and normalised: while this may not be an issue for some machine learning algorithms, this makes the calibration of standard rainfall-runoff models challenging, since biases in the mean and standard deviation cannot be assessed.
- furthermore, from a water resources perspective, being able to quantify water volumes is essential, but it is not possible here because of the normalisation of the data.
- 7-day streamflow averages are provided (instead of daily data for the CAMELS

datasets), which makes flood characterisation and modelling difficult, as the weekly averaging will greatly smooth out the flood peaks.
- in a comment posted on 22nd June, the authors clarify that "40 basins are having over ten years record, and the mean length of the continuous record is ~7 years", meaning that the time series are fairly short.
- while the authors claim to enable the community to "examin[e] the difference in hydrological behaviours across different catchments", the restrictions outlined above make this task particularly difficult, in particular because they hinder the computation of the most common hydrological signatures - I encourage the authors to compute and make available hydrological signatures based on the unaltered daily streamflow time series.

I understand that releasing the true streamflow time series is challenging, but some decisions made by the authors are puzzling. For instance, "for confidentiality, the names of these basins have not been announced", but shapefiles are provided and give (presumably) the exact location of the catchments. Likewise, the mean streamflow can be inferred quite readily from catchment descriptors, which makes me feel that the normalisation of the timeseries is unnecessary.

Hence, I recommend that the authors do not use the name CAMELS, as all the CAMELS datasets provide daily streamflow timeseries for their hundreds of catchments, which many in the community see as their most important characteristic. There are many alternative naming options, including exotic animal names (as illustrated by the recent LamaH dataset - https://essd.copernicus.org/preprints/essd-2021-72/).

Data availability/reproducibility: The abstract mentions that "complement code for generating the dataset will be open-sourced such that the user can generate meteorological series and catchment attributes for any watershed within contiguous China", yet the conclusion makes it clear that the forcing dataset SURF_CLI_CHN_MUL_DAY is only freely available for Chinese researchers (L444). This is a non-negligible constraint. Furthermore, I don't see a paper documenting the SURF_CLI_CHN_MUL_DAY dataset, and the link provided (http://data.cma.cn/data/cdcdetail/dataCode/SURF_CLI_CHN_MUL_DAY.html) leads to page in Chinese.

Overall, I encourage the authors to lift the restrictions they have imposed on the streamflow data. This would significantly increase the appeal and uptake of their dataset.