Author's Response

Date: 2021-07-26
Manuscript Number: ESSD-2021-71
Title of Article: Catchment attributes and meteorology for large sample study in contiguous China
Name of the Author: Zhen Hao
Email Address of the Author: zhen.hao18@alumni.imperial.ac.uk

Dear Dr Gudmundsson and referees,

Both reviewers gave very pertinent suggestions, and we thank them for their efforts to review this article. After being reviewed, we re-reviewed our work and re-examined its contribution to the entire hydrological community. We are writing to describe how we have revised the paper and the parts we have emphasized and clarified in the article.

We want to restate the importance of this work: though streamflow data is critical for large-sample hydrology research, large-scale catchment characteristics data are also vital. As stated in the initial CAMELS paper (which calculated the attributes, but the streamflow data was released in previous work):

"Although catchment attributes are routinely used when working with a handful of catchments, there is a growing recognition that a large sample of catchments can provide insights that cannot be gained from a small sample. Large-sample data sets enable us to concentrate on catchment similarities and on the formulation of conclusions that are valid for a large number of (gauged and ungauged) catchments. Individual catchments can then be considered to be part of a continuum of catchment attributes, which vary in space along several gradients (such as aridity or soil depth). Working with a large number of catchments enables us to study changes along different gradients and to better disentangle the effects of catchment attributes on catchment behavior. This is particularly useful for comparative hydrology, i.e., to identify how similarities and differences between locations influence ecohydrological processes. Further, large-sample hydrology opens new opportunities for data analysis and, for instance, makes it possible to explore interrelationships between catchment attributes on the basis of their spatial patterns, as exemplified later in this study using map comparisons."

There will be no large-scale streamflow data sharing within China in the foreseeable future, but Chinese researchers do have large-scale streamflow data. What hinders large-sample hydrology research is well-organized basin attributes data. The organization of the data into a catchment scale is the first of this kind in China. This work will also help researchers in other regions calculate watershed attributes by referencing the public code, especially attributes that are not trivial to calculate (e.g., p_seasonality). **As far as we know, there are already researches based on our data**.

To further improve the usability and influence of the code, we have reformulated the code such that the user can generate a basin's characteristics with just a "one-click" when the required source data are prepared, which will significantly improve the accessibility of the catchment attributes.

Due to the strict redistribution policy of streamflow data, we are afraid not to be able to release the original streamflow data and the mean streamflow. We must ensure that the source data is not released, but we want the released data to be useful, so we present the current solution. The current data can be used in such a situation: when it is desirable to verify the generalization ability of a machine learning model on a global scale, HydroMLYR (new name) can support the verification of the performance in the Yellow River Basin.

There are also some other shortcomings that can be fixed. Next, we respond to these questions one by one:

(1) **RC**: Wrong citation style when the citation appears at the beginning of a sentence. **AC**: "(Kratzert, Klotz et al. 2019) shows" should be "Kratzert, Klotz et al. (2019) shows" and similar for others. **Changes**: L55, L60, L71, L72, L74, L193, L196, L198

(2) **RC**: Texts in Figures 3,4,5 are hardly readable. **AC**: We have fixed it by redrawing these figures. **Changes**: Figures 3,4,5

(3) **RC**: The irrationality of using the Pearsons Correlation Coefficient to access the correlation between catchment attributes. **AC**: Although the Pearson's correlation can only provide a complete description of the association when both the two variables are standard, we think the most doubtful part of using it is that it assumes a linear relationship (a change in one variable will cause a proportional change to the other). Because there are too many variables, we can't plot the scatterplot one by one to check whether the relationships are linear. We think Spearman or Kendall's Tau may be more suitable in this case due to its wider application range. Even if the relationship between the two variables is linear, Spearman or Kendall can also return a very close result to Pearson. However, if the relationship between two variables is only monotonic, Pearson will have information loss. **Changes**: L595-L615 (Appendix B)

(4) **RC**: The suggestion of not using the name "CAMELS". **AC**: The paper title sounds like the title of a CAMELS dataset, and the use of "Normal-Camels-YR" might be misleading. We suggest using CCAM to stand for "China Catchment Attributes and Meteorology dataset" and HydroMLYR to stand for "Hydrology dataset for Machine Learning of the Yellow River Basin." The new names may avoid readers' wrong expectation of the data set and more clearly indicates the purpose. **Changes**: L1

(5) **RC**: 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 a page in Chinese. **AC**: We are sorry that we made a false statement about the facts at the beginning. We found on the registration page that foreign researchers can also register, but the interface is still in Chinese, which is out of our control. The SURF_CLI_CHN_MUL_DAY data was issued by the National Meteorological Information Center of the China Meteorological Administration (NMIC/CMA). The

data is quality controlled, and it is widely used in research in China. However, it does not have a related paper. **Changes**: L435

(6) **RC**: Some decisions made by the authors are puzzling. **AC**: In fact, the location of hydrological observation stations can be observed through remote sensing satellite images and then combined with HydroSHEDS's River network to determine their location. Then the boundaries of the basin can be determined based on the publicly available DEM, but we cannot release the names of these hydrological observation stations; these are sensitive information. **Changes**: We do not think it is necessary to explain this matter in the article.

In the past few days, we have made extensive efforts to reorganize our code. Combined with the data set that has been released, we are aiming to achieve two goals:
(1) Researchers can quickly obtain catchment attributes and meteorological time series of the local catchment from our data set.
(2) If the local catchment has a custom boundary, using our code can calculate the catchment attributes and meteorological time series quickly based on the given boundary. Our code currently supports one-click generation of all static attributes, as long as the required source data has been prepared according to the instruction, and the generation of the meteorological time series can also be quite effortless.

https://github.com/haozhen315/CCAM-China-Catchment-Attributes-and-Meteorology-dataset

Best regards,
Zhen Hao