Review Report

Journal: ESSD
DOI : https://doi.org/10.5194/essd-2021-72
Title: LamaH | Large-Sample Data for Hydrology and Environmental Sciences for Central Europe
Authors: Christoph Klingler, Karsten Schulz, Mathew Herrnegger
Date accepted to review: 2021-03-18
Date review submitted: 2021-04-04
Recommendation: Revisions

Summary

In this dataset, Klinger et al. 2021 provide hydrometeorological time series and environmental attributes for nearly 900 catchments across 170,000 km^2 in Austria and the Upper Danube. They compiled and processed an impressive array of indices and hourly + daily resolution datasets to enable large-sample hydrological studies in the area. By computing data for three types of catchments (full drainage area, immediate catchment, and reference non-impacted catchments), they allow the use of their dataset for several types of studies. The manuscript includes extensive discussion of the sources of data and processing steps, as well as considerable assessment of the uncertainties in the data. The dataset was easy to download and use, and appears complete and robust. I applaud this massive effort by the authors and recommend this dataset for publication, as I believe that it will further bolster the growing data ecosystems for large-sample hydrology.

Nonetheless, I point to several areas that deserve improvement and clarification in the manuscript as well as various (mostly small) inconsistencies in the dataset that need to be addressed prior to publication. Please find my major comments below, and more detailed comments on the manuscript itself in the attached PDF. I also provide summary files of my inspection of the data (.Rmd and .html) which do not need to be addressed by the authors but may be useful for them to reproduce the few errors I found.

I look forward to using this dataset and am available for future requests regarding the publication of this dataset.

Mathis L. Messager

Major Comments

Template from https://www.earth-system-science-data.net/peer_review/review_criteria.html

1. **Read the manuscript**:
   Are the data and methods presented new?
   Yes. To my knowledge, this is the first dataset of this kind for the region covered, using state-of-the-art datasets.

Is there any potential of the data being useful in the future?
Yes. This dataset will be useful for hydrological studies and other studies involving hydrology in the region.

Are methods and materials described in sufficient detail?
Almost all methods and materials are sufficiently described.

i. More detail is warranted to describe how catchments (A) were delineated for the gauges (L114, see my comment on attached PDF).
ii. The descriptions of several other important processing steps could benefit from more clarity (see attached comments, but e.g. L208).
iii. For reproducibility, the code for this study should be made available (e.g. on Github, and/or in the Zenodo repo) and a code availability statement included in the manuscript.

Are any references/citations to other data sets or articles missing or inappropriate?
No, the literature is well cited, and all datasets are referenced appropriately with a DOI or link when available. If available, please add a link to the HAO reference.

2. **Check the data quality**:
Is the data set accessible via the given identifier?
Yes, I programmatically downloaded the dataset directly from the identifier.

Is the data set complete?
Yes. I ensured that all files were included and matched the structure and description in the manuscript and metadata file. I inspected all main files in R and/or ArcGIS. Find attached an HTML and Rmd file of the data inspection process (but no need to address it beyond my comments on the manuscript PDF and herein). I also visualized a random sample of all time series data, which appear complete. I only checked the daily dataset but assume that my observations extend to the hourly dataset.

One addition to the dataset (gauging station attributes), which I think would greatly enhance its usability, is the inclusion of the river network IDs (e.g. HYRIV_ID) of the segments nearest to the gauging stations (based on the EEA and RiverATLAS networks). This procedure can be done semi-automatically and would not take long for the 859 stations in this dataset (first snapping to nearest segment in network, then checking drainage area ratio and manual adjustment of erroneous matches).

Are error estimates and sources of errors given (and discussed in the article)?
Yes. The amount of care given to describing sources of uncertainty and error is great in the manuscript. For instance, the inspection of the time series for glacial and/or anthropogenic impact on sub-daily variability in discharge is quite unique for this kind of dataset (to my knowledge).
However, it is important to note further in the manuscript that very limited data quality checking was performed by the authors for the streamflow gauging stations time series. I do not suggest that additional QA/QCing be performed, simply that this be mentioned as a disclaimer. At the moment, the only QA/QCing performed on these time series originates from the providing

agencies. Gauging stations time series are notoriously prone to artefacts (for various reasons), and additional QA/QCing is often needed (e.g. Gudmundsson et al. 2019, Zimmer et al. 2020). I recommend that additional resources be provided to readers for assessing the extent and nature of this checking if possible.

Gudmundsson, L., Do, H. X., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment. *Earth System Science Data*, *10*(2), 787–804. https://doi.org/10.5194/essd-10-787-2018

Zimmer, M. A., Kaiser, K. E., Blaszczak, J. R., Zipper, S. C., Hammond, J. C., Fritz, K. M., Costigan, K. H., Hosen, J., Godsey, S. E., Allen, G. H., Kampf, S., Burrows, R. M., Krabbenhoft, C. A., Dodds, W., Hale, R., Olden, J. D., Shanafield, M., DelVecchia, A. G., Ward, A. S., … Allen, D. C. (2020). Zero or not? Causes and consequences of zero-flow stream gage readings. *Wiley Interdisciplinary Reviews: Water*, *7*(3), e1436. https://doi.org/10.1002/wat2.1436

Are the accuracy, calibration, processing, etc. state of the art?
Yes, the extent and quality of data processing are impressive in their scope. I commend the authors' effort in bringing together so many data sources.

Without further explanation nonetheless, I am warry of the catchment delineation process and computation of river network characteristics. See my comment on the attached PDF. In brief, I am worried that multiple datasets were used, and that custom catchments were not delineated for the precise locations of the gauges (i.e. using drainage direction grids) but rather aggregated from existing catchment polygons which may not fully match the position of the gauges. See Lehner (2012) for an example protocol for catchment delineation.

Lehner, B. (2012). *Derivation of watershed boundaries for GRDC gauging stations based on the HydroSHEDS drainage network (Report 41)*. Global Runoff Data Centre in the Federal Institute of Hydrology (BFG). https://www.bafg.de/GRDC/EN/02_srvcs/22_gslrs/222_WSB/methodology_Lehner.html

Are common standards used for comparison?
Yes (e.g. Fig. 4)

3. **Consider article and data set**:
Are there any inconsistencies within these, implausible assertions or data, or noticeable problems which would suggest the data are erroneous (or worse). If possible, apply tests (e.g. statistics). Unusual formats or other circumstances which impede such tests in your discipline may raise suspicion.
Given the scale of the data processing involved, I can say that the dataset is robust overall and well put together. The inconsistencies I pointed out (see attached PDF, e.g. soil texture fractions > 1, NAs for 13 stations in beg_2017.csv table, etc.) do not worry me, although they need to be addressed.

4. **Check the presentation quality**:
   Is the data set usable in its current format and size?
   Yes, the formats (.csv, .txt, .shp) are all very standard and allow great interoperability, which is appreciated. The structure of the data is intuitive and easy to use. The dataset is not overly heavy and appropriate data storage formats were chosen.

   Small comments on format:
   i. At the moment, the directory containing the dataset is called CAMELS_AT. Please correct.
   ii. In Gauge_hierarchy.csv, NEXTUPID is character while NEXTDOWNID is integer. In 3_shapefiles/Gauges.shp, ID and NEXTDOWNID are character. Please make sure that these are consistent for correct joining.

   Are the formal metadata appropriate?
   I do not believe that "formal" metadata (e.g. ISO 19115 and ISO 19139) are provided with the dataset, beyond an information text file.

5. **Check the publication**:
   Is the length of the article appropriate?
   The article is long and could benefit from being more concise. I truly appreciate the extensive effort that the authors have put into the description of the dataset and area of study; it's a massive effort. Yet I have highlighted a few examples of sections that could be removed without significant loss of clarity/information. In particular, I think that many sections that focus on describing the distribution of variables and environmental characteristics of the study area are extraneous, as well as Table 1 (and the mention of regions in general; maybe transfer that content to another technical documentation together with the dataset?). The quantity of such information buries the key parts of the processing and/or data characteristics most relevant to the reader.

   If the authors deem it necessary to keep all, I recommend a re-structuring of the manuscript below to allow users to extract key aspects of the study more easily.

   Is the overall structure of the article well structured and clear?
   As mentioned above, I found it difficult to identify parts of the manuscript that are essential to understand the composition and assembly process of the dataset vs. uncertainty/quality checking and descriptive aspects.

   I highly recommend re-structuring it so that the essential information required to understand its creation and usage be concentrated in one main section, and additional uncertainty/quality assessments be in another section (e.g. water balance analysis, Budyko curve, etc.). In addition, it would be useful to more explicitly structure sub-sections of the manuscript that refer to the basic definition and processing of source datasets vs. more in-depth discussion of their provenance.

   Last, paragraph structuring within section could be better utilized to clarify the different steps (e.g. a paragraph for each type of basin delineation).

   Is the language consistent and precise?

In terms of writing, I think that some improvement is needed in terms of consistency and grammar. Again, I want to re-emphasize that this is a manuscript of considerable size and appreciate the challenges that come with harnessing such a vast amount of data. I highlighted several errors in terms of grammar or sentences that needed clarification in the text, but recommend an other copy-editing pass.

In terms of language substance, I pointed out some issues of understanding I had with the use of some terms that I think other readers could also struggle with (i.e., headwater catchments, orographic catchment).

Are mathematical formulae, symbols, abbreviations, and units correctly defined and used?
Yes.

Are figures and tables correct and of high quality?
Yes. I applaud the authors for the consistency and effort that went into putting these figures together. See my comments on the attached PDF for improvements I recommend.

***Finally**: By reading the article and downloading the data set, would you be able to understand and (re-)use the data set in the future?*
Yes, absolutely, and I certainly will. Good job!