Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2022-81-RC1, 2022

Comment on hess-2022-81

Anonymous Referee #1

Referee comment on "CAMELS-Chem: Augmenting CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) with Atmospheric and Stream Water Chemistry Data" by Gary Sterle et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-81-RC1, 2022

Evaluating the overall quality of the preprint ("general comments"),

Sterle et al. present a compiled novel dataset of water quality solutes and atmospheric deposition inputs for the CAMEL catchments. Their work augmenting existing and widely used CAMELS datasets is needed for further research in analyzing spatial and temporal water quality trends in minimally impacted watersheds. Existing papers have done large-scale water quality analyses, but few have provided open-access datasets and the breadth of solutes.

At its core, this is a data paper. As such, I think the methods need to be expanded. My comments primarily pertain to the data and methods, as I see that this is the paper's novelty.

The CAMELS dataset is widely used, and the addition of water chemistry provides the opportunity for analysis. However, I think the dataset could be improved substantially. This paper by Sterle et al. provides an excellent resource for the community.

Methods and Results

Length of dataset

In the paper, the dataset is stated to end in 2014 (see **Data Comments** below because I'm not sure if this is accurate). However, in many cases, solute and discharge data are available until the at least end of the NADP reporting period. Soon 2014 will be 10-years ago, and I worry the data will not be quickly obsolete and not used to its fullest capacity. The value of this dataset would be exponential if the authors were able to harmonize data from various agencies.

The USGS has developed methods to generate the longest timeseries possible for watersheds. However, I do not see these methodologies applied to these CAMELS watersheds.

Three different approaches can be used alone or in concert to expand the dataset.

- The following USGS report looked at the statistical differences between similar solutes. They were then able to merge solutes that were registered under different parameter codes but were statistically similar, allowing for a richer dataset.
- Secondly, the USGS report also has a methodology for pairing gauges, and sampling locations are not co-located.
- With dataRetrieval package in R you can pull USGS, state, tribal, and NGO water quality measurements. This can potentially expand the dataset. However a lot of manual data cleaning is required because of the poor metadata for non-USGS agencies (ex. nitrate can be reported as N or NO_3 , and not explicitly stated in the metadata) (Sprague et al. 2017). With the data harmonization from multiple sources, I would expect a section on the methodology of data cleaning.

- Oelsner, G. P., Sprague, L. A., Murphy, J. C., Zuellig, R. E., Johnson, H. M., Ryberg, K. R., Falcone, J. A., Stets, E. G., Vecchia, A. V., Riskin, M. L., Cicco, L. A. D., Mills, T. J., & Farmer, W. H. (2017). Water-Quality trends in the nation's rivers and streams , 1972 – 2012 — Data Preparation, Statistical Methods, and Trend Results Scientific Investigations Report 2017 – 5006. U.S. Geological Survey Scientific Investigations Report 2017–5006. https://doi.org/10.3133/sir20175006

Figures capture the various hydrological/biogeochemical metrics for select solutes in section 4. However, as a user, I find it challenging to evaluate whether the data would be sufficient for my use case. The authors have included some analysis of data coverage in section 3.1, however, since the strength of this paper lies in the data there could be more information to help users evaluate whether this dataset suits their needs.

There should be summary figures for all solutes, so users can adequately assess whether the dataset is appropriate for their use. I think the paper would benefit tremendously if the dataset had more metadata and signatures/summary statistics. Specifically, in Addor et al. (2017), the authors summarized many indices and described the indices in great detail (see Table 2 and Table 3 in Addor et al). I suggest summarizing information about (1) missing data/data gaps, years of continuous data, (2) low/high flow distribution, (3) FDC the WQ spans (figure 7), (4) seasonality of hydrology and solutes, and other metrics that the authors deem useful.

Data Comments

These comments pertain to the files on the google drive.

Data provided has inconsistencies in the way the date is reported. Example from camel_chem_v3.

- Gauge ID 14309500: sample_timestamp reads 8/15/67 15:00 whereas other dates are in 2010-12-11 format.
- Gauge ID 6447000: sample_start_dt appear to indicate that the data from 1950s. If so, this falls outside of the time period listed in the paper.

It appears that the data available in camel_chem_q_v1 ends in 2018. Please update the manuscript with the correct dates if this data is available.

The original CAMELS dataset provides Shapefiles, however, to allow for seamless merging of data, the header used to identify which column the watershed IDs are the same as the name used by the original CAMELS dataset.

The dataset provided should be able to stand on its own without needing the other CAMELs dataset. The watershed metadata should thus be included (area, outlet latitude and longitude, USGS gage number with leading 0s).

Individual scientific questions/issues ("specific comments"),

Line 84-87: Authors state that they have WQ data from 1980-2014. Later the authors state that the data is "for the same time period" as NADP data (1985-2019). Dates should be consistent.

Line 103: Section 2.2 should be written in a high-level abstract way. I find it unclear how these frameworks are applied to this data in the way it is currently written. I would be better to see more specificity and names of the database (ex. NADP, NWIS).

Line 125: It is unclear whether "daily average discharge" means a continuous dataset or just discharge measurements for the data that there are solutes. The dataset provided suggests the latter, but I think there would be value in providing the daily discharge timeseries for the same timespan of the solute data.

Line 135: In Table 2, deposition units are reported as mg/L. However, NADP reports their deposition in both concentration and kg per hectare. Are the units in Table 2 a mistake? If not, can you add some detail on the methods used to convert concentration to an area normalized load?

Table 1: Consider added the NWIS parameter code. For example, is "Nitrate, water filtered" the nitrate plus nitrite (00631) or just nitrate (00618)? There are many parameters for slightly similar solutes and it would help with reproducibility if the parameters codes were included.   Also, consider listing the difference between pH in the field and pH in the lab for users.

Table 1: Also consider adding more detail to units. For example, is nitrate mg-NO3/L or mg-N/L.

Line 204: EPA link is broken. I have had many issues with direct links where they are archived and become a dead end. I highly encourage the authors to find a paper with a DOI to support this sentence. As a starting point, you can consider:

- Baumgardner, R. E., Lavery, T. F., Rogers, C. M., & Isil, S. S. (2002). Estimates of the Atmospheric Deposition of Sulfur and Nitrogen Species: Clean Air Status and Trends Network, 1990−2000. In Environmental Science & Technology (Vol. 36, Issue 12, pp. 2614–2629). https://doi.org/10.1021/es011146g
- Lloret, J., & Valiela, I. (2016). Unprecedented decrease in deposition of nitrogen oxides over North America: the relative effects of emission controls and prevailing air-mass trajectories. Biogeochemistry, 129(1-2), 165–180.

Technical corrections ("technical corrections": typing errors, etc.).

- Subscripts for solutes should be consistent throughout the manuscript.
- Table 3 formatting caused solutes to be cut off.
- Line 18 and 149-151: 18 solutes listed in the abstract, 17 listed in Table 1 and in Line 149, and 16 listed in the text. Make them all consistent.
- Line 57: Remove (?).

- Line 60: Remove “CITE”
- Line 168, 174 and others: When referencing figure (ex Figure 2), please add the panel letters (a,b,c, etc.).
- Figure 2: Panel a, why does daily average discharge only have 393 watersheds while the original CAMELS dataset uses USGS discharge in the original CAMELS dataset?
- Figure 6: Regarding NO3, if arid and humid sites are a subset of all sites, I am unsure how the slope for all sites can be larger than both arid and humid.

Citation

Sprague, L. A., Oelsner, G. P., & Argue, D. M. (2017). Challenges with secondary use of multi-source water-quality data in the United States. Water Research, 110, 252–261.