Response to Reviewers

Two reviewers (including one from the previous round) have evaluated your revised manuscript and provided detailed comments. While they both agree that significant improvements have been made compared to the original version, they still find the manuscript has some shortcomings. Both reviewers noted a lack of details in the Methods and Results sections (mostly sections 2 and 3), which led them to comment (negatively) in different ways: one found that the lack of details made your work not reproducible, while the other felt that your dataset, as described, may appear as being no more than a "merge of USGS flow and water quality data", which is something that can be done using an existing R package (and hence is not really novel or value-added). Obviously, the CAMELS-Chem dataset is (or has the potential to be) *much more than a simple data-merging exercise*, but the manuscript in its current state is underselling that potential (I do share the reviewers' opinions on that). One of the reviewers sees your paper as mostly a data paper while the other thinks that submitting your paper to HESS (rather than Earth System Science Data (ESSD), for instance) means that the goal of the research should go beyond just the presentation of the dataset. Personally, I think that your paper would be most impactful as a hybrid one between those two formats, but this makes it tougher to write because you need to articulate not only technical details pertaining to the creation of the dataset, but also hydrological and earth system process details. Both reviewers have included numerous comments for your consideration, so I invite you to carefully look at them and see how you could address them. I do look forward to receiving and reading your revised manuscript and response document, which will be sent out for another round of review.

With best wishes,

Genevieve Ali

Dear Dr. Ali,

We appreciate your thoughtful feedback on our revised manuscript and the patience of the reviewers to improve the manuscript. We have worked hard to increase the research relevance of the manuscript by 1. Including research questions, 2. More clearly answering those questions, and 3. Giving more examples in Section 4 to highlight the broad research applications. We believe the newly revised manuscript fits the sweet spot between a research article and a dataset release.

Reviewer #2

General Comments
The authors describe an extension of the well-known CAMELS data set with water chemistry and deposition data. I was reviewing a first version of the manuscript already and still think that this is a super-valuable data compilation. However, the manuscript still has, from my point of view, some serious shortcomings. My expectation is that the manuscript shows how data was merged and gives an overview on the dataset. The method is, to my surprise, not reproducible and remains very vague. The results are mostly displayed in a visual way - I very much miss a quantitative overview on the basic descriptive statistic, on the sample frequency per year. I also miss a brief overview on what information are delivered with the original CAMELS data set. So, it is hard for the reader to say if the dataset actually fits his or her needs. Finally, I have problems with the embedded two examples. Motivation and methods are given in chapter 4 and are not aligned with the introduction and methods of the entire manuscript. Results of that examples are not much more than two plots and the statement that others may interpret that. So, overall I think this manuscript needs some serious revision.

We appreciate the reviewers thoughtful comment on the manuscript. We have added considerably to Section 2 to improve the reproducibility of the results. Section 4 is also better aligned with the introduction and methods are made more clear.

Specific Comments
Abstract
L20: Potentially remove the first "and" here.
Edited

L21ff: I cannot fully follow which constituent is given with full name and which is not - consider to homogenize that. Nitrate is given first as a full name, then as the short chemical form and then again as full name.
This was made more consistent.

Introduction
L45: GLORICH (among other water quality databases) is part of GRQA. Not sure how to address this here but I encourage you to make this more clear in this sentence.
Changed

L50-55: You mix water quantity data efforts such as CAMELS with water quality AND water quantity data efforts such as Ebeling et al. here. I think this need better separation and clarification. Why should an atmospheric deposition dataset be relevant for water quantity? Note that atmospheric deposition of nitrogen is part of Ebeling et al.

Added to line 60: “Furthermore, atmospheric deposition data is available for CONUS but has seen less inclusion in such data sets, despite the significant impact of atmospheric contribution to stream chemistry (Shao et al., 2020). “

L75: There is something wrong here. Please check the whole sentence.
Corrected

L85: Be precise - do you mean stream "flow" data?
Yes, corrected.

L87: This is a bit unclear. Why 2018, when CAMELS ends in 2014?
We are trying to maximize the available data.

L90ff: As already stated in the first review round I am not convinced to publish this as a research paper if the manuscript does not attempt to set itself a scientific goal. I think the two examples needs a motivation within the introduction section already and should be mentioned explicitly. The examples are not totally randomly selected as written later in chapter 4!
We understand this criticism and have added research questions that are more clearly dissected in Section 4.

Materials and Methods
Please note this section mas majorly edited.

L105: That is a nice statement but I wonder why it is made here? What is the consequence?
Changed this section completely.

L116: "Water resources data" is not entirely clear for me.
Changed

L96-114: This chapter leaves me a bit puzzled. This is not about "data sources and description". Would there be a more fitting header? Where are actual multiple data sources described?
Majorly edited

L126f: I thought discharge data is the key of CAMELS (see also line 47)? Why was discharge not available in all 671 catchments anymore while it was available when CAMELS was published?
There are periods of the CAMELS records without discharge records.

L128: How many catchments had data before and after 1980-2018? Or is this coverage part of the result section?
We focus our discussion on 1980-2018 but add Figure S3 to also give information on the dataset before 1980.

L121ff: The methods should describe you workflow in a way that it is reproducible for the readers. This is not the case. Also matching fig. S1 is not of big help here. I think you need to state what exactly you did to the data. Fig. S1 lists aggregation, filtering, cleansing, profiling, joining and sorting. At least the first three processes are not described in the text.
Agreed. Section 2 was majorly rewritten.

Results
L141: I would not call discharge a general water quality parameter.
Changed

L143: Also give full names of Cl and the cations.
Not sure we follow this suggestion. We have consistently use abbreviations such as Cl.

Table 1: You state for some constituents "Water, filtered and total" - is that visible in the data base? Why is magnesium measured in the suspended sediment? This makes it hard to be analysed similar to all other cations. Why is not unit and database abbreviation given for discharge? Why is alkalinity not stated here?
Thanks for catching a mistake with Mg. Alkalinity was added to this table.

L146: Mention the fig. S5 here specifically. It takes a while finding the right figure when just referring to "supplementary materials"
Added

L148f: Something went wrong here. 325,477 catchments?
Corrected

L155f: Interesting that you start with the spatial distribution. For sure very important. However, does it make sense to start with descriptive statistics on the data? A table such as table 1 but station median, percentiles, number of observations...? Especially as all catchments aim at pristine areas and do not incorporate too much direct human impacts.
We have added what is now Table 3.

Fig. 1: Given the different catchment size, I recommend to give discharge per area (e.g., mm/a) in 1a. Otherwise this is not very informative. I struggle with the colors vs. size. More intuitive for me would be the value represented by colors and the CV by symbol size. However, this is up to you. For pH it does not work well - but here you can maybe scale the symbol size between min and max similar to the other plots.
We agree that colors versus size is in the eye of the beholder. We elected not to change the discharge units.

Fig. 2: Why is this coming after fig. 1? Seem to be a better fit for the methods or an earlier point in the results.
We switched figure 1 and 2.

L158: These are maybe not the best examples as DO is having strong physical constraints and pH is a log-unit.
Agreed, these were changed to Cl and Na.

L183ff: From point of view of the USGS these are impressing numbers. However, the reader would be more interested in the average number of samples per station per year.
Agreed. We have summarized this in Table 3.

L188: Can you give some of the published coverage of different hydroclimates from the original CAMELS references? What range do you really cover here?
We added Figure 5 in the previous revision to exactly summarize this concern.

L225: Further above you stated that you had a large number of chemistry samples what do not have a matching instantaneous discharge observation. How do these numbers (here >90%) fit together?
Changed.

L240-245: This section mostly contains data source, interpolation methods, GIS-matching. All this belongs to the method section and not to the results.
Agreed and moved.

Example analyses
L275ff: All this needs a proper method description in the methodology chapter. I strongly miss a proper description of the actual results (only a weak verbal description and a link to Fig. 7) and a scientific interpretation of the results. Just stating that more sophisticated methods may reveal underlying controls is not enough here.
This section was rewritten

L278: Maybe it slipped through my read but what is FDC?
Added in methods.

L303ff: All what I stated for the first example is also valid for the second one.
Agreed and edited.

To Authors:

The CAMELS-Chem dataset will be a useful tool for catchment research, and I look forward to seeing this paper published. What appears to be the most novel and intensive part of this work is the addition of atmospheric deposition inputs to the CAMELS catchments. Mirroring comments from previous reviewers, I think that presenting this work as a data paper is the most appropriate path. As such, I also still have some comments about increasing transparency of the methods.

As someone familiar with accessing large amounts of USGS flow and water quality data from the NWIS, I think that the paper should make it more clear how the dataset they present is not just a compilation of flow and water quality for the same group of catchments. Be very explicit and detailed with what was done for data cleaning, interpolation, gap filling, etc. I find the language about ETL in Section 2.2 confusing, mainly in that the end product that I saw (three CSV files on Hydroshare) seems fairly straightforward, whereas the described end product (a relational database using PostgreSQL) sounds very fancy. I fully acknowledge my ignorance of the ETL methodology, but as someone who would be very excited to use CAMELS-Chem, I'd like this section to paint a picture of what exactly is going into the production of your dataset/database, and what I would be downloading to use it.

Regarding the deposition data, that section is also scant on details. What is the resolution of the input data, and what resolution are you reporting the data? The data was also not available for me to review in your Hydroshare link, so I was unable to comment on its formatting.

Again, from my perspective, doing a quick merge of USGS flow and water quality data would be quite easy for someone using the R package "dataRetrieval", and so I think you need to make clear that your dataset is value-added: mainly from the addition of deposition data, but also through your data processing/cleaning procedures and the construction of the relational database. Make sure to clearly sell CAMELS-Chem to the audience! I look forward to seeing this work progress.

We have worked diligently to address the concerns mentioned by rewriting the methods. We have also changed the introduction to better highlight the value of the dataset.

Comments:

1. Line 18. Maybe in the abstract say how many sites are in CAMELS-Chem? See Comment Line 97.

Added

2. Line 24. “Annual deposition loads and concentrations” are promised, but they are not included in the included database, nor are they mentioned in Methods section. See comment from line 239.
   Included in database

3. Line 75. “Thus, we now have There are opportunities”. This text should be fixed.

Fixed

4. Line 97. Cite the number of sites from CAMELS, and then CAMELS-Chem (516 sites cited at Line 84)

516 is the correct number

5. Line 101. “most of its sites are drawn come from” USGS HBN. Remove the word “come”. Also, how many sites from HBN?
   Changed to HCDN to match Newman et al. description.

6. Line 103. How many sites from NWIS or HCDN?

They are not mutually exclusive, but essentially all of them.

7. Line 116. You list the challenges: missing data, mis-matched sample times, inconsistent parameter names, or varying units of measure. Of these, later in the paragraph I only see sample times and units of measure discussed. Did you fix parameter names? Or fill missing data? Please elaborate.

Much more detail was added to this section.

8. Line 127. How do the 506 and 488 become 516 sites? More detail please.

Only 506 have coincident discharge and water chemistry but 516 have water chemistry.

9. Line 130. “impute”. Could you use a more colloquial form for this sentence to improve readability?

We clarify that imputing is fixing missing data.

10. Line 131. Forgive my ignorance, but does the Hydroshare upload represent your final data repository, using PostgreSQL? Having looked at your CSV files, which I found well organized (see some minor comments later about the dataset), I’m not sure if there’s something I’m missing.

You are correct, but the database is now available for download.

11. Line 148. Could you include the number of catchments and number of measurements for all the parameters, as you've done for Si and DOC, into Table 1?

New Table 3

12. Table 1. Units and abbreviation and USGS code are missing for discharge

Corrected

13. Table 1. In a response to a previous reviewer comment about merging solutes, you justify using unfiltered and filtered parameter codes (e.g. chloride, sulfate, nitrate). This justification needs to be in the manuscript, explaining why there are multiple parameter codes in the last column of Table 1. See also my comment on the database, about my confusion about the column "total_no3" and others.

Corrected with citation for combining constituent codes.

14. Line 225. This sentence needs to be corrected (referring to right-most column or bottom-most row). This does not reflect the information in the table. Should it be second column, and then second row? I might advise moving Q to first row/column before Temp.

Corrected

15. Table 3. As mentioned elsewhere, leading zeros should be present for USGS parameter codes.

Added to database as gauge_id2

16. Line 239. Section 3.5. Atmospheric deposition data. This paragraph should be included under Section 2 Methods. In describing the NADP data, what is the data resolution (temporal, spatial)?

Added to method

17. Table 4. Table 4 should probably be in Methods, rather than Results?

Added to methods as Table 1

18. Figure 1: The inset histograms referenced in the figure caption (and shown in manuscript version 1) are no longer present. They should be added to the figure.

Removed from captions (in Figure 1).

19. Comments on the database. Mirroring previous reviewer comments, the data should be able to stand on their own, and so the following comments are requesting context and explanation to improve readability.

All of the following comments were addressed by a new data release on Hydroshare. We appreciate the reviewers attention to detail.

19.1. The paper discusses atmospheric deposition data (measured concentrations and annual flux estimates), but those data do not appear anywhere in the dataset.

19.2. In the “metrics” file, the following columns have no explanation: q_inst, q_15, q_derived, q_derived_note, q_daily_note, q_daily_cd, q_inst_cd, q_15_cd, measure_unit_code, sample_start_dt, sample_start_time, sample_timestamp, q_inst_ts, q_15_ts, inserted_ts, updated_ts, gauge_id2

19.3. In the “dataset” file, there is a second column “gauge_id2”. This column is not described in the “metrics” file, or how it is different from “gauge_id”

19.4. Leading zeros. The gauge IDs should have the proper number of leading zeros. The USGS parameters should also have leading zeros. This applies to all CSV files in your database. This is critical for users not familiar with USGS standards, and might be looking for “940” instead of “00940” (for example).

19.5. The columns total_cl, total_no3, and total_so4 are confusing. My guess is that these columns are filtered and unfiltered values, harmonized? These columns are not mentioned in the manuscript Table 1, nor are the methods for calculating the column values in the manuscript.