

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2  
<https://doi.org/10.5194/hess-2022-81-RC2>, 2022  
© Author(s) 2022. This work is distributed under  
the Creative Commons Attribution 4.0 License.

## Comment on hess-2022-81

Anonymous Referee #2

---

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-RC2>, 2022

---

Overall comment:

This paper describes efforts to amend the known and quite successful US-wide CAMELS dataset with hydrochemical and deposition data. This effort is overdue and would greatly enhance the usage of the original CAMELS data as well. Having stated this I have to admit that the manuscript is not convincing to me. It fails to state what data (e.g., constituent codes in the USGS data) was used in what exact way (missing description of data evaluation, conversion, filtering). It visualizes data coverage but does not state number of observations and number of stations in a consistent way. So overall the reader is left rather unclear about the whole data handling process and the outcome. While I like the idea of three examples what this data could be used for, this leaves me rather puzzled. This is submitted as a research paper but does not really come up with research. For me this manuscript would rather fit the purpose of ESSD (Environmental Science System Data) as a dataset description than HESS as a research paper. For the latter it would have an appropriate structure and depths but still would need to acknowledge the details comments below.

Specific comments:

Abstract

The abstract needs to transport the content and motivation in a better way. It misses quite some information e.g. resolution of provided data. It is not clear what is really

provided in this new data set and what is already available from CAMELS.

Line 18: The linkage to the original US CAMELS dataset remains unclear in the abstract. Is this an addition to the original one as indicated by title and text here or is this something completely new.

Line 24: This is odd. Either give the exact number or >. Rounding to something easier to read would probably best.

## Introduction

From my point of view the introduction should state some examples of existing water quality databases. There are recent advances here such as: GRQA: Global River Water Quality Archive (Virro et al., ESSD), GLORICH (Hartmann, J., Lauerwald, R., and Moosdorf, N.: A Brief Overview of the GLObal RIVER Chemistry Database, GLORICH, *Proced. Earth Plan. Sc.*, 10, 23–27, <https://doi.org/10.1016/J.PROEPS.2014.08.005>, 2014. a, b) or QUADICA (<https://doi.org/10.5194/essd-2022-6>)

Line 31: There are quite a number of examples of nation-wide to continental scale water quality studies using more than a single catchment. I disagree here that availability of datasets does not go hand in hand with usage of this.

Line 35: Check citation formatting here.

Line 35: Can you be more specific on the "issues" mentioned here?

Line 42: Year in citation missing.

Line 43: Use the complete name of CAMELS here.

Line 56: Check the question mark here.

Line 60: Check citation here.

Line 83: I think you should better introduce the idea to provide atmospheric deposition data.

Line 86: Can you specify why to stop in 2014? Because the original CAMELS data also have that time frame?

## Material and Methods

Line 94: I would expect a citation to the NWIS data sources mentioned here.

Line 97: What is meant with the "two datasets"?

Line 97: "for each observation" seems unnecessary here.

Line 97: The match with the CAMELS station does not explain the promised "geographical coverage". Please adjust this sentence.

An option could be a reference to a figure showing the distribution of stations, maybe including the ones from the CAMELS data that are not amended with chemical data.

Line 102: From my point of view the 2.2 chapter on the methods is not informative and needs some revision. The reader does learn some fancy names but not what exactly was done with the data. The linkage of methods to the provided text-files remains unclear. This does not fit the previously stated wish to make everything reproducible by providing scripts. See also details below.

Line 103: I cannot follow this argumentation.

Line 109: Is ETL an established method? If so, please reference this. If not: Is there really a need for a fancy name for a standard data transformation method needed that uses a scripting language in a reproducible way?

Line 112: I do not see the link from this method to the resulting data provided, which are easy-to-use but simple text-files. This paragraph seems to be overly complicated.

Line 116: Similar as before. Either reference this as established methods or leave it. These abbreviations are not used further in the manuscript - so why introducing them here?

Line 124: Is type of discharge match (paired daily, instantaneous...) indicated in the dataset?

L131: "once per day" at the same time? Is this what is meant here?

Line 135: Again a reference to the dataset would makes sense for me. It would be, moreover, helpful for the readers to know about the absolute number of stations the interpolation is based on.

L 137: I do not understand what "align" means in this context. Please be more specific.

Line 139: Somehow strange to mention Table 2 first in the text while Table 2 comes second.

Table 1: It would be helpful, if not done elsewhere, to state the USGS parameter codes as this often causes confusion. The caption is not fully matching the column names, e.g. is abbreviation=attribute? Also it would be good if this is also consistent with or clearly linked to the column names in the provided data files, which seems to be not the case.

Line 145: The "+" seems unnecessary.

Dataset description

Line 151: This description says all what is needed. No further need for a table here.

Line 151: Please use subscript in the constituent's abbreviations similar to the table where the constituents are introduced. This applies to the entire manuscript.

Table 3: I do not get the meaning and idea of Table 3.

Line 164: This chapter would profit from a table (maybe combined with table 1) that lists number of stations and number of observations and maybe median number of observations per station for each of the constituents.

Line 171: I was not aware of this varying foci. Is there a reference for that?

Line 175: This sentence needs to be checked - the logic is not clear.

Line 186: While the previous chapter states reasons for different data density and distribution. Here, however, suddenly spatial pattern in the deposition is described. This belongs elsewhere. The chapter should describe the meta-information only.

Trends in...

Line 226: "not" missing?

Line 234: This reads quite strange. You compute cq relationships to check if there are cq relationships? Consider revising.

Line 250: I do not get the meaning of "overlapping solutes" here.

Line 251: Isn't this part of the A (attributes) in the CAMELS dataset?

Line 127: I very much like the analysis of FDC coverage. For me this is, however, not just an addition to the cq analysis but rather part of the dataset description in the former chapters.

Line 269: "response of the relationship"? Consider revising.

Line 270: What are uneven collection dates?

Line 299: This statement seems to be unrelated to stream chemistry and thus not directly a result of CAMELS-Chem.

Line 302: This sentence is very long. What exactly is meant by large-scale response? The observed response is always local, isn't it?

Line 404: This sentence is redundant, please revise.

Code availability:

I think the authors may already give an idea which repository they are aiming at.

