Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-43-RC2, 2022

Comment on essd-2021-43

Anonymous Referee #2

Referee comment on "The Surface Water Chemistry (SWatCh) database: a standardized global database of water chemistry to facilitate large-sample hydrological research" by Lobke Rotteveel et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-43-RC2, 2022

The authors present a newly created database on chemical composition of surface waters. The database is comprised of several database sources from which specific parameters/variables are extracted and unified for the specific purpose to provide a data base for surface water acidification research.

The collection and harmonization of data on water chemistry is very important to the research community, as it enables more refined global analyses of matter fluxes, temporal developments, climate change impacts, any many more.

The manuscript addresses an important data topic, which makes it worth to be published. However, due to the points stated below, I recommend a major revision.

Data quality
I would argue, from a personal viewpoint, that if the goal is to provide global coverage of data to enable global cross-boundary evaluation of surface waters, it may not be very important to have a high data quality, as the available amount of data will level out "outliers" or differences in the data analyses from a statistical viewpoint.

Data harmonisation
The calls for a unified approach in all future data collections are very noble, but I doubt that they will be heard. Data producing authorities very often have their own, historically grown structures and formats, that are so convoluted and unpredictable that it would be and hopeless to expect a globally unified data structure

Data selection
The authors state that the parameters were specifically selected to evaluate surface water acidification, however I would argue that the most important parameter in this regard is missing: total alkalinity (TA). This is reported in some of the used sources, even if it may be in awkward units sometimes. The TA is fundamental for the understanding of the carbonate system and the interaction of CO_2 and natural waters. Alternatively, dissolved inorganic carbon could be included, or both parameters, where available, to be able to calculate the missing parts of the carbonate system (TA and pH or DIC and pH enable the calculation of DIC or TA, respectively). Furthermore, the inclusion of TA would enable the calculation of a charge balance, which could provide an indicator for the data quality.

Database structure and presentation
I really appreciate the approach of publishing the scripts for the database of Github. This makes the work very transparent and should be an example for all scientists working with complex data processing.

The chosen format of the data is slim and straightforward, however, for the average enduser, the relational style of the files may present a potential problem as data cannot be filtered and used as is, but have to be transformed. It may be an advantage (not a requirement) to provide a python script that converts the data into the "classical" column-row-format. It may, however, increase the filesize to an extent that makes it hard to handle.

Regarding the units, the choice of weight units is okay but may lead to the need to recalculate to molar units as this is needed in geochemical calculations (e.g., charge balance, ratios, chemical formulas)

Text quality
There are several typos, duplications and wording issues in the text. I mention some of them below. Overall, the text could benefit from a revision, which clears out the errors but more specifically narrows focus on the specific arguments for the need of a new and harmonized database.

Specific comments

L8 2x "identify"

L18 Define the need for more data collection – how would that improve global models? Little data from arid regions may also be due to the fact that there are less surface waters

L19 “Environs”

L21/22 2x “address”

L29 “a number projected…” is meant to refer to the 4 bln people, but as it stands in the text it rather refers to “at least one month”

L30 “these resources” – which?

L36 Define “transboundary problem”

L47 When I comes to the fate and behavior of compounds in natural water, I would argue, the catchment scale is a good and proven approach. I may not understand the term “transboundary” in your sense, but why should be look transboundary if fluxes are “confined” in catchments anyway. Isn’t this the very idea of catchments to have all waters included in one larger scale area?

L49ff Yes, catchment waters will be influenced by land cover and geology, but so are observation on larger scales.

L51 “affected”

L79 I understand the point that the authors want to make here, however, the example may be a bit too tightly defined. Looking for “water chemistry database sweden” yields the website of the water information system VISS (https://viss.lansstyrelsen.se), I don’t know if data is extractable there but it seems that it is a good starting point. With this approach and a slight variation in search terms, more data should be discoverable.

L94f Can you state how much data was discarded, in %? Maybe leave the data in the dataset but provide a flag so that users can decide based on their needs?

L107 “simplied”

L107 2x “reduce storage requirements”

L129 Replace “standardized” with “harmonized” as probably most coordinates adhere to some kind of standard.

L146 Cost may be one reason but also, these are the most relevant parameters for many fields of research.

L148 What do you mean with under-reported results? Unclear.

L150f If no location data, I can understand the point. But w/o method information, it could still be interesting data in a global context (see argument above).

L156 Unclear logical connection between data gaps and discharge dependency.

L168 “people”

L168 “who collected” -> “collecting”