Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-51-RC2, 2021

Comment on essd-2021-51

Anonymous Referee #2

Referee comment on "GRQA: Global River Water Quality Archive" by Holger Virro et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-51-RC2, 2021

General comments

Virro et al. describe aggregating and harmonizing five national, continental and global datasets that can be used for global water quality models. Among the five selected datasets the GEMSTAT itself is a global database containing harmonized data from the contributing countries. The authors follow the ETL approach (Extract-Transform-Load) to bring the different data sets together. The methods are well described and data, metadata and scripts are available at the given websites. The authors suggest in their conclusions to transform the set of CSV files to a relational database in future what would further improve the usability of this dataset. I encourage the authors to do it. Also I like the idea to develop an online dashboard for GRQA. From a global modelers perspective the derived GRQA can be easily used. The authors selected 42 specific parameter relevant for modelling nutrients (water temperature, oxygen, nitrogen, phosphorous, carbon compounds ...). Still, GRQA cannot solve the problem of data scarcity in Africa, Asia and South America and also the suggested machine learning cannot help here.

In general, the paper is well written and supports the publishing of GRQA. One issue for me is that it was not mentioned that WATERBASE is already integrated in GEMSTAT and how the authors dealt with this. There must be a lot of duplicate data from this fact. But the given procedures should have found all the doubled data.

I support the publication of this paper. Some minor issues are listed below.

Specific comments

line 55: please list the parameter used here “most important water quality parameters” or refer to your table 7

line 71: “... only eight parameter matched the data set” – you mean one of the 42 selected parameters? Pls clarify.

line 73 – 74: please give this numbers (how many parameter matched the set, site count range, mean time series length per site, average number of observations per site) for all data sets in a table and refer only to the parameters that were used

line 99: “WATERBASE has the shortest timeseries ... 2008 - 2018” - as far as I remember

there are nutrient data available starting with 1992; please check again the disaggregated data; e.g. see the graphs given under this link

https://www.eea.europa.eu/data-and-maps/daviz/rivers-nutrient-trend-4#tab-dashboard-01

The graph is based on data from WATERBASE.

line 125: "introduction: nutrients, carbon, sediments and oxygen"-Please refer to table 7 here.

line 217: I understood that WATERBASE was included into GEMSTAT; please see https://www.waterandchange.org/en/european-water-quality-monitoring-data-in-gemstat-database-undergoes-major-update/

Also here the time frame is given with 49 years for WATERBASE; please check and write a short paragraph how you dealt with this issue.

Figure 2: Why is the percentage of observations (1898 - 1970) not given her? If no data exist before 1970 it should be mentioned here. The colors in the legend don't correspond to the colors in the plot. E.g., I can only guess what dark green is. Also I wonder what I can learn from this plot. Maybe the authors change this plot and show the time series available per continent instead. That is actually where I'm interested in as global modeler.

line 222-228: I wonder how can the same station be reported to different databases with different position information? In case it is the same station then both stations should contain the same time series? And how are you dealing with stations that are very dense (<1km) but not the same? For example at a tributary that is going parallel and there is one station in the tributary and one in the main stream before tributary is coming in. Did you find such cases and how did you deal with it?

line 296-297: It is not clear to me why the IQR test outliers are removed from the plot? For illustrative purpose – what does it mean? The plots would not look much different wouldn't they? So the outliers are not shown or are removed from the data? And actually for TSS the outliers can be really data e.g. before the crest of a flood or at the beginning of discharge reach a region specific threshold TSS can get extraordinary values.

Figure 4: e.g. TSS – 18.9% outliers removed from the plot – this 18.9% refers to all data? So it could be that 18% outliers are in GEMSTAT and the rest in the other datasets? Do I understand it right? please clarify in the figure text.

line 307 – 313: Please delete this paragraph. The focus of this paper is on merging data together rather than assessing the content. Therefore I would remove lines starting with "DOC concentrations are lower ..." and ending with "point sources". Even if Figure 7 is interesting from the content perspective I would remove it from this paper because it is not the main focus. In my opinion it is sufficient to mention that this kind of plots are available for all 42 parameters in the GRQA dataset.

line 320: What do the authors mean by "using the wrong form"? This is not clear to me. You should have checked the forms before using it in your codes and I'm sure you did. So how can it happen that a wrong form was used? Please explain.

line 322: Why are units in GLORICH in µg/l a problem during conversion process? The authors transformed it to mg/l. I cannot see a problem here. Please clarify the discussion.

line 352: I doubt that ML methods alone can help to fill the gaps in Africa and Asia without

support of any measurements in the regions. So, the argument given in line 359 contradicts the statement given in line 352. Maybe a combination of remote sensing techniques and water quality modelling (including ML) could help. Please revise.

line 380: "The dataset is expected to have yearly updates after publishing" – are you sure that this is realistic? I would rather advise to remove this sentence. It is not necessary to state something like this in a paper, you can just do it if the capacity is available.

line 383 – 384: I like very much the idea to put GRQA in a database and to make it accessible via an online dashboard.

Technical corrections

line 268: table reference is missing

line 269: given numbers of sites per parameter ("15 (POP) up to 90792 (pH)") don't correspondent with the numbers given in Table 7. Please clarify.

line 312: reference is missing