Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-51-RC1, 2021

Comment on essd-2021-51

Anonymous Referee #1

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-RC1, 2021

General comment

This manuscript present workflow and data for harmonizing and joining several major water quality data basis on international and national level to a consistent dataset. The authors carefully describe their workflow and present selected examples on the spatio-temporal coverage of their data sets and suggest further improvements. This is a good example and a showcase what problems scientists have to face when joining datasets. The authors are very careful in making their workflow reproducible for scientists who want to work with the data, e.g. by flagging but not removing outliers. The resulting dataset is unique and very helpful for water quality researchers. I would consider the data as high quality with a few drawbacks mentioned below. While I very much appreciate the work the authors put into the data, I have some concerns regarding the manuscript itself. Overall I found the introduction rather weak. The authors build the justification for their work mainly on water quality modelling with a special focus on machine learning. For me this is to narrow. Any research rely on data, be it the mere statement of a global or regional statistics of a certain constituent, data-driven exploration of controlling factors behind observed patterns or statistical or even mechanistic models. I moreover miss a reference to the need for open data and the FAIR principles. Considering the workflow I greatly miss the handing of values below the limit of detection/ quantification. This is a major issue for all water quality studies and is left entirely open here. I hope the authors can make use of these general comments and the more detailed specific comments below.

Specific comments

Abstract

Line 1-4: I find the introduction rather weak. Some more specific words would be helpful to define the problem, to say what is already there and to what extent this manuscript is going beyond the status quo.

Line 2: It is not immediately clear that "current" study is referring to this manuscript. Consider rewording.

Introduction

Line 13: I disagree here. Why is water quality modelling an integral part of ecosystem health monitoring? Why starting directly with modelling? Models need a conceptual basis that arises from data and interpretation of mechanisms behind the observations. I think there is great value in the data analysis itself before making the step to modelling.

Line 21: What model inputs do you mean here? Please specify. Is it rather calibration data? Or input in terms of drivers (such as precipitation drives a rainfall-runoff model)? Or do you mean parameterization for processes such as a nutrient uptake?

Line 28-33: Here you seem to jump to hydrological models (quantity, not quality) without further mentioning that. To what extent can this be transferred to water quality models?

L34ff: I see the point for putting emphasis on machine learning approaches. However, this seems to be the major justification for large-scale or large-sample datasets and the major outlet of these data. Here I disagree - ML is one possibility but surely not the only justification for the need for water quality data with a wide spatial and temporal coverage.

Line 40: Do you really mean large-scale here? Or rather a wide spatial coverage?

Data

Line 71f: What dataset "that had been previously collected from the other sources" are you referring to? Is that showing up later in the manuscript? Maybe introduce that in the preceding section?

Table 2: Would it make sense to state the date of download here? The databases are still actively fed with new data, right?

Method

I miss a description how the detection limit was handled. This is very crucial. Often numbers such as <0.01 mg/L are given. For some constituents this is rather the rule than the exception.

Line 120: Are outliers just detected (flagged?) or also removed? Why are time series characteristics derived before removing duplicates? Are duplicate station or duplicate samples ment?

Table 4: For me this table does not make sense. Why not stating all conversion information but only an example? The meaning of x1 to x2 should be mentioned in the header.

Line 155: Surely this decision was made wisely. However, it is not clear for the reader why one km is the threshold here. Can you explain that? In a headwater catchment of 2 km2 size a shift of the sampling stations by 1 km would probably not treated as one joint size, right?

Line 206: This is maybe not the best example/ justification. If an agricultural spill has a long-lasting effect on water quality, it would not be an outlier but create more than one elevated values. Consider adjusting the example here.

Line 210ff: What about physical impossible values? Negative concentrations, water temperature above 100°C or concentrations above the solubility of a constituent? Would it make sense to check and remove those as (if they occur) they will influence the percentiles/ quartiles?

Results

Table 6: Attribute parameters such as "upstream basin area" and "Drainage region" have not been mentioned before. However, this is a very crucial information for researchers working with the data. Choosing sites according to their catchment sizes is common with the consequence that this information needs to be reliable. Was that just taken over from the original datasets?

Table 7: Number of digits is relevant here! Is that fixed for all constituents to the same value? If yes, is should be e.g., 0.010 for median phosphorous, not 0.01. This decision should be part of the text.

Line 267: Table reference is missing.

Line 291ff: Usually concentrations are considered to be lognormally distributed. A skewness does thus not purely result from outliers. You should mention that non-normal distribution rule here.

Figure 4 is less informative than figure 3 but basically shows the same. For me this is not necessary. Potentially both figures can be combined.

Line 312: What is the question mark referring to?

Figure 7: Here I realize the point of the physical possible range of parameters. Oxygen can, to the best of my knowledge not be much higher than 15 mg/L - sure there can be oversaturation. But I would claim that 502.75 mg/L is impossible. Would those value not hamper your outlier statistics?

Line 330ff: This is a good idea to come up with suggestions for improving datasets. However, I strongly recommend to make this part of the problem definition in the introduction and the objectives of this study.

Line 352ff: The spatial lack of data is for me not directly connected to the adoption of machine learning. Yes, those techniques may can come up with an estimate. However,

data-driven models are only as good as their training data coverage. For a prediction under boundary conditions/ landscape properties that have not been sampled you will need a mechanistic approach. The whole discussion here is a bit misleading. I don't think the solution for spatial gaps is not gap filling but rather measurements or incentives for countries to share data.