Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-19-RC2, 2021

Comment on essd-2021-19

Thorben Amann (Referee)

Referee comment on "Introducing GloRiSe – a global database on river sediment composition" by Gerrit Müller et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-19-RC2, 2021

The manuscript presents a data compilation of the composition of suspended solids in rivers. The authors have collected a comprehensive set of data, which covers almost all regions of the world. This should enable a plethora of new studies on sediment transport to the oceans, riverine biogeochemical cycling, weathering fluxes, and many more.

The manuscript is timely fills a much needed gap for global studies. I recommend the publication after minor revisions.

I have one general point, which I stumbled upon: The title says the database is about sediments, and then it is actually only about sediments (L71: "Riverbed sediments "), where no data on suspended matter (L68: "suspended sediment") was available. While I understand the differentiation, I find it a bit confusing when I just read the title. I am not concerned about this, I just want to point out, that there is potential for misunderstandings, which may be resolved by a slight change of the used terms. I leave the decision to the authors.

Following the step-by-step guideline for reviewers:

Are the data and methods presented new?

Yes. To my knowledge, there is no other comprehensive data compilation on river suspended matter.

Is there any potential of the data being useful in the future?

Definitely, as stated above.

Are methods and materials described in sufficient detail? Are any references/citations to other data sets or articles missing or inappropriate?

Everything is well described and comprehensible.

Is the article itself appropriate to support the publication of a data set?

Yes

Check the data quality: is the data set accessible via the given identifier?

Yes.

Is the data set complete?

Yes, with the limitations described in the MS itself (Section 4)

Are error estimates and sources of errors given (and discussed in the article)? Are the accuracy, calibration, processing, etc. state of the art? Are common standards used for comparison?

This doesn't really apply here. But shortcomings or problems merging different data sources into one comprehensive database were discussed in the MS.

Is the data set significant – unique, useful, and complete?

Consider article and data set: are there any inconsistencies within these, implausible assertions or data, or noticeable problems which would suggest the data are erroneous (or worse). If possible, apply tests (e.g. statistics). Unusual formats or other circumstances which impede such tests in your discipline may raise suspicion.

Although I didn't experience any problems using the dataset with Python/Pandas, it may be advisable to change the headers to names containing no special characters (like µ, a dash, a smaller than sign…). Programs like ArcGIS, for example, do not like those characters in the header.

Generally, in all files: The use of spaces, underscores or no space between parameter and units is not consistent. Units are also not given consistently (could be derived from documentation, but better give unit in header). I recommend a second screening of the headers to unify the appearance.

Overall, I found some inconsistencies in the data using Python and the package Pandas Profiling (https://github.com/pandas-profiling/pandas-profiling). I will point out some found issues here, but strongly recommend looking into a tool for exploratory data analyses (another recommendation: https://github.com/sfu-db/dataprep) to find flaws in the dataset that relate to format, data types or other formal issues.

Specific points (very selective, there may be more):

SedimentDatabase_ME_Nut.csv

- no Sample_ID/Location_ID/SeaCat/Observation/type/Sample type/Basin_ID/Original_Unit ME/Treatment Method from line 2411 to end, maybe I missed an explanation, but as without identifier, the data is rather useless, isn't it?
- The csv ends with 1 useless (empty columns)
- Column "filter size_>µm" contains 88 values named "cent". Is this correct?

SedimentDatabase_Minerals.csv

- Header: tottal mafic – remove 't'
- The csv ends with 4 useless (empty columns)
- 8 Filter size (> µm)
- 9 Sieve size ((<) µm) -- inconsistent use of parenthesis
- The abbreviations used in the headers should be written out in a table in the

documentation

SedimentDatabase_TE.csv

- Column *Be_ppm* and maybe others contain values with a less-then sign. This may be correct to report like this but makes it hard to process the data with software like Python, Matlab et al, as they will handle the entire column data as string or object. The user then has to manually find and replace the values.

Is the data set itself of high quality?

Check the presentation quality: is the data set usable in its current format and size? Are the formal metadata appropriate? Check the publication: is the length of the article appropriate? Is the overall structure of the article well structured and clear? Is the language consistent and precise? Are mathematical formulae, symbols, abbreviations, and units correctly defined and used? Are figures and tables correct and of high quality?

Everything looks well suited for publication. The Figure 1 quality should be improved. I think the exemplary part (Section 5) is very extensive and goes into a whole lot of detail. I feel this could be shortened a bit (but not left out, just maybe moved to the appendix), to focus on the results that can be achieved. Also, it could be nice to have an overview table with the parameters included in the database, together with the basic statistics like count, mean, median, min, max, percentiles…

Is the data set publication, as submitted, of high quality?

Yes, from my perspective it looks well put together. Just the data itself needs some more cleaning/trimming, as described above.

Finally: By reading the article and downloading the data set, would you be able to understand and (re-)use the data set in the future?

Yes.

Specific comments

89ff Maybe good to reference specific databases like the Glorich or the new GRQA (under review in the same journal: https://doi.org/10.5194/essd-2021-51)

92 "Conversed" ï□ converted

92 Define "properly". What is the reasoning behind the conversion? I guess there is no one doing it not properly on purpose.

103 How were the basins identified?

108 How was this done with Google Earth? Visually?

125 The figure deserves a higher resolution to avoid compression artefacts.

139f What was the reason for the decision to exclude coarser grainsizes? I imagine it is because you assume they don't get suspended, but it would be good to have a short explanation on the criterion.