Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-199-RC2, 2021


Earth System Science Data Discussions

# *Interactive comment on* "MOSAIC (Modern Ocean Sediment Archive and Inventory of Carbon): A (radio)carbon-centric database for seafloor surficial sediments" *by* Tessa Sophia van der Voort et al.

**Anonymous Referee #2**

Received and published: 5 January 2021

In general, I welcome the proposed database and can see its value and utility. However, I do have several points to raise to the authors that should be addressed before publication:

1. The narrative in the Introduction forms a case for support for the need and uniqueness of the database on the one hand, whilst on the other slips into scientific argument of what could/should be done with the data. Both articulations are reasonable, but confuse the reader somewhat. I suggest toning down the suggestions on what can be done with the data. Overall, the paragraphs starting line 82 and line 120 seem largely



redundant. Similarly, i was surprised not to see reference to recent reviews and opinion pieces about sediment carbon (e.g. Snelgrove et al. 2018, TRENDS IN ECOLOGY & EVOLUTION, 33, 96-105; Middelburg, 2018 BIOGEOSCIENCES 15, 413-427) to reinforce what we know and what we don't know. these, and other similar summaries should be incorporated into the text.

2. MOSAIC - minor point, but this acronym is a little unfortunate as it matches the MOSAIC expedition in the Arctic (https://mosaic-expedition.org/ ), a significant programme that will have a long legacy in the literature. I suggest altering the acronym to avoid this overlap, and suggest the authors consider using a title rather than an acronym that incorporates the description of the exactly what is in the database.

3. Line 146 (and then Line 170)- I see the intention of the database, but how often will it be updated and what data quality controls are in place? re line 170, how with the new information gel with the older data, and will efforts be made to back fill the missing data?

4. Paragraph starting Line 164 - A very important aspect of any database that has extracted information from the literature is that the search terms and process of selection criteria needs to be repeatable and absolutely clear. This is of fundamental importance and needs to be explicitly stated in the this section with supporting information in the supplementary material. How were the 200 papers found, selected and checked for data? What search engines and search terms (including any refinements) were used, and how were quality controls implemented? How many papers did the initial search yield, and how was the final subset arrived at? When was the database accessed? Does this database contain data from other databases? What downstream processing of the data, or meta-data, was necessary? e.g. were units converted, how was lat and long derived/converted to the same projection, how was a position assigned to biogeographical zones etc? All steps need to be explained. This is an essential area that needs to be articulated in detail to ensure the authority of the data. The authors need to convince the reader that these data are the ones to use. This is probably the

most important aspects of my commentary that needs addressing fully. Section 2.1.2 needs significant amendments with a focus on attention to detail.

5. Line 177-180 - this is admirable and will be beneficial, but at present does not exist. This aspiration should be omitted from the current description. Instead, the authors should add in the Data Accessibility section that updates will take place (how often? when?) and how to access the latest version of the database. I assume that each iteration will have a documented history and version number thats traceable? If not, this needs to be implemented from the outset.

6. Line 186 - can each individual datapoint be traced back to the individual source (paper)? It will be important that users of the data can look at the context of each datapoint by going back to the original source if necessary. in other words, is there a unique identifier that matches the data value to the specific paper from which it was extracted? This is essential and needs to be included if not already done so.

7. Line 221 - how are submitted data quality checked? make this clear here.

8. Line 228 - how exactly are unexpected values determined? How is this reconciled with unexpected, or outlier variables, that are nevertheless real? Need to reassure the reader that the data is not being sanitised to some pre-determined criteria or parameters.

9. Data quality control - this section needs expanding, as stated earlier, to include quality controls at the point of data collection. The current section only lists quality control post collection. In addition, this section would benefit from some explanation/justification of the detail, supported by citations where necessary/appropriate.

10. Section 2.3.5 - it would be beneficial for the supplementary material to include an "idiots guide" for how to complete a search and extract the data for a simple and more complex query example. For example, what are the step through processes to extract a global dataset versus just one region, or whatever is likely to be a common query.



This should be made readable and accessible to users that have never used SQL or programming, or that have little or no experience of extracting data. The video is a useful addition in this regard, but a manual type addition to the supplementary material would be helpful.

11. Section 3.1 - much of this section is unnecessary and not particularly helpful. the description of the distribution of data is only relevant to the database as it now stands, but as highlighted in the papers, the database will be updated. hence, such statements will be misleading at the point of the first update. Instead, purely descriptive statistics that relate to the database structure (i.e. not interpretative information) should be presented, such as the number of observations for each variable, categorised by region, water depth and other column headings in the database. Presently, it is hard for the reader to understand what the database contains without entering the database itself. As made above (point #1), this section morphs from being a database description to a paper thats interpreting the data. In my opinion, as interesting as the summaries are, the latter has no place here. If the authors wish to interpret the data, they should write a separate contribution and publish elsewhere.

12. Section 3.2 - this can be condensed significantly, many of these points have been made in the Abstract and Introduction. The text would also benefit from reaching out to other fields, perhaps offering other areas that these data may be relevant to that have not received attention previously.

13. Section 3.3 - this section is quite weak and not very compelling. It is not entirely clear whether (i) the data contained in this database is a subset of the other databases mentioned, (ii) how these data differ from other inventories and what the pros and cons of these data are in relation to specific areas of research (maybe include reference to other databases that may form good companions to these data), (iii) and why a user should opt for using these data? Some aspects of these matters are listed, but only in very general terms that lack specifics. Much more explicit arguments need to be made here.

14. Section 4 - add a sentence that states what version of the database this paper is referring to/describing, and how often users can expect updates to the database (e.g. periodically, annually?). I suggest it will also be advantageous to state how errors can be reported.

15. Section 5 - this section is repetitive of the sections above and does not add anything new. This section needs revising to pick up from where the Introduction left off.

16. Table 1- the database contains 8706 entries with latitude and longitude, but only about half of these have a water depth associated with them - could those that do not have a water depth be estimated using, for example, Google earth based on the lat and long co-ordinates? I note the comment re GEBCO, but the same comment made earlier about the state of the database at the point of publication versus aspirations stands.

Overall, I am supportive of the communication, but as the manuscript now stands it does not include sufficient detail about how the data were derived and forms an incompatible mix of existing versus aspirational database properties. I would see both of these as moderate revisions.

---

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-199, 2020.