Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2022-324-RC1, 2022

Comment on essd-2022-324

Lukas Jonkers (Referee)

Referee comment on "BENFEP: a quantitative database of benthic foraminifera from surface sediments of the eastern Pacific" by Paula Diz et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-324-RC1, 2022

Diz et al present a synthesis of benthic foraminifera assemblages from sediment core tops from the Eastern Pacific. Their data product is much larger than what has been compiled for other areas. It hence provides a comprehensive view of “recent” benthic foraminifera biodiversity in a large oceanic basin. The data can be used as a baseline of benthic foraminifera diversity and to aid the interpretation of benthic foraminifera assemblage composition in sedimentary time series.

The paper is well written and logically ordered, the figures are mostly clear and the tables are useful and necessary. My main questions relate to the taxonomic harmonisation, where I think more detail is needed. I have a few additional questions and small remarks that I hope will help to improve the manuscript.

Introduction

The motivation to focus on the East Pacific remains somewhat unclear. The processes and variability that the area is exposed to/characterised by (L31-39) do not seem to be unique to the Eastern Pacific. It would hence be good to be clearer about the scientific motivation to build a database for this specific region.

The authors also mention other syntheses. What is the scope of linking them? I understand that this is no trivial exercise, but I think a few lines somewhere (in the outlook) might be beneficial to the manuscript.

Taxonomy

Taxonomic data are tricky because of the existence of different taxonomic schools and because the nomenclature keeps changing as new species are discovered. The authors have done a great job in standardising data from many different sources. This must have been a long process. Still the data is complex, perfectly illustrated by the fact that the authors recognise 1071 taxa, yet their dataset contains 1502 columns with abundance information. Because of this it would be good to provide a bit more detail on how taxa were mapped onto the WoRMS taxonomy. Was this done using the synonyms provided in WoRMS, using expert knowledge, something else?

Whilst looking at the synonym list and the taxon names in the data file I noticed the occurrence of many variants and subspecies, which are treated as separate categories/taxa. This of course risks inflating biodiversity and renders comparison of different studies challenging as subspecies and variants need not be consistently recognised. It would be good if their meaning and how they are treated would be described in the methods of the harmonisation.

In addition, could part of the reason for percentages summing to >100 be related to the reporting of species and subspecies without removing either the subspecies or the species prior to summing? This is a common issue in planktonic foraminifera datasets.

Some of the taxa appear to be fossil (e.g. https://www.marinespecies.org/aphia.php?p=taxdetails&id=927451), how are these treated? Are they real and therefore indicative of bioturbation/sampling of old material, or do they reflect taxonomic confusion?

The authors have chosen not to preserve the original name and have provided a synonym list in the supplement. The exact changes to each data set are therefore no longer traceable. How do the authors enable/envision future updates to the taxonomy or application of a different taxonomic framework where species lumped here are recognised as separate taxa? And would it be possible to include this information somehow to make the taxonomic framework more flexible and future proof?

One way to address many of these issues is to convert the data to long format. This would allow different taxonomies to exist in parallel (and hence preserve the original). Moreover by having species, subspecies and varieties in separate columns (perhaps even genus to capture sp/spp) the hierarchy of the taxonomy can be preserved in a way that makes it easier to work with because it facilitates easy grouping at different taxonomic levels. I don't claim that this is the only solution, nor that it is absolutely essential to rework the entire database, but it would be good if the authors reflect more on their methodology and describe how they dealt with these issues or how they would recommend users of the database to deal with them.

Limitations

There is very little discussion about what "living" actually means or how it is deduced. I think this could be useful for the non-specialist. Conversely, little attention to how old dead could be. Some more explanation and reference to discussion in the literature would be good in this regard. Paragraph 3.6.3 goes some way, but does not cite much literature.

The same paragraph also mentions that core top samples may not be representative of the most recent sediment if the sediment water interface is not captured. With the majority of the samples collected using gravity corers this could be a real issue, especially when dead and living specimens are not separated. Some discussion is probably warranted. And finally, if living really means living, could the database be used to trace shifts in the assemblages over time given that the database covers a long period of sampling? Would this be a potential use of the data that could be mentioned in the introduction?

Is the preservation potential of all species equal? I imagine that calcareous species are prone to dissolution and that agglutinated species are not always well-preserved. Some discussion/guidance on this matter would be good.

Usage notes

The database is very heterogeneous. There are differences in the sampling instruments, sampling depths, picking method and in the level of detail in the taxonomy (some studies appear to report subspecies and variants, others don't) and the data is sometimes non-

numeric. This is of course not the authors' fault, but it poses some important limitations on/challenges for the use of the data. I therefore miss some usage notes or recommendations on how to deal with this heterogeneity. For instance:

- How to handle non-numeric data;
- How compare samples analysed using different size fractions;
- What is the meaning of empty cells;
- What to do when assemblages don't sum to 100%;
- How to derive meaningful indices of biodiversity given the differences in the level of taxonomic detail;
- etc.

Some of these questions are perhaps trivial to the authors, but not all users of the data will be benthic foraminifera specialists and they might benefit from more detailed user instructions.

Data

The format (xlsx) is proprietary and thus therefore not strictly adhering to the FAIR data standards. I recommend making the data available as some kind of text file that is universally readable and more likely to remain so.

The species names in the headers appear inconsistently formatted and not to map one-to-one to the names in the synonym list in the supplement (use of underscores and brackets). I realise that this is a pain, but it would enhance machine readability if the formatting were made consistent.

Another, very small remark, is the use of special characters in the column headers, which complicates (machine) reading and the authors could consider replacing them with plain text.

Minor comments

L10: consider adding “benthic” at the start of the sentence.

L29: I don’t understand what “figures” means here?

L58: reword, something like “rendering interpretation of the fossil record more meaningful.”

L171-172: I assume that the database can be handled in any software, especially once converted to text format. Why focus on these two examples only instead of highlighting that it can be analysed using any software on any platform?

Fig. 2 legend: the colour scale does not, I presume, show the log10(number of species). I think it shows the number of species as normal numbers, but with a logarithmic colour scaling. I would just remove the (log10).

Fig. 4: perhaps this is a figure where I wonder what the exact purpose is. Do we learn much about the data and why is that useful?

L223: these size ranges seem inconsistent with the figure below, those shown in the figure are wider. Why? And are the “>” signs needed?

Fig 7D: are there really samples with 0 species? And how exactly was this figure made, see my comments above on user instructions.

Fig. 7E: I am not sure in which direction the abundances are summed in this figure (rows, columns or both?). If the authors want to show the depth distribution of selected species then I would expect that each row should sum to 100%, but that seems not the case.

L287: “A few samples…” I counted 423 and 98 >105%, that is not a few. Just provide the numbers.

L331-332: something went wrong with the sentence “It is recommended It would be…”

L333-334: the authors should be bolder here, providing taxonomic information is crucial

for the reusability of (any) taxonomic data. So it is not only of value to the "specialised foraminifera community" and unclear taxonomic information compromises the data and prevents their reusability. Using data only once (in one study) really is a waste of time!

L340: I recommend some kind of versioning scheme so updates can be more easily recognised.