Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2022-324-RC2, 2022

Comment on essd-2022-324

Anonymous Referee #2

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-RC2, 2022

The database focuses on combining and taxonomically standardizing existing benthic foraminiferal abundance data from the North Pacific. These types of data compilations are useful to the community especially when they make available data that were not previously digitized and when they provide taxonomic standardization as done here.

This contribution could better inspire others to use the data by describing the types of studies the data allow, and guiding the reader through what subsets of the data are important for which types of analyses. The authors do note some things, like raw counts are best for studies requiring sample size standardization (rarefaction), but what types of analyses could span all the types of data? The geographic scope of the data is also prescribed and a description of what particular question the geographic scope is uniquely appropriate for answering would put the decision to focus on one region in better context.

My comments below primarily concern the usability of the data and structure of the database because the easier it is to use, the more likely it will be used and contributed to.

Comments on data usability

Supplement 1 indicates which names used by the original authors were synonymized with a given taxonomic name from WORMS, however it does not indicate which authors (or publications) used which names. If future taxonomic revisions are made such that previously synonymized names are assigned differently, there would be no way to deconvolve the data.

If the data were provided as long data, then each entry could have both the synonymized name from WORMS and the name from the original authors. It would also be clear where

subspecies were consolidated if other researchers wish to consider that aspect of diversity. Because the data are sparse (have many zeros), using a long data format would also reduce the file size (not that it is very large as is). Further, adding new data to the database would be substantially easier because it could just be appended to the end of the existing data without having to add or modify columns. Using long data would also allow new data to be uploaded as separate compatible files without modifying the original contribution and allow other workers to more easily use the resource.

The authors note that R could be used to transform the data such that it is usable in ArcGIS, but do not recommend a function nor describe issues the user may encounter when doing so and how to avoid those issues. It would, of course, be better to provide the data in long format and then tell the reader how to turn it into wide data for ecological analyses in R. It would also be helpful to tell readers how to amend the formatting of the provided data table so that is it is readable in R. As presently formatted, a user could not simply read it into R.

The authors note that the database will be continually updated. What will be the mechanism of update? Will the Pangea file be replaced or will multiple versions be available through the same Pangea number? Is there a mechanism that will keep the database location stable? How will different versions be annotated such that version can be tracked? If errors are found and corrected in existing data, how will these changes be documented?

Given the data are provided in different size fractions, some guidance on whether these data should be analyzed together or if they must be analyzed separately would be warranted. How should users handle sources that record only calcareous taxa? Is the scope of the data source clearly noted in the metadata and connected to the main data file such that a user could easily subset data with different characteristics? Information is provided in the remarks columns, but not in a standardized way that would allow easy subsetting.

Guidance on how to handle non-numeric data would also be warranted. What are the recommended process for including or excluding these entries? Would the recommendations be different depending on the situation or scientific questions? I would recommend using different non-numeric symbols for the different meanings. For example, if x does not mean <1% in all cases a user could not treat all x's in the same way and would have to carefully determine which x's mean what. In some cases, the meaning of the x does not appear to be described in the Notes column and/or is not consistently described in the same Remarks column such that automating that process would be difficult. For example, Bandy_Arnal_1957 has x's, but no indication as to what they mean in the Remarks. For Smith_1964 the Remark is in Remarks 2 rather than Remarks 1 like most of the remarks concerning x's.

In the BENFEPfile1 with the main data, sometimes there is a "0" in a cell but often the cell is empty. Is there a different meaning for a "0" as for an empty cell? When converting from wide to long data, entries with a "0" would be retained and need to be removed by

the user unlike the entries where the cell is empty. Formatting all data consistently would be a best practice.

Line 168: I'm not clear what is meant by "ranked abundance of individuals" and the N100, N200, N300 categories based on the text description. Looking at the spreadsheet, it seems this designates whether a sample had at least 100, 200, or 300 individuals in the total sample. Ensuring the language used in the text is the same as what is used in the spreadsheets would reduce any confusion.

Line 159: If the original authors of the data sets do not note whether a specimen is calcareous or not, placing it under "Indeterminate calcareous" may be in error. An "Indeterminate unknown" category would be more accurate. This is also another example of where the language of text does not match the language of the spreadsheet; in the spreadsheet there is a column "Calcareous Indeterminated" but no "Indeterminate calcareous."

The column labeled "total" is confusing as there is not an accompanying column for the units of that total. Units of density may vary and it would be useful to know the actual units rather than just "counts per volume unit."

Line 104: I'm not clear what the authors consider to be quantitative vs. qualitative data. Does quantitative include raw counts, relative abundances, and densities whereas qualitative is just occurrence? Or does qualitative means something else? Perhaps just stating the type of data would alleviate confusion.

The BENEPqual datafile lists localities with metadata and notes the type of data it provides, often noting that the data are available in or are semi-quantitative or are given a species groups. I am not clear what the authors feel a user should do with this information and some description of how this table could be used would be helpful. If these are localities where at least species lists could be obtained, providing the occurrence data rather than just the metadata would make this table of some use for diversity studies.

I notice that Culver and Buzas's Smithsonian Contributions are included in the reference list, however not all the papers they drew from are present in BENEP or BENEPqual. What was the criteria by which sources in those compilations were excluded? Although the authors describe the method by which they search for the data, the criteria by which found sources were rejected is not described and need to be for a user to understand which scientific questions the data are suitable for answering.

If the original authors collected surface sediment samples and simply did not stain the sample to determine what is living, would that go under "living and dead" or "dead"? It

seems that "dead" should only be applied to samples where staining was done and the stained and unstained individuals were counted separately. If the original authors simply did not do staining a surface sample will usually contain both living and dead individuals and I believe it is typical to consider this samples as "living and dead". In the case of Morin 1971, for example, both living and dead data are given for the same number of samples suggesting the samples were examined for both and they were counted independently, but in Niensted, 1986 only "dead" is noted even though the sample is 0-2 cm and presumably would contain living individuals, just none were confirmed living. Some description in the text as to what "living" and what "dead" mean in the text would be warranted and the data attributes made to match how benthic foram workers typically use the terms.

The authors include information on the picking method and it would add clarity if they also described how different picking methods affect the species that are found such that a user may decide how to subset their data. Why is it important that the preparation methods changed over time (Figure 4)? How does that limit what can be done with the data? Does N/A in Figure 4 actually mean "not given" rather than "not applicable"? Similarly, how does the collection device potentially affect the data or is this not relevant to how a user might subset the data (Figure 3)? How does the distribution of living and dead sampling in the database along latitude impact the analyses that can be done (Figure 2)? While these facts about the data are useful, discussion of how to use the facts (beyond it showing where more data should be collected) would greatly help a reader.

Minor notes:

Line 22: Last sentence of abstract seems to be missing something and could use revision for clarity ("studies dearth of quantitative data" … do you mean "with quantitative data" or "without quantitative data")

Line 26: Please cite the source for the Large Maine Ecosystem scheme.

Line 29: What are "protection figures"? Please revise for clarity.

Line 102 says "391 benthic foraminiferal entities (those classified to genus genera level)." But line 236 says "394 benthic foraminifera individuals identified as genera level" the numbers are inconsistent.

Line 202: I'm not clear how a sample taken at 0-2 does not contain the surface of sediment and is thus "deeper sampling." Surely a "surface sample" must also have some width to it.

If the original data were given as a proportion rather than a percent, was this converted to a percent and reported as a percent even if not originally given in that form? If so, information about the precision of the originally reported data may not be preserved.

Figure 5: I don't quite understand the notation in the legend. Would a sample sieved at 150 um be colored red or blue? I assume the (vs] are telling me the answer, but I'm not familiar with the notation.

Figure 6: The color scaling is a bit odd because all they are not equal width. The blue bar would only be 5-10 species whereas all the others are wider bins. What would this plot look like if it didn't only have valid species but also contained the sp. A, sp. B designations for each genus? This plot may show more about taxonomic resolution of the data than anything real about the diversity of genera. Some discussion about how to appropriately interpret this plot would be warranted.

Figure 7: Plotting absences in C is a bit odd because an absence at a locality does not necessarily indicate that the taxon did not live there. Blue dots on a blue ocean are very hard to see and the majority of points are some shade of blue or green in this figure. In D, it seems that "biodiversity" may be heavily affected by sample sizes and taxonomic decisions and probably should just be called "number of taxa present in database" rather than some measure of "diversity."