Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-394-RC1, 2022

Comment on essd-2021-394

Anonymous Referee #1

Referee comment on "A novel specimen-based mid-Paleozoic dataset of antiarch placoderms (the most basal jawed vertebrates)" by Zhaohui Pan et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-394-RC1, 2022

As authors stated, this is a data description article.

It must take authors much time in these specific data collection and sorting. The open-accessed data is really a contribution for the communication of palaeontology and evolution. I believe it will be very useful to the data-driven study of the geosciences and knowledge discovery. In this aspect the manuscript is worthy of a good publication. I highly recommend its publication.

However, there are still several key issues in this work, which, I think, require authors' more and further work. I just give my comments here, as a layman reader of fish fossil study. I suggest authors give more explanations in the text, there are too many scientific terms that are difficult to normal readers. Examples are seen in figure 1, figure 3, and related discussions, and others. Another suggestion of mine is more focusing on the dataset, which is the key point of the whole study. A good description of dataset makes readers understand its potential in analyzing.

The title, 'Dataset of antiarch placoderms (the most basal jawed vertebrates) throughout Middle Paleozoic', why Middle palaeozoic? In the text, we know the geological range of placodems is from the late Silurian to the Late Devonian. why not limit the time range before the fossil group? Authors need to give explanation. In the description of the dataset, the geological background is very important but not well given in the text.

This dataset was extracted from the DeepBone database, or the present dataset is a subset of it. The purpose of this study, as a data description study, is to show the dataset and its potential using, not giving much attention to the analytical result. The explanation of the data elements, its geological background, and data preparation, make up the key contents of the study, which, in this respect, the study should give more information. The using and analyzing data, in this study, are actually only examples.

The valuable feature of the present dataset is its unique and abundant records of Silurian to Devonian Antiarcha. A simple comparison is given in this study (section 3.1). But I think that authors can go further.

Fossil occurrence-based dataset is better for analyzing fossil organism diversity and distribution. A lot of paleobiological study just prefer fossil occurrence data. GBDB is geological section based (Xu et al., 2020, ESSD. the publication year is 2020, but in this study it was written as 2021) and better in stratum correlation, but its data can be exported to fossil occurrences. For the present data analyzing examples, I see that authors are still using the fossil occurrence data (figure 5 and related text). What is the unique merit of the fossil specimen-based dataset? Why the present dataset or DeepBone chose the specimen-based data structure?

Just because this structure, we see that in data spreadsheet, there are many duplicated lines, except the first column showing the specimens identifiers. Every line shows one specimen, we know the common fact that one specie from one locality normally corresponds to several fossil specimens.

Additionally, the elements in the table 1 are not all corresponding to those in the first line in the data spreadsheet.

Line 16 and other, "The dataset consists of 64 genera and 6025 records, covering all antiarch lineages". Why authors do not mention the number of species? Such thing occurs in all the text. Here "6025 records', I guess, means 6025 pieces of fossil specimens. I think such causes confusing because that it needs further definition, especially to define the basic unit (element) of the dataset. In the sections 2.4 and 2.5, figures 3-5, what are the Antiarcha records? Are they individual species, localities? Or specimens? It is only obvious that basic element of the diversity analyses is the fossil taxa (figure 6).

Line 19 and other, "data of Antiarcha", "structured data of…: what does this mean? What data? here also need definition.

Line 21, "including testing hypotheses", actually, using data is not 'testing' something but showing something.

Section 1, authors should emphasize the significant of the present dataset, not only the fossil group. Such two points are closely related but different.

Lines 49-50, “Explaining the spatial and temporal distribution of early vertebrates is the prerequisite to understand their biogeographic exchange”. The normal sequence is, collecting data – analyzing and showing the distribution – recognizing pattern, the last step is probably the explaining you called here.

Line 116 and others, the TrackPoint V 7.0, I only searched this software in the method part of Xu et al., 2020. Palaeogeography, Palaeoclimatology, Palaeoecology. 560. 110029.

Figure 5 needs to be improved, currently it is not clear and hard to get information.

Line 221, “Based on our dataset, the oldest record of”, are you sure that using dataset can conclude the time range result of a fossil? The section 4.1 seems not quite related to the present study. please reconsider it.

Line 256, Eem event, needs explanation.

Section 6, specific and definite conclusion is needed.