Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-212-RC1, 2021

Comment on essd-2021-212

Anonymous Referee #1

Referee comment on "Harmonized chronologies of a global late Quaternary pollen dataset (LegacyAge 1.0)" by Chenzhi Li et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-212-RC1, 2021

This paper describes a pollen records dataset, including explanations and descriptions of the dating methods involved in creating the dataset. The global coverage of this dataset is impressive and the presentation of the manuscript is quite good. There are some minor issues with accessing the data, and some considerable issues with the associated code attached to this paper. While the general shape of the manuscript is good, I encourage a stronger focus on the data itself. These papers are most useful as upfront descriptions of data which requires a slightly different structure than a research articles. Specifically, I would recommend reshaping the intro and the abstract especially to put the data at the forefront, i.e. lead off with statements declaring the dataset, and what it is--for example, putting the name and description of the dataset as the first sentence in both.

The description of dating methods needs to be expanded briefly, including explicitly defining terms such as "reservoir effect" or clarifying what "insufficient carbon" is. Lead dating is lacking description of methodology as is luminescence. Please also include how these dating methods add to measurement uncertainty in the data. Are uncertainties included?

DATA (PANGAEA) - this dataset looks to be in good shape and is well-documented when i look at the site the DOI takes me too. When I download the .tab delimited file though, it is really tough to parse. Is there a reason this is in .tab format? A comma separated (.csv format) would be more universal, but I defer to the authors here if there is some subfield specific reason .tab format is better. Admittedly though, I found it difficult to work with this format when downloaded directly. The html web formatted table was easy enough to read.

CODE
The R code that accompanies this data paper and package is highly problematic from an open-code, data sharing perspective. It is formatted for personal use and not up to

community standards. The main issue is the beginning call of `rm(list=ls())` This command cleans out and removes all entries in a user's memory and R workspace. Jenny Bryan wrote an excellent piece on why this snippet of code does not work for project based workflows (https://www.tidyverse.org/blog/2017/12/workflow-vs-script/)

The major problem with this becomes apparent a couple of lines down when there are 'fixed' calls to data files that do not exist anywhere--nor can I find them. So running the code is impossible.

I would recommend using URLs for those code calls so that when the code is run those data are imported directly from their fixed, online locations. The fixed DOIs from where your data are stored could be used.

This area of this manuscript/data must be addressed. Additionally, the code is commented adequately, and follows a fairly good syntax, formatting structure. I applaud that. The repo in GitHub though has no readme and no documentation there. That really needs to be added. You could include a lot of what is in this paper, in the data metadata write up elsewhere. I would also encourage including a copy of this manuscript as well as copious amounts of links.

A big ask, which I think would take this next level, is to include a vignette or markdown file showing how to work with his data that includes a small, worked example.

In the current state, I cannot run the code, which gives me pause on my recommendation.

Specific comments:

line 44 - the phrase "calibrated and uncalibrated" is confusing.

line 65-75 - it would be advisable to have these variables in a table with further descriptions.

79-80 - repeated use of references to "most common"

Section 2.3.1. - for this type of paper, consider leading this section off with what you have

as your final sentence, then describing it. "...all age relationships in our data set are constructed using Bacon..." then describe why and what and how.

line 139-141 - where did the latest calibration curves come from? this sentence lacks context.

Section 2.3.4 consider laying this section out using bullets or with some kind of work design flow infographic.

* just a note format your units with super- and subscripts, not / notation

lines 167 -... Consider again bullets or something instead of a numbered list inside of a paragraph.