Comment on essd-2021-212
Anonymous Referee #2

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-RC2, 2021

Most analyses using Neotoma or other archived pollen data are dependent, at least to some extent, on the chronologies. The available chronologies have variable quality: some record have an uncalibrated chronology, others have a Bayesian chronology. In many cases the uncertainty on the chronology is not available, or if it is, just the upper and lower credibility interval. To synthesise pollen data from several datasets, it may be necessary to harmonised the age-depth models, a huge amount of work. Once such harmonisation is presented in this current manuscript.

As far as I can tell, the chronologies are not archived, but instead the metadata needed to make the chronologies. This is probably a good idea as it encourages the user to check the parameters.

One important result is that "95.4% of records could be improved ". However, it is unclear what objective criteria were used to determine whether the new model represented an improvement. The criteria need to be explicitly stated.

The metadata and code are available on github (Zenodo.org would be preferable for the final version).

The data are arranged in wide format, with a set of columns for each date. This is not the ideal way to arrange the data, as it makes the code much more complicated to deal with this structure, and will need extra extra columns adding in the future to cope with new sites. A better setup would be to have the data in long format, perhaps in multiple files that can be joined by the dataset ID.

At present, datasets are marked as being marine or otherwise. At least in principle, there could be datasets where some dates are on marine fossils, and others on terrestrial macrofossils. Marine should be a property of the date, not the core.

Setting a constant reservoir age for a core without error seems risky to me. Do you check these are only occurring with bulk dates, as there should not normally be a problem with terrestrial macrofossils.

The code is presented in a single script. This is fine if someone wants to make
chronologies for all datasets, but often people will want to run a subset of the sites, and may find the script difficult to use. Breaking the script into functions than each do one task, perhaps compiled into a small package with help files and other documentation, would make the product more useful.

The code includes fixes for issues the authors found in Neotoma. Please report these issues, and make sure the code to fix them (e.g. line 187) will work safely when the error is fixed.

Figures 4-6. The x-axis, labelled 0-1, 1-2 etc, is rather cluttered and difficult to read. Please consider other ways to label them. The relative lack of outliers in the pre-LGM sediment is interesting, but probably represents over-fitting the models to sparse data.

Figure 7 has a minuscule font: I need to view it at 200% before I can read it. No need to repeat the legend four time - once is enough and leaves more space for everything else.

Minor points:

Line 202 "with fewer than 2" Maybe rewrite as "only one" for clarity. Consideration should be given as to the minimal number of dates that can give a good chronology - I would be cautious using a model based on only two dates.