Response to comments of Anonymous Referee #1

1. General comments

Reviewer comment: 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.

Reviewer comment: *(1) 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.*

Response: Thank you for your suggestion. We have revised the abstract section according to your suggestion.

New text (line 15-17): 'We present a chronology framework named LegacyAge 1.0 containing harmonized chronologies for 2831 pollen records (downloaded from the Neotoma Paleoecology Database and the supplementary Asian datasets) together with their age control points and metadata in machine-readable data formats'.

The introduction is now in line with other ESSD papers on similar topics (Sánchez Goñi et al., 2017; Cao et al., 2020), i.e., 1. introducing the potential of available pollen databases, 2. stating the research gap why pollen databases cannot be fully exploited (i.e., missing harmonized chronologies), 3. summarising this study.

Reviewer comment: *(2) 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?*

Response: We have expanded the description of the dating methods, and how these dating methods add to measurement uncertainty in the data, please see section 2.2.

New text (line 187-190): 'Reservoir effects: the uptake of old carbon by aquatic plants, mosses, or shells either originating from, e.g., limestone in the catchment ('hard-water effect') or slow ^{14}C

exchange between the atmosphere and ocean interior, can result in too old radiocarbon dates (Philippsen, 2013; Philippsen and Heinemeier, 2013; Giesecke et al., 2014; Heaton et al., 2020)'.

New text (line 84-96):

'*Radiocarbon dating:* most records were dated using radiocarbon-based methods (^{14}C dating, conventional or accelerator mass spectrometry, Christie, 2018), coverimg the time range of ca. the last 50 kyr BP (before present, where 'present' is 1950 CE). However, the accuracy and precision of the radiocarbon dates depend on the calibration curve, taphonomy, and dating materials (Blois et al., 2011; Heaton et al., 2021).

***Lead-210 dating:* the uppermost part of some lake records has been dated using a radioactive isotope of lead (lead-210), which has a half-life of ca. 22 years and provides useful age control for the last 75-150 years. However, the abundance of other radioactive isotopes (e.g., Caesium-137) affects the accuracy and precision of the calibration curve for lead-210, resulting in temporal uncertainty (Appleby and Oldfield, 1978; Cuney, 2021).**

***Luminescence dating:* archaeological materials, loess, and river sediments have often been dated via luminescence, including thermoluminescence (TL) and optically stimulated luminescence (OSL), which cover time scales from millennia to hundreds of thousands of years (Roberts, 2013). Due to the systematic and random errors in the measurement process, the luminescence ages have at least 4-5% uncertainty, which widens with increasing time (Wallinga and Cunningham, 2015)'.**

2. Data (PANGAEA)

Reviewer comment: *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.*

Response: Thanks for your suggestion. All datasets can be downloaded in .csv and .tab format. We improved the documentation on how to access the dataset in 'Code and data availability' and in the readme text.

New text (line 313-320): 'Seven supplementary datasets (Table S1-S7, in comma-separated values format) and one readme text about the LegacyAge 1.0 are accessible in the navigation bar

‘Further details’ of the PANGAEA page (https://doi.pangaea.de/10.1594/PANGAEA.933132; Li et al., 2021a). We provided the chronological control points metadata (Table S1), prior information of dates from publication (Table S2), Bacon parameter settings (Table S3), original chronology metadata from the Neotoma and Cao et al. (2013, 2020) (Table S4), LegacyAge 1.0 chronology (Table S5), description of the comparison of original chronology and LegacyAge 1.0 (Table S6), and record references (Table S7) respectively. All datasets are already in long data format that can be joined by the dataset ID’.

Readme text: ‘Please select 'Further details' on the left navigation bar of the webpage to access the dataset in .csv format’. The 'Download Data' on the bottom navigation bar of the page can also download the dataset as tab-delimited text and view the dataset as HTML (shows only the first 2000 rows), stored in the PANGAEA. However, PANGAEA may rename the variables of the uploaded dataset to match its database format. These new variable names may have special characters that do not match the requirements of R, so we highly recommended downloading the original file (in .csv) we uploaded before running the R code’.

3. Code

Reviewer comment: *(1) 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/)*

Response: We revised the code according to your suggestion, i.e., we removed rm(list=ls()) memory clean. While coding, we were unaware that this could be a problem, so thanks for your input and the link to Jenny Bryans’ work. Furthermore, we now store metadata, code, and results on Zenodo (https://doi.org/10.5281/zenodo.5815192), which better supports open-coded and data sharing.

Reviewer comment: *(2) 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.*

Response: Thanks for your suggestion. We revised the code and reduced the input files to three tables, defined in the first 51 rows of code. We also used URLs for the calls to import these three input files directly from PANGAEA (Supplement Table S1, S3, and S4; https://doi.pangaea.de/10.1594/PANGAEA.933132).

New code (line 38-51):

‘metadata <- read.csv2("https://download.pangaea.de/reference/111158/attachments/Table-S1_chronological_control_points_metadata.csv", stringsAsFactors = FALSE, sep = "\t", dec = ".")

parameter <- read.csv2("https://download.pangaea.de/reference/111160/attachments/Table-S3_bacon_parameter_settings.csv", stringsAsFactors = FALSE, sep = "\t", dec = ".")

AgeDepthPollen <- read.csv2("https://download.pangaea.de/reference/111161/attachments/Table-S4_original_chronology_metadata_by_pollen_records.csv",stringsAsFactors = FALSE, sep = "\t", dec = ".")’

Reviewer comment: *(3) 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.

Response: We apologize for this. We revised the code and rephrased this part of the text.

Description of the uploaded file in new text (line 321-324):

‘The R-code for calculation and comparison of chronologies with embedded manual, metadata for code runs, Bacon output graphs of each record, graphs comparison of original chronologies and LegacyAge 1.0, and a short shared-screen video of the R-code to show the usage on two example records are accessible on Zenodo (https://doi.org/10.5281/zenodo.5815192; Li et al., 2021b)’.

Description of code usage in new text (line 331-333):

‘As for the R-code, users only need to set the working directory where the Bacon results will be stored and input the record ID of interest to run it successfully. The manual and shared-screen video on R-code usage could provide helpful guidance for users, with or without some R-experience’.

4. Specific comments

Reviewer comment: *(1) line 44 - the phrase "calibrated and uncalibrated" is confusing.*

Response: We rephrased this part of the text. This particular part of the sentence was deleted.

New text (line 46-48): ‘Furthermore, the chronologies have been established using a variety of methodologies, and the quantification of temporal uncertainty, particularly between records, remains a challenge (Blois et al., 2011; Giesecke et al., 2014; Flantua et al., 2016; Trachsel and Telford, 2017)’.

Reviewer comment: *(2) line 65-75 - it would be advisable to have these variables in a table with further descriptions.*

Response: Now, all metadata variables are listed in Supplement Tables S1 and S4 at PANGAEA (https://doi.pangaea.de/10.1594/PANGAEA.933132), the related variable descriptions are listed in the readme text.

Reviewer comment: *(3) line 79-80 - repeated use of references to "most common"*

Response: We supplied references where we used most “most common” before.

New text (line 84-87): ‘*Radiocarbon dating:* most records were dated using radiocarbon-based methods (^{14}C dating, conventional or accelerator mass spectrometry, Christie, 2018), coverimg the time range of ca. the last 50 kyr BP (before present, where ‘present’ is 1950 CE). However, the accuracy and precision of the radiocarbon dates depend on the calibration curve, taphonomy, and dating materials (Blois et al., 2011; Heaton et al., 2021)’.

Reviewer comment: *(4) 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.*

Response: We followed your suggestion.

New text (line 121-128):

'We used the Bacon software (Blaauw and Christen, 2011) to establish continuous down-core chronologies from the age control points. Bacon fits a monotonic autoregressive (AR1) model to age control points using Bayesian methods to combine information from the control points with prior information on the statistical properties of accumulation histories for deposits, e.g., a prior distribution for the mean accumulation rate and how it varies (Blaauw and Christen, 2011). Several other approaches are available for age-depth modeling, including linear interpolation, smoothing splines, and other Bayesian methods, e.g., OxCal (Ramsey, 2008) and Bchron (Haslett and Parnell, 2008). However, Bacon has become one of the most frequently used and compares well with other methods (Trachsel and Telford, 2017, Blaauw et al., 2018)'.

Reviewer comment: *(5) line 139-141 - where did the latest calibration curves come from? this sentence lacks context.*

Response: The latest calibration curves (IntCal20, SHcal20, Marine20; http://calib.org/), are already included in Bacon.

New text (line 148-152):

'To transform the measured ^{14}C ages to calendar ages, the latest calibration curves, approved by the radiocarbon community (Hajdas, 2014), were used in Bacon routine: IntCal20 (Reimer et al., 2020; Heaton et al., 2021) and SHcal20 (Hogg et al., 2020) to calibrate the terrestrial radiocarbon dates in the northern and southern hemispheres, respectively; and Marine20 (Heaton et al., 2020) for the 38 marine records included in our dataset (Sánchez Goñi et al., 2017)'.

Reviewer comment: *(6) Section 2.3.4 consider laying this section out using bullets or with some kind of work design flow infographic.*

Response: We laid this section out using bullets following your suggestion.

New text (line 161-183):

'(1) The prior for the accumulation rate consists of a gamma distribution with two parameters, mean accumulation rate (acc.mean; default 20 yr cm^{-1}) and accumulation shape (acc.shape; default 1.5). For the acc.shape, we accepted its default value as higher values resulted in a more peaked shape of the gamma distribution. A first approximation of the acc.mean was calculated as the average accumulation rate between the first and the last date of each record, combined with the prior information of dates, which is more reasonable than using a constant value.

(2) Bacon divides a core into many vertical sections of equal thickness (thick; default 5 cm), which significantly affects the flexibility of the age-depth model, and through millions of Markov Chain Monte Carlo iterations estimates the accumulation rate for each section. Blaauw and Christen (2011) indicated that models with few sections tend to show more abrupt changes in accumulation rate, while models with many sections usually appear smoother but are computationally more intense. We run Bacon for six section thicknesses (2.5 cm, 5 cm, 10 cm, 30 sections, 60 sections, and 120 sections), optimal values after numerous tests, with and without core-top age resulting in 12 initial chronologies for each record.

……'

Reviewer comment: *(7) * just a note format your units with super- and subscripts, not / notation*

Response: We changed '/ ' to superscript ($^{-1}$).

New text (line 161-162): 'The prior for the accumulation rate consists of a gamma distribution with two parameters, mean accumulation rate (acc.mean; default 20 yr cm^{-1}) and accumulation shape (acc.shape; default 1.5)'.

Reviewer comment: *(8) lines 167 -Consider again bullets or something instead of a numbered list inside of a paragraph.*

Response: We laid this section out using bullets following your suggestion, same as before.

New text (line 186-207):

'(1) Reservoir effects: the uptake of old carbon by aquatic plants, mosses, or shells either originating from, e.g., limestone in the catchment ('hard-water effect') or slow ^{14}C exchange between the atmosphere and ocean interior, can result in too old radiocarbon dates (Philippsen, 2013; Philippsen and Heinemeier, 2013; Giesecke et al., 2014; Heaton et al., 2020). In addition to the reservoir ages reported by the original authors, we also identified some additional records for which there is likely a reservoir effect through modern correction and linear extrapolation (Wang et al., 2017). We then subtracted the reservoir age as a constant from all ^{14}C dates of an affected record, excluding those derived from terrestrial macrofossils. We may have underestimated the number of such records due to the difficulty of estimating the reservoir age where the sediment surface was eroded or used for agricultural purposes.

(2) Waterline issues: stratigraphic records do not always start at a depth of 0 cm, for example, if the uppermost part of the core is lost, if the record is only a part of a longer sequence, or if the depths are measured from the water surface instead of the sediment surface, leading to the so-

called waterline issue. Accordingly, we adjusted the uppermost depth of the chronology based on information collected from the original publications and Neotoma.

……’