Response to comments of Anonymous Referee #3

1. General comments

In this work the authors present their attempt to harmonize mainly radiocarbon-based chronologies of continental climate records. The harmonisation is with respect of age-model software usage, calibration curve usage, which is a very valuable task. Furthermore, harmonisation is performed with respect to parameters used for the age-depth modelling software. As far as I understand, the authors use the age-modelling software Bacon for age-depth modelling of a huge quantity of records. Before modelling, the cores were manually evaluated in terms of complications, such as radiocarbon reservoir effects, water lines, etc.

While I appreciate their approach, I think there are some things to be improved before suggesting this piece of work for publication.

2. Data (PANGAEA)

(1) Furthermore, I am not able to find age-depth profiles on their provided Pangaea-page. I thought the authors did all their work (handling reservoir effects, water lines, deciding for the best thicknesses to be applied, ...) in order to provide a homogeneous age-depth data set. And according to their paper, they spend a lot of efforts to evaluate the datings etc of all records. It would be a pity, if they would not share this. Or is the user supposed to start from scratch again? Even if it 'only' means to run their script – if I understand the code structure correctly, the user has to run all of their thousands of records, even if the user is only interested in one or two records. Especially, as this means to run 'millions of MCMC iterations' (line 120) which cannot be that cheap as even admitted by the authors: "... it needs much supervision and computing power" (line 122). Why not provide all age depth models (including uncertainties) in addition to all meta data and code? Or at least enable the user to only calculate the age-depth models of the records they are interested in?

Response: 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). We provided the chronological control points metadata (Table S1), prior information of dates from literature (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. Furthermore, the R-code for calculation and comparison 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.5793936). Note that we moved to store them from GitHub to Zenodo. Zenodo provides a persistent DOI to make the work easier to cite, supporting the data from Github repositories, as supported by referee 2. Thus, readers can obtain chronologies directly using Table S5 we provided or use the script to calculate several or all of the records they are interested in. We also encourage readers to check the parameter settings.

(2) Another critical question is about the final age models. As I cannot find them, nor are able to run the R script, I have to ask: Which depths intervals do you choose to save for the homogenised age-depth models? In the paper you mention the effect of choosing different levels or depth intervals on the goodness of the model data and that some are better suited than others. However, I even wonder, why a user should care about having the age-depth relationship on a fixed sampling interval? If I want to work with other paleoclimate data, I am interested in an age-depth model, which provides dates at depth, where the proxies were measured. Is the output of your script arranged in a way, that this could be easily accessed? Unfortunately, this is not mentioned in the paper. Or do you expect the user to apply some (more or less) fancy interpolation algorithm to assign ages for the proxy depths?

Response: Two Bacon parameters need to be clarified, 'thick' and 'd.by'. Bacon will divide the core into many vertical sections (by default of thick=5 cm thickness) during calculation, which significantly affects the flexibility of the age-depth model. Since our dataset contained 2831 records, it was unrealistic to establish the age-depth relationship for each record using different section thicknesses. To batch process, we finally selected six thicknesses tested many times for most chronologies (ca. 85%). We have also adjusted the section thickness for some records, please refer to supplement Table S3. Of course, the reader can check or modify the parameter settings to generate a higher quality chronology. Another parameter, 'd.by,' i.e., depth intervals at which ages are calculated. Table S5 includes all records' the calibrated ages (mean, median, minimum, maximum) at each centimeter. Readers can assign ages for the proxy depths in two ways: applying the interpolation algorithm in the results we provided, or modifying the parameter 'd.by' (default=1 cm) to recalculate it. In summary, the six section thicknesses (2.5 cm, 5 cm, 10 cm, 30 sections, 60 sections, and 120 sections) mentioned in the manuscript affect the flexibility of the age-depth model, which is different from the depth interval of the chronology.

3. Code

Usually, such a data set and code is generated to be used. Unfortunately, I cannot find any description or manual, how to access the age-depth models. Nor is it possible for me to run the R-script. I admit, I am a R-noob, but I think, application should be properly described with at least a short manual for users with some R-experience (or even noobs). This does not have to come with this publication, but it should at least appear on their github space next to the R-file.

Response: We apologize again for this. We reorganized the code and reduced the input files to three tables (Supplement Table S1, S3, and S4) defined in the first 51 rows of code together with an embedded manual. The reader can calculate chronology for all records or some records of interest by modifying lines 35-36. We used URLs for those code calls so that when the code is run, those three input files are imported directly from PANGAEA (https://doi.pangaea.de/10.1594/PANGAEA.933132). Also, all readers can download these files from PANGAEA or Zenodo (https://doi.org/10.5281/zenodo.5793936) to a new folder and insert the path of the folder to the folder definition at the begin of the code. Additional to the embedded manual, we provided a short shared-screen video in Zenodo to show the usage on two example sites. The embedded manual and the screen video should be helpful as readme/documentation, and now it should be possible to run the code easily.

4. Figure

Fig. 7: Please provide information about which of the twelve generated age-depth models for each record you show here! Would it be possible to show one additional age-depth realisation, which fits less good with the measured ages. Only to give the reader an idea about the effects of the choice of depths intervals.

Response: Thanks for your suggestion. As you read from revised Figure 7, there are two section thicknesses in the title of each figure, 'best' and 'poor' separately. Also, one additional age-depth realization with relatively poor performance has been included for comparison. Preference was given to models that fitted the dates well, had small mean uncertainties, and good runs of Markov Chain Monte Carlo iterations (i.e., a stationary distribution with little structure among neighboring iterations as indicated by the traceplot of the joint likelihood) when choosing the 'best' model for each record.

5. Specific comments

(1) L16 and 46: Please elaborate a bit more on what you understand by 'harmonized chronology' already this early in the manuscript. I am pretty, sure, that different people understand different things under this term. I mean later in the paper it becomes clear, what you understand by this term, but I think it is worth to highlight this already in the beginning of your work.

Response: As you understand, 'harmonized chronology,' i.e., using the same strategy for consistent inference of age and age uncertainty. We also elaborated this term a bit more in the introduction section, see line 48 of new text.

(2) L27-28: This sentence needs more explanations. Maybe not here in the abstract, but below in the according text passages. Please find a more detailed comment below.

Response: Yes, only the final result of the comparison is only shown here. The criteria for the preferred models are that the model fitted the dates well, had small uncertainties, combined dates with prior information (e.g., geological and hydrological setting, environmental history), and calibrated with the latest calibration curves, see line 222-225.

(3) L69-74: You provide quite some detailed information on metadata, which I appreciate a lot. However, I doubt that putting those data all in one file is the best option. I agree with referee 2 to splitting this file up in several is maybe more appropriate and easier to handle. At least keep this in mind for any potential future improvements.

Response: We reorganized the metadata into three supplement tables: the chronological control points metadata (Table S1), Bacon parameter settings (Table S3), and the original chronology metadata from the Neotoma and Cao et al. (2013, 2020) (Table S4). Also, readers can learn more information about the variables in the table from the readme text.

(4) L155: 'acc.mean' is possibly 'acc.rate'?

Response: The correct abbreviation for mean accumulation rate is 'acc.mean', see line 164. We have made the change in the text and apologize for the confusion.

(5) L158: 'We tested six thicknesses (2.5 cm, 5 cm, 10 cm, 30 sections, 60 sections, and 120 sections) …'. I am not very familiar with Bacon. But, why would you want to test those 6 sampling intervals? I mean, the proxies of the cores were measured at specific depths - wouldn't it be more suitable to only interpolate to those depths, where proxy data exist? Actually, this is the data, I would be interested in. But it seems, that this is missing completely. What do you suggest to finally obtain the ages at those depths?

Response: As mentioned earlier, these six section thicknesses (2.5 cm, 5 cm, 10 cm, 30 sections, 60 sections, and 120 sections) affect the flexibility of the age-depth model, which is different from the depth interval of the chronology. All readers can assign ages for the proxy depths in two ways: applying the interpolation algorithm in the results we provided, or modifying the parameter 'd.by' to recalculate it. For example, we assigned ages to pollen samples by interpolation (https://doi.pangaea.de/10.1594/PANGAEA.929773). As you can read from the supplement Table S5, we provided the estimated age at each centimeter, which provides the possibility for other proxies (not only pollen) to interpolate ages at a specific depth.

(6) L159: 'artificial surface age', Why would it be necessary to add an artificial date? I don't know if I understand the concept of adding an artificial date correctly. Stating things like this sounds very arbitrary. Or do you mean you added another age-constraint due to the assumption that the core sedimentation was active until core recovery? And that the additional age constraint is the year of core recovery? If yes, please consider to specify accordingly.

Response: Yes, you are right. If the core was collected from sites where sediment was still accumulating, the core-top age could be one significant time control for the chronologies. Therefore, an estimated artificial surface age (-50 + -30 cal yr BP) was used if the core-top age cannot be obtained from the sampling date in literature or original chronology in Neotoma. We have also made changes in the text, see line 137-147.

(7) L159: 'generating 12 age models for each core'. Just to make sure I understand correctly. Your code provides 12 age-depth models for one core. Are all provided in output files?

Response: Yes, our code initially outputs 12 age-depth models for each record. We only provided the 'best' chronology for each record to PANGAEA (https://doi.pangaea.de/10.1594/PANGAEA.933132; Supplement Table S3), also the Bacon output graphs of each record in Zenodo (https://doi.org/10.5281/zenodo.5793936). You will get the best chronology for each record if you run the script directly. Meanwhile, if you want to get multiple age-depth models for each record, you can do so by modifying the column '*Resolution.cm*' or '*Resolution.section*' of Table S3. The embedded manual and the screen video should be helpful as readme/documentation, and now it should be possible to run the code easily.

(8) L170: I think, C exchange between dissolved C-species in water and atmospheric CO2 is not responsible for 'too old radiocarbon dates'. Instead, this process counter balances to some degree the effect of the arguments listed earlier in this sentence.

Response: We agree with you. However, slow ^{14}C exchange between the atmosphere and ocean interior, can result in too old radiocarbon dates, which we originally wanted to express.

Radiocarbon dates of a terrestrial and marine organism of equivalent age have a difference of about 400 radiocarbon years, i.e., marine radiocarbon reservoir effect, see line 185-188.

(9) L171-173: For some records you added your evaluation of reservoir effects. I appreciate this a lot, but I think it is worth to add a column in your metadata file and mark those records. This would allow a better transparency about what is your evaluation and which information came from the original studies.

Response: Readers can view this information in column *'Reservoir'* of supplement table S3, or view type 2 in column *'Category'* of supplement table S2 to learn how this information was obtained.

(10) L184: For the use of radiocarbon dates for modelling purposes, you followed 'in most cases the suggestions in the original publications'. Please consider – again for a better transparency - to provide information (maybe in your metadata file), for which records you did not follow the suggestions of the original publications.

Response: We rejected or added dates based on prior information collected from the original publications and Neotoma. As you can read from the last two columns of supplement Table S2, all kinds of prior information are listed here.

(11) L189-191: 'For each record, 12 age models were visually assessed. Preference was given to models that fitted the dates well and with small uncertainties when choosing the 'best' model for each record (Blaauw and Christen, 2011; Blaauw et al., 2018).'. This is a lot of work for thousands of records. You are sure, that you did this all correctly for this large amount of records? I wonder if it would have been more objective to apply a short statistical test on this. I mean, most likely a simple least square test between age model and ages of dated depths would do a better and faster job. Also the 'small uncertainty' argument would be most likely more precise and faster to obtain, when calculating the mean uncertainty instead relying on visual assessment.

Response: Yes, you are right. We visually evaluated the 12 initial age-depth models for each record following the Bacon manual's common method, which took a lot of time. Preference was given to models that fitted the dates well, had small mean uncertainties, and good runs of Markov Chain Monte Carlo iterations (i.e., a stationary distribution with little structure among neighboring iterations as indicated by the traceplot of the joint likelihood) when choosing the 'best' model for each record. This work was visually evaluated by two individuals independently according to a unified standard, and then the results of both were combined to reduce the error. We also tried least-squares initially, but it is dangerous to choose the best model only based on its results. A significant disadvantage of the least-squares method is that it is greatly affected by the

disturbance of outliers. If we only choose a model with the least-squares, this model may have significant uncertainties due to the overfitting dates. Finally, we decided to use the visual method because Bacon's output of the graph is apparent. Specifically, we can quickly check the result of Markov Chain Monte Carlo iterations, the overall picture of uncertainty, and how well the model fits the date. If the model can fit the date well, it is actually also an application of the least-squares idea, but we judge subjectively rather than statistically. We also calculated the mean uncertainty of each model for each record. The reader can view the mean uncertainty (95% confidence ranges) of the 'best' model and uncertainty at each centimeter of each record at supplement Table S5.

(12) L203: Who did the evaluation about what a reliable date is? You or the original authors? I can imagine, that this is a difficult task, especially for cores from others.

Response: We assessed all dates based on prior information, as authors usually report all ^{14}C dates from a sequence, even if some are deemed inaccurate. We also fully respect the original authors' comments because we are no more familiar with the sites than they are.

(13) L247-248: 'where original chronologies outperformed LegacyAge 1.0, …' How do you know, which model approach outperforms the other? How can you measure or evaluate this? Do you have knowledge of the 'true sedimentation history' of all those records to be able to judge this? Which one do you choose from your 12 ones/core? I think it is very crucial to provide more details on this issue. Or, in case you wanted to express a different thing with this expression, please consider to rephrase this sentence.

Response: The newly generated 'best' calibrated chronology of each record were compared with original chronologies taken from the Neotoma and Cao et al. (2013, 2020) datasets (Supplement Table S4) to evaluate the performance of the new models. The criteria for the preferred models are that the model fitted the dates well, had small uncertainties, combined dates with prior information (e.g., geological and hydrological setting, environmental history), and calibrated with the latest calibration curves. We have added it to section 2.4 in the text (line 222-225).