Responses to Timothy J. Heaton

Dear Timothy J. Heaton,

First of all, we would like to thank you very much for taking the time to review our manuscript entitled "Improving age-depth correlations by using the LANDO model ensemble" in detail. We appreciate your comments on how to improve the current version of our manuscript and your advice to use LANDO for purposes other than lake sediments.

In the attached .pdf response file (supplement), we provide detailed replies to each individual comment and provide our proposed changes and adjustments to the current manuscript that we will carry out and show within the revised manuscript version. We also point out potential future work beyond the scope of the current version. We have therefore highlighted your comments in black and *italics* and highlighted our responses in blue.

Thank you once again for taking the time to review our study.

On behalf of all the authors,

Gregor Pfalz

---

2 Major Conceptual Comments:

2.1 Suitable Types of Data

Implicitly it seems as though when entering ^{14}C determinations, the data are calibrated against the IntCal20 atmospheric calibration curve - possibly with the application of a reservoir age (although more on that a bit later). Calibration against this IntCal curve is only appropriate for NH atmospheric samples, or for lakes where the reservoir offset is independent of ocean circulation (in such cases the surface water depletion occurs as a potential consequence of both the release of old, but not necessarily dead, organic carbon from soils and peats; and dead inorganic carbon, a hard water effect, entering the lake from its inflows/groundwater). For data from open oceans, one must use the Marine20 calibration curve - it is not appropriate to apply a constant reservoir age to ^{14}C samples from the open oceans since the open ocean environment considerably smooths/filters the (radiocarbon vs calendar age) variations seen in atmospheric signal. Such smoothing does not occur with the application of a constant reservoir age.

However, there are many applications beyond simply lake sediments where one might wish to use age-depth modelling. Age-depth models are frequently used in ocean sediment cores and in archaeological sites. This broadens the potential scope of LANDO. While the introduction discusses only lake sediments, it is not explained as to when the internal calibration process is appropriate and when it is not. Some explanation is needed here as it is likely users will come across LANDO in other contexts beyond simply lake sediments. Further, permittting the users to select the marine calibration curve (with an appropriate ΔR) would increase the applicability of the tool.

I also note that potentially some of the cores, shown in Figure 1 and used in the third case study, look like they might be more general open ocean cores than solely lake sediment. Is this the case? In which case they really should be calibrated using the Marine20 curve - this may also have an effect on the sedimentation rate estimates around the Holocene-Pleistocene boundary since the (open ocean) marine reservoir age is known to change between these two periods - see Figure 4 and 7 in the Marine20 paper.

We agree that for open ocean core data you need to calibrate the data with the Marine20 calibration curve. However, all sediment cores included in this study are lake sediment cores. We do not have a single marine sediment core. However, it is already possible to calibrate dates in LANDO using the Marine20 calibration curve. As we state on Page 11, Lines 326-328: "*To include age determination data within the plots, LANDO internally calibrates the radiocarbon data with the "BchronCalibrate" function of the Bchron package (Haslett and Parnell, 2008; Parnell et al., 2008) with either the IntCal20 (Reimer et al.,* *2020) or Marine20 (Heaton et al., 2020) calibration curve.*" If the user enters "14C marine fossil" as "Category" for the input data (please see Table 1, page 4, "material_category"), LANDO automatically applies the Marine20 curve to this age point and uses the "reservoir age" as ΔR.

So far, we have only used the IntCal20 curve since all our sediment cores are in the northern hemisphere. If users intend to use LANDO for purposes other than lake sediment cores, we would need to make it clearer for the user which calibration curves are applicable. However, this is beyond the scope of this article, which focuses on Arctic lake sediment cores. We may address this in a future update to make LANDO accessible to other areas.

2.2 Reservoir Ages

The way that reservoir age is applied in LANDO seems to use a different definition of reservoir age to that commonly in use within the ^{14}C community. For standard IntCal/MarineCal radiocarbon calibration, the reservoir age (at calendar age θ cal yr BP) is defined as the difference between the radiocarbon age of dissolved inorganic carbon (DIC) in the mixed surface layer of the water at that location, and the radiocarbon age of CO_2 in the Northern Hemispheric (NH) atmosphere. In other words, for IntCal and calibration, the reservoir age is measured in ^{14}C yrs and is applied to the ^{14}C determination before calibration.

In this paper however, it seems that the LANDO reservoir age is defined as the difference between the calendar age obtained by calibrating directly against the IntCal curve without any adjustment, and the true calendar age. In the preparation step you use the difference in calendar ages between the unadjusted model and the top of the core.

Applying a constant offset in the ^{14}C domain before calibration of a sample is equivalent to an assumption that a constant proportion of the ^{14}C in that sample arises from inorganic carbon (e.g., the hard water effect). This is not quite true if you simply shift the calendar ages. For old sediment cores (from the pleistocene) or sparsely sampled cores the difference between the approaches may be relatively small - especially for cores that are incredibly long. However I believe this will cause confusion to users.

Thank you for your comment. We assume that there is a slight misunderstanding. Regarding the procedure: First, we assume that there is no reservoir age. We use the hamstr modeling system, which internally calibrates the radiocarbon dates with either the IntCal20 (or the Marine20) calibration curve using the "BchronCalibrate" function of the Bchron package. Then we use hamstr to determine the uppermost layer (0 centimeter of the depth within sediment core = depth below sediment surface) based on the calibrated dates. This means that we build an age-depth model with no reservoir age as input. If we detect a difference between predicted model output for the uppermost layer (e.g., 1200 cal. years BP) and the actual age we know from the expedition (e.g., -68 cal. years BP – for an expedition in 2018), then we assume that the difference between the two values is the reservoir age (e.g., 1268 cal. years BP). LANDO then adds the new reservoir age to the input file, which means that for the actual model runs, all modeling systems within LANDO apply the reservoir age to the ^{14}C determination before calibration.

We would also like to mention that the reservoir correction is an additional option in LANDO. As we states on Page 5, Lines 133-134 "***In the absence of a known reservoir age or recent surface sample**, we used available radiocarbon data points and a fast-calculating modeling system to predict the age of the upper most layer within a sediment core.*" If users know the reservoir age and still use the reservoir correction option, then this would either not produce a suggestion from LANDO (as desired and predicted output match) or LANDO would suggest a higher reservoir correction that user can either accept or ignore.

2.3 Application on Inconsistent Cores: Example 2

It is my very strong belief that no one should be trying to fit an automated age-depth model to the data as shown in Figure 3. It is clear there is something highly unusual and unexplained regarding the measurements in this core. I am not sure if this is due to certain techniques disagreeing e.g. OSL being different from ^{14}C (since the data are plotted in the same colour and I cannot tell which dates are which). The correct approach would be to go back to the measurements and determine what is going on. In my view, suggesting that data as inconsistent as this can be resolved by forcing them through a range of models (which may or may not happen to select all path through the data) is highly dangerous. This will encourage users to do similarly rather than investigate the root cause of such issues. I believe this case study should not be used for this reason.

In such inconsistent sets of data, I imagine which route the models take through the data will be highly dependent upon their initialisation and, in the case of MCMC, are very unlikely to mix (as can be seen by the narrow uncertainty bands on each individual curve). I do not see this as a case of model averaging (where there are fundamentally different modelling assumptions which are all plausible which lead to different results) but *rather luck as to where you initialise each method. In this case perhaps it works ok as the methods happen to choose what look like the most extreme paths however I would think this is to a large extent fortuitous rather than by design. None of the methods individually fit the data at all well. I would argue that trying to average over a lot of models which are all individually catastrophically bad fits does not add much strength. I much prefer, and would suggest any user takes, the approach of Vyse et al. to go back and look at the raw data and understand what is happening.*

Aside: When plotting the data in the cores as solid circles (e.g., Fig 2) I would find it helpful to colour code according to the type of dating used (e.g., OSL, ^{14}C ...). This would make it much easier to identify and understand outliers/inconsistencies. Currently, the text around lines 405 - 410 are not understandable as a reader does not know which are the OSL and which are the ^{14}C dates.

Thank you for this comment. We are aware that Figure 3 is not the solution to our case study No. 2. We ran our simulations several times and can therefore assure that the path taken does not depend on the initialization of the MCMC. We argue that the path taken by each individual model depends on the underlying methodology of each age-depth modeling system. The results shown in Figure 3 represent the unoptimized solution for an inconsistent set of data. An inconsistent set of data is challenging for any age-depth modeling system, as we have shown in Figure 3. One option is to remove questionable dating points based on geoscientific knowledge for the specific region/lake. According to Lacourse & Gajewski (2020), 33% of the studies they reviewed rejected one or more 14C ages before modeling, primarily because of age reversals. With LANDO, we present an alternative where we use uninformed models and – in case of scattered age dating points – an adapted fuzzy change point detection method (originally proposed by Holloway et al. (2021)). The results shown in Figure 4 represent the optimized solution to case study No. 2. We agree that it is important for any study to look at the data and examine the reason behind the age scattering. We will therefore add a note to our revised manuscript that users should investigate the scatter in addition to using LANDO. If there are still dates in question, users can remove dating points from the input file and run LANDO again.

And thank you for the additional comment. Yes, we will continue to improve the plot function in LANDO.

Reference:

Lacourse, T. and Gajewski, K.: Current practices in building and reporting age-depth models, Quat. Res., 96, 28– 38, https://doi.org/10.1017/qua.2020.47, 2020.

Hollaway, M. J., Henrys, P. A., Killick, R., Leeson, A., and Watkins, J.: Evaluating the ability of numerical models to capture important shifts in environmental time series: A fuzzy change point approach, Environ. Model. Softw., 139, 104993, https://doi.org/10.1016/j.envsoft.2021.104993, 2021.

2.4 Method Description

While I appreciate that the paper is about the development of the coding, I think there needs to be a short description of each age-depth model for the user. This should not repeat the original papers but just give an intuitive explanation so the reader is aware what they are applying, and what the specific assumptions of that technique are. Ideally this section should include an informal discussion of the strengths and weaknesses of each modelling approach. With such a section, a user might be able to decide whether certain models are more appropriate than others for their setting.

Thank you for this suggestion. We will extend the part about the age-depth models in the method section (Page 3, Lines 103-108) by a brief description about the models.

2.5 Checking convergence

I do not know enough about the specific implementation of the age-depth models in their own packages but is there a way of passing information on model fit to the user. In particular, some of the methods rely upon MCMC and one needs to be assured of convergence; while the frequentist approach (undatable??) may also give some measure of model fit. It may be that the underlying code itself (BACON, BChron, clam, ...) does not provide this but is there a way to obtain information in LANDO that the results of the individual models are appropriate/have converged/or fit?

For example in case study 1, I wondered if it was realistic to have a sedimentation rate that varies from 0.002 to 2.486 cm/yr where the raw ^{14}C dates suggest an inversion? I do not know enough about this location but is it instead possible that some of the models are not fitting well, have been run with inappropriate parameters, or *have not converged properly. This is even more so in case study 2 where I would hope that individually all of the methods would tell you that they are not fitting the data well.*

I recognise this is more about the underlying code to which you link than the LANDO implementation | so there may be nothing you can do to resolve this.

About convergence: Some of the packages provide convergence information in separate methods within each implementation. However, there is no standardized reporting standard for convergence against which we can compare the models. First, we would have to develop such a reporting standard. Then we can build a separate pipeline within LANDO that tracks these convergence values for each sediment core and each modeling system, as our application should continue to work for multiple sediment cores. As this is more a larger feature request, we may address this in a separate technical note manuscript in GChron and update of LANDO. In the initial version of LANDO, there are two extras settings enforced to ensure that models behave appropriately. For Bacon, for instance, we increased the parameter "ssize" as a precaution to ensure good MCMC mixing, as suggested by Blaauw et al. (2021). In clam, we already use the information on the fit to determine the best model for each core, which clam provides as a direct output. About the example: The values you gave refer to the possible values of the outermost two-sigma ranges when considering all models for the entire core length. Both values do not occur at the same time. At the point of inversion suggested by the calibrates ^{14}C dates, we see high sedimentation rates and only a few models include the dating point at 114.75 cm in their two-sigma uncertainty range. We will rephrase this in the revised manuscript version.

Reference:

Blaauw, M., Christen, J. A., and Aquino Lopez, M. A.: rbacon: Age-Depth Modelling using Bayesian Statistics, https://cran.r-project.org/package=rbacon, 2021.

3 Suggested Additional Information

- *I would suggest you need to provide the link to the software very clearly and explicitly in the Introduction. This is what most users will want and currently you have to get to the end to find out how to actually access the software.*

  Thank you for this suggestion. We followed the rules of the GChron manuscript preparation guidelines that said that we must present the link to the code in the code availability section. However, if it is possible, we will also provide one in the introduction.

- *I think it worth perhaps adding a clear caveat that it is not appropriate for a user to try all the age-depth models and then simply select the single answer that they like the most in terms of fitting with a particular hypotheses. I know you are not proposing any user do this but a warning might help ensure proper use of LANDO.*

  Good idea. We will include a warning in the discussion section.

- *You seem to have missed the opportunity to discuss the practical differences between the various age-depth models in your examples. Can you identify features that always seem to be present for some models? For example, in case study 1, the red and green sedimentation estimates seem to be much more extreme than the other models. Is this a consistent feature? Are there reasons for this?*

  We will add a description of the practical differences in the method section about the age-depth models.

- *I do not understand Section 2.5.3 - this needs to be written much more clearly. Does this relate to the way that the proxies and the Holocene boundary are used to filter unreasonable models?*

  We understand that the section might feel a bit out of place. We will move it to a different section number and rename the section to "Further analysis - Sedimentation rate development over time". We wrote this section to describe our further analysis that we performed on the multi-core dataset. This section does not apply to the proxies that we use to filter unreasonable models.

- *Important: Table 1 seems to be lacking what I would guess is the most important piece of information - the depth of the measurements. The format of the data on github (and column labelling of the .xls spreadsheet also does not quite correspond to that given in Table 1 - although it is pretty self-explanatory how they transfer).*

  We included the depth information of the measurement in the "measurementID". In Table 1, we define the measurementID as "*Composite key composed of a unique CoreID, a blank space, and* ***the depth below sediment surface (mid-point cm) with max. two decimal digits of corresponding analytical age measurement*** *- example: "CoreA1 100.5", when users obtained sample of CoreA1 between 100 and 101 cm depth*". The measurementID is an essential key / identifier within our database, which we introduced in Pfalz et al. (2021). However, to accommodate other scientist, we have split the measurementID in the example spreadsheet into its two components: CoreID and Depth mid-point (cm).

  Reference:

  Pfalz, G., Diekmann, B., Freytag, J.-C., and Biskaborn, B. K.: Computers and Geosciences Harmonizing heterogeneous multi-proxy data from lake systems, Comput. Geosci., 153, 11, https://doi.org/10.1016/j.cageo.2021.104791, 2021.

- *I think it worth adding a note that all users of LANDO should also reference the underlying methods (and their papers) not simply LANDO.*

  This is a point, which we unfortunately missed. We will include this additional remark in our revised manuscript.

4 More minor comments

- *I find the use of "correlations" in the title an odd choice. Would it not just be better to have a title "Improving age-depth modelling by using the LANDO model ensemble"?*

  Agreed. Where possible, as part of the submission process for a revised version, we will change the title to "Improving age-depth relationships by using the LANDO model ensemble".

- *I agree with one of the comments from CC1 that the parameters chosen for each model are critical. However, I do slightly disagree that this is entirely the responsibility of the user. A reliable method should provide good default values, or ideally an automated method to select good parameters (possibly using e.g. cross-validation??)*

  We agree with both Reviewer No. 1 that parameter selection is crucial and your comment that a reliable method should provide good default values. As we said on Page 22, Lines 572-575 "*But we also wanted to simplify the process for users who do not have in-depth modeling knowledge. By using the default values, we can compare models based on their ability to work with the available data.* ***On the other hand, we are sure that the developers have set their default values based on systematic testing."*** Our aim is to compare age-depth modeling systems given solely the input data. Finding appropriate parameters for each of the individual modeling systems would require techniques, such as grid search, and would be highly dependent on the given input data. This would require direct involvement of the developers of all modeling packages to make suggestions to user on the possible parameters for given input data.

- *Table1 on pg4 - What is the relevance of thickness? It seems unclear to me what thickness of the sample layer means, where is the actual sample within the specific layer? If it is an average over the entire layer then the methodology will presumably become more complex since the ^{14}C determination relates to the average over the respective time period rather than a single calendar year.*

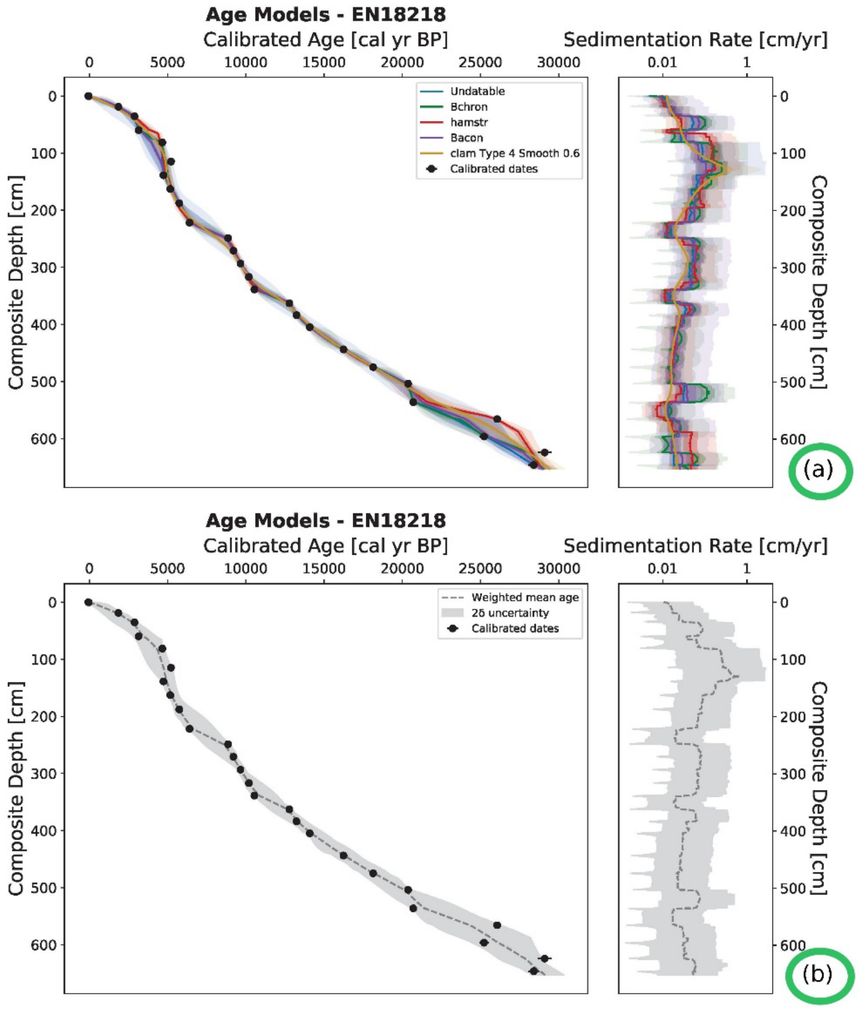
  Thank you for spotting this error. Thickness actually refers to the thickness of the sample used for age determination, not the layer. "*Thickness of the layer*" should be "*thickness of the sample slice*". In most cases measurements of ^{14}C sediment bulk samples require a sediment slice of at least one centimeter. However, we have some data entries consisting of bulk samples from larger slices, presumably to ensure a successful measurement due to the low carbon content within the sediment. Four out of five age-depth modeling packages within LANDO require thickness as an input parameter. Therefore, we included this parameter in our "necessary" category. The models do not internally average over the entire slice, but instead apply a distribution for each sample. We will correct the description of thickness to "***Thickness of the sample slice used for age determination in [cm]***"

- *Should one also have the option of selecting a specific radiocarbon calibration curve? This seems to be particularly relevant for marine ^{14}C samples where one might want to use a Marine curve with a ΔR*

  As we stated before, it is already possible to calibrate dates within LANDO using the Marine20 calibration curve. In the current manuscript, however, our focus is on terrestrial lakes from the northern hemisphere. In an updated version of LANDO for other use cases, we will include the option for selecting the calibration curve individually.

- *Figures lack actual panel labellings such as (a) or (b), e.g., there is no Figure 2a*

  All figures with panels should have a panel label to the right of the sedimentation rate curve. Please find below Figure 2, where we highlighted the location of the label with green circles. We apologize, if you have received a copy our manuscript without these labels.



- *Line 367'ish - you need to make clear that the mean/median figures relate only to this section (108 - 133 cm in depth) not the entire core.*

  Thank you for spotting this inconsistency. We will rewrite this sentence to "*All models revealed highest sedimentation rates for the interval between 108 and 133 cm. Mean values ranged from 0.242 cm/yr* *(hamstr) to 0.764 cm/yr (clam)* ***within this interval****, whereas the median sedimentation rate varied between 0.107 cm/yr (Bacon) and 0.314 cm/yr (clam).*"

- *When plotting the data in the cores as solid circles (e..g Figs 2) I would find it helpful to colour code according to the type of dating used (e.g., OSL, 14C ...). This would make it much easier to identify and understand outliers/inconsistencies. Currently, the text around lines 405 - 410 are not understandable as a reader does not know which are the OSL and which are the ^{14}C dates.*

  Thank you for this suggestion. Since we want to reserve colors for possible additions of age-depth models to LANDO, such as OxCal, we will present an alternative solution in the revised manuscript. We will change the legend of LANDO to use different symbols to indicate which category the dating point belongs to, e.g., "14C terrestrial fossil" dating points as a square, "14C sediment" dating points remain as circle, or "other" (such as OSL) dating points as triangles. We hope this will improve the readability of our plots.

- *Line 534'ish | Please identify and explain this unusual core in the section discussing he modelling rather than in the conclusion. I am presuming this is the observation in Figure 6 that has a latitude of around 75° where undatable has a huge difference from Holocene to Pleistocene which is not replicated in the other models.*

  We wrote on Page 21, Lines 533-536: "*For instance, during the examination of the Holocene and the Pleistocene sedimentation rates (Figure 6), we noticed that one sediment core (PG1228) had an extremely high mean sedimentation rate for the Holocene dataset in Undatable. Similar to the second case study ("Inconsistent sequence" – CS2), we found scattered age data points for this sediment core, which influenced the modeling process of Undatable.*" The paragraph you are referring to is in the discussion section, which discusses all the different scenarios, not in the conclusion section. We identified the core as PG1228, explained that the deviation in sedimentation rate is due to scattered age data points and – similar to case study No. 2 - the different path Undatable takes through the scattered data points. However, due to the community comment, we will revise this graphic with the changed input data and ensure we mention any new observations.