Review of "Improving age-depth correlations by using the LANDO model ensemble"

January 12, 2022

1 Paper Summary

The paper describes a front-end Jupyter Notebook which allows users to fit multiple different age-depth model (BChron, BACON, clam, hamstr, and undatable). Currently these models are all only available in individual packages (I believe four in *R* — BChron, BACON, clam and hamstr; and one in MATLAB — undatable). This separation of the age-modelling approaches into individual packages has hindered the community from testing, using and comparing the various age-depth methods.

The aim of the work is also to allow the user to perform model averaging whereby the output of the various age-models can be combined. This is based upon the idea that one should investigate the sensitivity of the output to the specific modelling assumptions which underlie each age-depth approach — none of these assumptions are likely to be entirely correct and so one should see how much difference they make. By creating an ensemble, one can investigate the spread between the various models. While all the models are wrong, hopefully by considering lots of them you can understand how much of a difference the various modelling assumptions make. From that one might aim to get an idea of the further uncertainty due to age-depth model dependence. If all the methods show similar results this boosts confidence, if they do not one should recognise this model dependence.

The work is a valuable contribution to the community who I am sure will benefit from being able to fit multiple models through the same front end. I do however have some points regarding the presentation and case studies which I think need addressing.

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 affect 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.

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.

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.

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.

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.

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

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

- 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.