Responses to Bryan Lougheed

Dear Bryan Lougheed,

Firstly, we would like to thank you very much for taking the time to review in detail our manuscript on age-depth relationships in our LANDO model ensemble. We are very pleased that you acknowledge the scientific significance and implications of our study.

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

Thank you again for investing your valuable time in helping us improve our manuscript.

On behalf of all the authors,

Gregor Pfalz

Main points

The main issue with the manuscript as it currently stands, in my opinion, is not related to the software itself or how it is described, but how the manuscript uses the LANDO software in an exercise in interpreting the performance of the various age-depth modelling software packages. When comparing the different packages, the authors state:

> *"To lower our impact and to avoid introducing biases in the modeling process, we used the default values from each modeling system as our own default values (Blaauw et al., 2021; Blaauw, 2021; Parnell et al., 2008; Dolman, 2021; Lougheed and Obrochta, 2019)."*

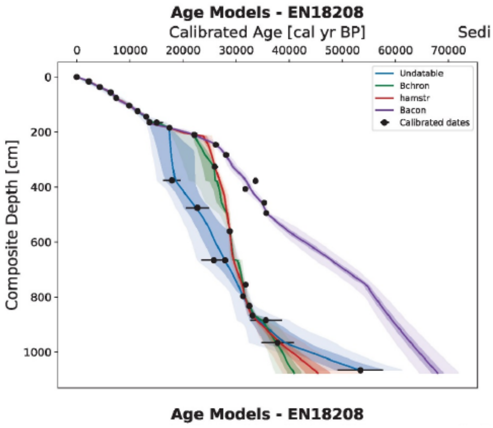
The above highlights the general issue with the parts of manuscript that compare age-depth modelling software packages. All of the age-depth model software packages in the manuscript are compared using "default" settings, but all of the packages have settings for a reason, namely that they should be adjusted. So it is possible that the age-depth software packages are not compared on their merits. I note that the LANDO software has the option to adjust the settings for each software package, so I am not describing a limitation of LANDO here.

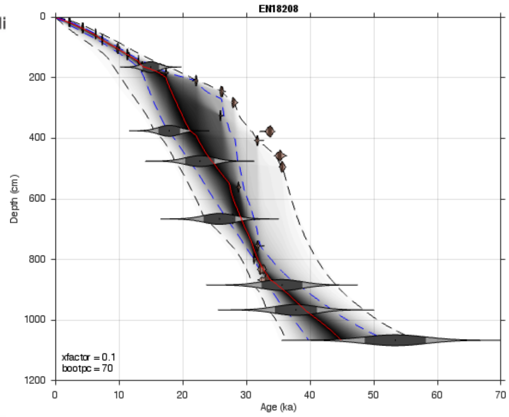
I can give an example about how using "default settings" can affect the interpretation in the case of Undatable. Figure 3 in the Pfalz et al. manuscript suggests that Undatable exclusively follows the younger dates between 200 and 600 cm, and the authors mention something similar in the manuscript in lines 408 to 410 of their manuscript.

While it is true that the GUI version of Undatable displays some settings in the data entry windows when the GUI first boots up, these are by no means "default values", but rather starting/dummy values in the GUI. The Undatable paper (Lougheed and Obrochta, 2019) discusses that bootpc (bootstrapping percentage) should be increased in the case of large age-depth scatter or age reversals. Indeed, dealing with scatter in this way is stated in Lougheed and Obrochta (2019) as one of the main advantages of

Undatable. Seeing as core EN18208 contains such scatter, I have rerun Undatable using a bootpc of 70 (after Gregor Pfalz kindly shared the input data with me), with the following result, with Pfalz et al. Figure 3 shown for comparison:

Pfalz et al. Figure 3 *EN18208 with Undatable, 70% bootstrapping*





In the above example, the Undatable uncertainty range expands to take into account the scatter of the dataset, and between 200 and 600 cm the highest probability area shifts more towards the centre of the age-depth scatter. This is the intended philosophy behind the deterministic Undatable, namely that the uncertainty range of the age-depth model should increase so that the scatter of the age-depth points is taken into account, i.e. 95% of the age-depth points should feasibly be located within the 95% uncertainty range of the age-depth model.

Other age-depth modelling packages also have their own settings and approaches.

We agree that the parameter selection can directly affect the interpretation of each use case. However, in this manuscript, we focus on how well the models perform on specific input data. As you have shown, increasing the uncertainty band for Undatable would certainly allow all dating points to be included, but still, the mean values of the model and higher probability ranges would not match the model developed by Vyse et al. (2020).

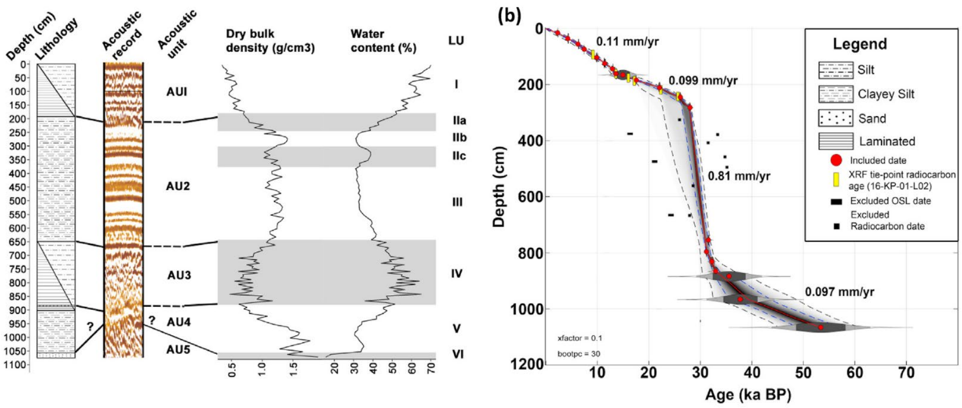


Figure of age-depth model by Vyse et al. (2020)

If one considers that Vyse et al. (2020) used Undatable for their age-depth model but removed dating points prior to execution, this shows that the model performance depends primarily on the input data. But we also agree with Reviewer 2 that a reliable method should give 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.***" 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.

It is further questionable which of the two approaches (large uncertainty range vs. narrow uncertainty range) the user will want. We agree that a larger uncertainty band is desirable for paleoenvironmental reconstruction, when propagated into proxy interpretations. However, Lacourse & Gajewski (2020) found that "*[a]lthough 84% of the papers in our literature sample showed uncertainties on a plot of the age model, none made explicit use of the uncertainties in their paleoenvironmental reconstructions".* Since we made sure that user in LANDO can change the settings in each of the modeling systems, which you thankfully mentioned, they can examine the impact of each input parameter on their model and input data. Overall, we agree with your comment and recognize that this is a broader topic that potentially feeds further research projects related to data science methods.

Reference:

Vyse, S. A., Herzschuh, U., Andreev, A. A., Pestryakova, L. A., Diekmann, B., Armitage, S. J., and Biskaborn, B. K.: Geochemical and sedimentological responses of arctic glacial Lake Ilirney, Chukotka (far east Russia) to palaeoenvironmental change since ~51.8 ka BP, Quat. Sci. Rev., 247, 106607, https://doi.org/10.1016/j.quascirev.2020.106607, 2020.

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.

Other points

A small point regarding interpreting a lack of age-depth reversals as "undisturbed sediment"... Following bioturbation theory (e.g. Berger and Heath, 1968) when the sediment is fully uniformly mixed throughout the deposition history, downcore multispecimen / bulk samples will produce age-depth points that are in chronological order, i.e. lacking age-depth reverals. In other words, a lack of age-depth scatter is not an indicator for undisturbed sediment (despite perhaps 90+% of the literature assuming otherwise).

Thank you for this comment. As you rightly pointed out, most of the literature suggest that term undisturbed sediment refers to a lack of reversals. We will change the name of the case study from "undisturbed sediment" to "continuously deposited sediment" to ensure that our work also considers for bioturbation theory.

When describing the performance of the age-depth models, the text describes the age-depth models from top down, whereas most of the algorithms operate in the direction of sedimentation/time, i.e. from bottom up.

Thank you – we will change the order in how we describe the performance of the age-depth models in the revised version.

In the age-depth model figures, calibrated dates are indicated by black dots with error bars. Please add some information in the legend or caption about what the black dot is (median or mean calibrated age?), and the error bars (+/- 1sigma, i.e. symmetrical error bars, or the central 68% range, i.e. asymmetrical error bars).

Thank you for the suggestion. We will update the figure captions to further describe the black dots - for instance in Figure 2: "*Generated output from LANDO for sediment core EN18218 (^{14}C data from Vyse et al., 2021). Panel (a) consists of a comparison between age-depth models from all five implemented modeling systems (left plot) and their calculated sedimentation rate (right plot). Colored solid lines indicate both the median age and median sedimentation rate for all models, while shaded areas represent their respective one-sigma and two-sigma ranges in the same colors with decreasing opacities. Panel (b) shows the ensemble age-depth model (left plot) and its sedimentation rate (right plot). The dashed line in panel (b) represents the weighted average age estimates (left plot) and the weighted average sedimentation rates (right plot) for the ensemble model, while the grey area represents the two-sigma uncertainty, i.e., the outermost limits of two-sigma ranges from all models. Both plots on the left of (a) and (b) show the depth below sediment surface on the inverted y-axis as composite depth of the sediment core in centimeter (cm) and the calibrated ages on the x-axis in calibrated years Before Present (cal. yr BP, i.e., before 1950 CE).* ***Black circles within (a) and (b) indicate the calibrated ^{14}C bulk sediment samples with their mean calibrated age using the IntCal20 calibration curve (Reimer et al., 2020), and their one-sigma uncertainty as error bar.*** *The plots on the right display the sedimentation rate in centimeter per year (cm/yr, x-axis as log-scale) against the depth below sediment surface as the composite depth of the sediment core in centimeter (cm, inverted y-axis).*"

In addition, 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" dating points as triangles.

The "optimised" age-depth model in Fig 4c takes what can be described as a middle route through the age-depth points, but with very small confidence intervals. It could be argued that such small confidence intervals mask the scatter of the age-depth determinations, and therefore the true geochronological uncertainty. This is more of a philosophical point, however, seeing as some age-depth packages try to find an optimised route between age-depth points with minimal age model uncertainty (e.g. Bacon, Bchron, OxCal), whereas others also expand uncertainty to take into account the scatter in age-depth points (e.g. Undatable). An argument can be made for either approach, but in a manuscript that compares all the different types of approaches, it would be useful to point them out.

We agree that there are several ways how geoscientists approach the age-depth modeling process, as we mentioned in the previous answer. In our case of Figure 4c, the optimized output has a similar form to the original age-depth model from Vyse et al., 2020. However, we agree with your suggestion on pointing out the different approaches. We shall add another section to the discussion part of the revised manuscript, to ensure that users can decide which approach they prefer.

Reference:

Vyse, S. A., Herzschuh, U., Andreev, A. A., Pestryakova, L. A., Diekmann, B., Armitage, S. J., and Biskaborn, B. K.: Geochemical and sedimentological responses of arctic glacial Lake

Ilirney, Chukotka (far east Russia) to palaeoenvironmental change since ~51.8 ka BP, Quat. Sci. Rev., 247, 106607, https://doi.org/10.1016/j.quascirev.2020.106607, 2020.

Regarding the title, "correlation" would (in my mind, anyway) refer to a statistical relationship of some kind between age and depth. So perhaps replace "correlations" with "relationship" or "models"?

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