Dear Editor(s) and Authors,

I have read the manuscript 'The conflict between sampling resolution and stratigraphic constraints from a Bayesian perspective: OSL and radiocarbon case studies' in detail. It is a well structured and well-written manuscript, and the scope fits Geochronology. You address a standing topic and issue in age-depth modeling, and your contribution is relevant. While your conclusions are not entirely new, these are based on real and well selected datasets, and your manuscript is a welcome part for the scientific literature. **I clearly support publication in Geochronology after revisions.**

We would like to thank the referee for such positive feedback. We would also like to emphasize that the review comments provided below are – in our view – very constructive, and we thank the referee for such a thorough review.

The authors use two case studies to point out challenges with Bayesian age-depth modeling – and rightfully demonstrate that the tested models are not without bias and artifacts. I particularly like – and agree with - the repeated call for applying common sense and looking at data with an experienced eye of a geochronologist even when applying models, e.g. as (quotes from your submission follow) 'As frustrating as it may be, in our view none of the tested models can tell us anything better than the actual data themselves', and as 'when testing any chronological model, it is of utmost importance to compare the model outcome with the input data.'. I fully agree and find this an important lesson: look at data, know possible issues – and then think if a model may help and/or is of any help.

Thanks again!

The final statement 'Our study shows that this goal [make use of prior observations to refine the precision, accuracy and robustness] is difficult to reach and that using models to correct measurements appears to be dangerous.'. Well – that really depends on the case and individual data structure in my opinion, and such a general statement should be at least softened when based on two datsets only (why is no reference to the often used models BChron & Bacon made?), and few datasets which are indeed challenging.

Agreed – we have changed our wording, by specifying that according to our case studies, it is high-resolution datasets that push existing models to their limits (more below on BChron and Bacon).

With this I come to my main criticism of this manuscript: the arbitrary selection of models, seemingly influenced by previous work of the authors. When speaking of luminence modeling I ask you to refer to ADMin (https://www.sciencedirect.com/science/article/pii/S187110141730047X) – probably the model least affected by the spread effect (?), but at the same time slow/unsuitable for large (and these?) datasets.

We have read the article suggested by the referee (Zeeden et al., 2018); however, we find the ADMin model slightly difficult to describe, since it is not defined by any equation. In addition, the nature of errors (systematic VS random) is ignored while the total uncertainty budget is fixed, which seems contradictory. Calibration errors are not random, by definition; so, when in Figures 2 and 6A the model predicts 100% of uncertainty as random, it makes no sense from a physics point of view (there has to be some systematic error). Moreover, the probability density is increasing with the fraction of random uncertainty, suggesting that this function is not integrable – thus, such a function cannot be considered as a probability density. Finally, about the spread effect: it is present in the ADMin model, by definition (since this model only accepts sequences of ages strictly in order); it is indeed visible on the ages presented in Fig. 4 of Zeeden et al. (2018).

Generally, I disagree with the BChron and Bacon models not even being mentioned, as these are really often used.

The reason for not mentioning BChron and Bacon is that our article deals with age estimation under ordering constraints, not with age-depth modelling. These two issues are quite different and, from a modelling perspective, independent questions: in the first case, an age is estimated for each measured sample from a suite, whereas in the latter case an age is estimated at any given depth in a profile – in light of the measured samples. Notably, in our second case study (as is the general case when working on archaeological sites), depth is not relevant – simply because stratigraphy does not follow depth/altitude.

Further comments.

References to *Ramsey* should in my opinion be to *Bronk Ramsey*

Agreed. We found the two occurrences in the literature, but after checking papers by the mentioned author himself, we now refer to Bronk Ramsey.

Line 84: ‘event model of Lanos and Philippe (2018)’ – could you please introduce this one – it is less known than the one by Bronk Ramsey which you introduce in detail

Agreed.

Line 115: please explain ‘Theta matrix’

Agreed.

160ff: Can BayLum model 14C ages!? - that would be different than luminescence modeling, because here 'only' the 14C age is used?

Yes indeed, BayLum does allow modelling 14C ages – as detailed in Philippe et al. (2019).

In Fig. 3 (and others) please include original ages. The problem is that Fig. 3, like Fig. 4, are outputs generated by software (Chronomodel for Fig. 3 and BayLum for Fig. 4); so, unfortunately, we cannot change these figures.

Generally, I find your figures would benefit from clearer explanation in captions, and systematically placing units on axes - ideally all would be on the same age (ka or BC ,please dont mix here).

Agreed.

Abscissa of Fig. 3 : space before bracket missing

Agreed.

Figure 5 and its explanation: ordinate unclear. Why was this only done for BayLum?

Agreed – we have reformulated the caption of Fig. 5 and added more explicit explanations in the text. About why this was only done with BayLum: because the comparison on OSL data with other models would be blurred by the fact that only BayLum starts from aw measurements of OSL data and includes shared errors across OSL samples. In other words, only BayLum takes the specificities of OSL ages into account, whereas all other models treat OSL ages as Gaussian probability densities (with only random errors). So, while the comparison between OxCal, Chronomodel and BayLum is justified for radiocarbon datasets, it is not the case for OSL. We stated this in our introduction (l. 97-100): 'To compare chronological models when confronted with high-resolution datasets, we turn to radiocarbon dating. Indeed, in BayLum OSL measurements are combined in a hierarchical model linking regenerative doses, individual equivalent dose estimation, the central dose parameter of interest, etc. Conversely, OSL ages can be included in OxCal and in Chronomodel, but only in the form of Gaussian probability densities (in practice, an age and its uncertainty). *In particular, unlike in OxCal and Chronomodel, in BayLum it is possible to model shared errors across OSL samples arising from, e.g., equipment calibration errors*.' Note: we added the sentence in italic to make our point clearer in the text.

Line 278: please explain the phase structure here. We are not sure what the referee is suggesting here, since we reproduced the model of Bayliss et al. (2015) and explicitly refer the reader to this original publication.

284: 'between samples OxA-9893 and OxA-23251' – please mark in Figure so that these can easily be found. Agreed, we find that this is an excellent suggestion and we hope it will help readers recognise similar features in their own studies. We actually added two brackets, in black, to highlight the (in our view) two clearly visible concentration effects: one concentrating ages towards younger periods (bottom bracket, between samples OxA-9893 and OxA-23251) and one towards older ages (between samples OxA-9776 and UCIAMS-103138). We added a sentence corresponding to the second sample set in the main text.

286f: I disagree with your statement 'These two bottom-most samples are PL-980252A, whose age lies outside the calibrated age of all samples above' – the densities do overlap. Agreed, there is indeed a tiny probability that sample PL-980252A has an age consistent with the samples above. We added 'almost entirely' in the sentence.

Chapter 3.2.2., and Fig. 8 limited to the lower 17 samples – was the model run for all or these samples? The model was run for all samples, but the figure only shows the 17 bottom-most sample; otherwise, the figure is very difficult to read. We are now more explicit in the text.

Line ~322: please highlight where the spread effect is pronounced why

Same as above regarding the suggestion to highlight the concentration effects, we thank the referee for this very helpful comment. We added another bracket in blue to highlight he spread effect visible in the top part of the sequence. We also added text to describe our observation: prior to modelling the ages stratigraphy, all calibrated ages are essentially the same; yet, in the modelling output, the ages are inconsistent because they are spread apart.

Fig. 10: units on both axes missing – please also include original dating - either as distribution or mean ages. We tried this suggestion, but reading the figure becomes very difficult because of the large number of samples. We believe that adding another dataset is only useful with a smaller number of ages (Fig. 11). Finally, we added units (years) on the x-axis – the y-axis does not have any unit.

352ff: given that Chronomodel and OxCal partly do not overlap the praising of larger uncertainty alone seems unjustified. Here we disagree, because uncertainties estimated with OxCal appear very small (too small in our view) compared to measurement uncertainties. Since the two models disagree, we feel like it is OK to write that 'Chronomodel appears to lie on the cautious side of things' and that this is perhaps an advantage.

In chapter 4.1. I find a prominent feature missing: The duration of the sequence when using OxCal is much shorter than when using BayLum or Chronomodel. This is worrying in my opinion, and the OxCal results seem much more similar to original ages than the BayLum and Chronomodel results. Especially the outer model ends seem unrealtistic long in BayLum and Chronomodel. The spread effect of the whole sequences seems therefore best captured by OxCal.

Agreed – to some extent. Indeed, BayLum and Chronomodel are strongly affected by the spread effect – while in OxCal it is balanced by a concentration effect. Even if we have not quantitively estimated the duration of the archaeological sequence of Çatalhöyük, we agree with the referee that BayLum and Chronomodel will overestimate this duration. This being said, OxCal will underestimate it (since it underestimates the start of the mound). Overall, it is difficult to say whether OxCal or BayLum best captures the duration of the whole sequence – both models are presumably in error, in two opposite directions.

In line 415 I suggest reference to

https://www.sciencedirect.com/science/article/pii/S0277379103003160

https://journals.sagepub.com/doi/full/10.1177/0959683616675939

As stated above, we are not dealing here with age-depth modelling – so we feel like referring to these two articles is no justified.

It is really good to see the computer code in Supplements. Yet I am wondering why this is only the case for one of the two examples. Further, R code would benefit from better documentation, please do so that also non-R-familiar colleagues can follow what is done why.

Agreed in principle; but the OSL data is not ours, so we cannot publish it. Hence, it will not be possible for readers to reproduce the calculations on the samples used in this study.

Further, I would like you to provide results (data plotted in Figures) in Supplements. We are not sure which data the referee is mentioning here. Which data? In our view, the ten figures already included in the article suffice; and all data are provided in tables for the reader to produce additional figures if deemed necessary.

I am aware of issues with suggesting literature in the review process, and I am asking the editors to have a critical look at these – yet I ask you to consider including the information contained within the suggested literature in your manuscript.

References

Zeeden, C., Dietze, M., & Kreutzer, S. (2018). Discriminating luminescence age uncertainty composition for a robust Bayesian modelling. *Quaternary Geochronology*, 43, 30-39.