Dear Dr. Wichern and Co-Workers,
the reviewer was pleased with the overall revision of the manuscript but suggested a few more minor adaptations, which I believe will be easy for you to implement. It seems the reviewer made a technical mistake and entered their suggestions in a form that is only visible to the editor. I am therefore pasting the very helpful comments here into my response to you. Please be invited to upload a new version of your manuscripts after minor revisions. Please provide a manuscript version with changes tracked along with your submission. Let me know in case you have any questions.
Best
Tina Treude

Reviewer #1 Comments:

Since the first version, the reframing of the introduction makes the scope and purpose of this paper much clearer. Overall the paper is much easier to read than before. I applaud the authors revisions and encourage acceptance following minor/technical revisions detailed below. All these suggestions should be able to be addressed by adding a single sentence or few words.

The authors are glad to hear that the revisions have improved the clarity and readability of the manuscript. We want to thank the reviewer for their helpful comments, and we have incorporated these further points of improvement into the manuscript.

Line 64: "predicted for moderate IPCC pathways"…. We are already > 400ppm? Is this a relevant mention anymore?

It is true that this is no longer relevant for the atmospheric CO_2 concentrations, but it remains relevant for (short-term, end-of-century) temperature changes. We have amended the text to reflect that this comparison focusses on temperatures only.

Figure 1 : the glauconitic sand and clayey sand are hard to distinguish. Is there any area that is NOT clayey? Make this visually more distinguishable?

Most intervals have some clay content, except for the Luchtbal Member. The top of the Oorderen and the entire Kruisschans Member are more clay-rich than the rest of the Lillo Formation. The horizontal lines that indicate clay content have been made longer and thicker. Unfortunately we had to revise some of the stratigraphy at the last moment, so the figure has been altered in general.

Line 120: "Their" should be "they are"

Thank you, this has been corrected.

Line 171 + following paragraph: how much material (weight, mg or ug) is typical for one drilling transect (one sample)?

This information is already noted at the beginning of section 3.5.1. Line 162: 'Samples of approximately 100-300 µg were taken from the outer surface of two specimens..'.

Line 216 – this bracketing correction for ETH3… Is this supposed to be equivalent to the ETH1/2 slope correction step? I can see how that is partially getting at the same issue – you are basically making a line through two points that are (d47 of measured ETH3, D47 of measured ETH3) and (d47 of 0, true value of ETH3) to access that slope, except it does not take into consideration the d47 of your samples. If that is different from the d47 of ETH3, then you are either over (0 d47 ETH3). I doubt this will make a big difference in the end, but I have not seen this type of correction applied before and it might merit slightly more explanation.

We carried out the initial ETH-3 correction to average out variability within single runs (so, across several hours). ETH-3 was chosen for this bracketing as it is closest in composition to the samples, which is stated earlier in the paragraph. Only after that we apply the d47-D47 correction with the other two ETH standards as well. The latter was done with ETH standards that were measured over a span of several months.

The text has been edited to state explicitly that this ETH-3 correction is related to intra-run variability. We hope this clarifies the difference between the two.

Line 248 – what about the offset between d18Oc from Gasbench and MAT253 shown in Figure B6… its something like 0.5 permil to my eye. Does this contribute to the uncertainty of 0.15? or did you correct for this offset first before looking at the variability?

This offset (indeed ca. 0.5 per mil) is related to the fact that the GasBench d18O data required a correction for Mass 44 variability (Fig. B3). The MAT253 data did not require such a correction. After this correction, GasBench and MAT253 data gave similar values. Only after this correction did we look at the uncertainty within the data.

We clarified the text to say that all standard deviations were calculated after the Mass 44 correction was carried out (section 3.5.2). We also explained in the caption of Figure B6 that the offset stems from the fact that the GasBench data shown here are not corrected for Mass 44 variability yet. We hope that this clarifies the concerns about this offset.

Line 325 – could add a supplemental plot of Fe vs. Sr. Are the higher Fe outliers correlated to the lowest Sr values? What thresholds do you use to define low and high Fe. Some of the Fe values of 1000-4000 seem high to me. I see the bulk of the data is low, suggesting good preservation, so I don't think this changes the conclusion.

We do not define single 'threshold' values as these can have their own issues, rather we look at the overall distribution of the Fe and Mn kernel densities. These are strongly skewed towards low values, with only a few higher datapoints. We therefore conclude that the shells are overall low in Mn and Fe. The caption of Figure has been edited to explicitly mention this skew towards low values, while acknowledging that there are some high value outliers.

The reviewer is right in that it may be interesting to look at Fe vs Sr correlations. However, this is not within the scope of this paper as 1) the Fe and Mn density distributions already suggest good preservation, and 2) EBSD was employed to look at and rule out minor, localized diagenesis specifically (which is what correlated Fe and Sr in a few select areas might point to).We therefore have decided against incorporating such a plot in the supplments.

Figure 7 – dark and light growth bands shading is hard to distinguish in color. Do you need light growth band shading at all? What are the red and blue zones. Those could be included in the legend as well or labeled on the plot with text (instead of only in the caption)

The contrast between light and dark growth bands has been increased, so that it is hopefully easier to distinguish. The growth bands are necessarily, as in specimen SG-126 they can be linked to winters (dark) and summers (light), which suggests slower growth or even stops in winter (see the discussion in section 5.3). The red and blue zones are now also explained with text on the plot. We hope these changes clarify the figure in question.

Figure 8 – why did you choose not to assign a peak around 21mm distance from umbo as an annual peak? Its about as high as other peaks later on in the shell. How would including or not including this peak change your results? Does it lead to a worse fit with the gompertz and von B equations?

The peak at 21.02 mm was not incorporated, as it, despite its height, only consisted of a single datapoint. The other counted 'years' all consist of a peak, or at least a slope, of two or more points. This was, however, not explained in the text. We agree that this should be clarified.

It only minimally changes the output of the Gompertz function (from a maximum height of 34 to 31 mm). The change for the VB function is more significant (from 52 to 40 mm). The residual errors are worse for this fit: 0.6 instead of 0.5 for the VB equation, and 0.8 instead of 0.6 for the Gompertz equation.

To be more transparent, we have included an alternative growth fit in the appendix (Fig. B7). The text has been amended to explain why this peak was not incorporated into the main figure.

Line 432 – this discussion could have citations of de Winter 2021 (climate of the past, optimizing sampling strategies in…) and Zhang and Petersen 2023 (paleoceanography and paleoclimatology, clumped and oxygen isotope sclerochronology methods tested in..) where the trade offs are discussed.

Thank you, the authors agree. Both citations have been incorporated into the relevant paragraph.

Figure 9 – why not state numbers here. What are the D47 means you show in the large symbols. What is the T that corresponds to? Could add two more likes for D47 = ___ and T= ___ under the d18O<0.9 text.

The authors have explicitly chosen not to convert the D47 values to temperatures, as their large errors render them essentially meaningless. This is explained at the end of paragraph 4.4.3. We have not made any changes based on this comment.

Line 445-446 – Can you independently define an uncertainty that is appropriate? You say it "reduces the conf. intervals"… to what? I think you are trying to indicate that the error bar plateaus in size. There are big reductions between N=0 and N=20 and then smaller reductions as N=20 to N=40. Maybe reword this part to make this clear.

Yes, the authors did indeed mean that the error bar plateaus in size. A sentence has been added to clarify this.

Line 452 – what is a "reasonable level"? This error is NOT reflecting only measurement uncertainty but also the seasonal range. A highly seasonal place would never be able to get a small uncertainty no matter how many samples you ran.

The authors agree that this phrasing is ambiguous. We have changed it to 'a few degrees Celcius' to distinguish it from the larger error bars (some of which correspond to 10 degrees Celcius or more. This is of course no longer useful when the temperature in question is only 13.5C).

The second point that the reviewer raises is important, and we have added a sentence explaining that part of the error originates from seasonality itself and it therefore cannot be reduced even with more data.

Line 505 – "usually homogenous"…based on what? And "similar to A. nysti"… Cite something here in these places?

Citations have been added to both statements.

Line 590 and paragraph – How does the 12C MART from d18Oc compare to that estimated from D47-seasonality (the numbers you fail to include in figure 9)? If the modern seasonality in that region is smaller than 12C, and the D47-seasonality gives you a smaller range, why did you rule it out? Is there any info on how modeled Pliocene seasonality relates to modern seasonality for this region (saw an abstract by one of the authors…)? Would you expect it to be more or less? It may be that true seasonality is low and the D47 is MORE accurate… you are placing extra weight on d18Oc-based measurements, perhaps more than seems right. Why not just say "D47-based seasonality is X, d18Oc

says Y, it's probably somewhere in the middle". If you suggest your d18Oc range is likely an underestimation, are you suggesting the seasonality in the north sea was much GREATER during the Pliocene? Why would that be?

See one of our previous comments. It is true that the actual number from D47 would probably be 'somewhere in the middle' (strongly depending on which binned cut-off you pick…), but with the large error bars we could not use these data to say anything meaningful about Pliocene seasonality anyway.

Many of the factors that may result in an underestimation of seasonality for d18Oc, as discussed in section 5.5, also apply to clumped isotopes (minus the d18Osw influence, but with the added caveat that you would likely always need to combine many summer and winter datapoints that may not all represent 'peak' summer or winter). This also has been added to the discussion in section 5.5. So, seasonality is more likely to be under- rather than overestimated in both approaches.

Finally, there are several studies that also suggest higher-than-modern seasonality during the mPWP in the North Sea. These are cited at the end of section 5.5. We therefore feel that what our limited data suggests regarding seasonality is rooted in reality. However, we do not want to go into further detail on any climatological factors that could explain this higher seasonality, as our data is too limited to do so.

We have not made any changes to the text regarding this comment, except for a few sentences on section 5.5 on clumped isotope limitations.

Line 600 – here you discuss d18Osw fluctuations in terms of d18Oc seasonality estimates, but is there anything further you can say about the average d18osw value you reconstructed? How does it compare to waters in this region today, on average?

The average $\delta^{18}O_{sw}$ of 0.10±0.88‰ VSMOW is within the range of ca. -0.3 to +0.3‰ for the modern North Sea, although it is on the heavier side for coastal areas which are closer to -0.2 to -0.3‰ (Harwood et al., 2008). We added this information to the manuscript.

Figure B2- great figure. Could add that in the drilling example #3 that not only are the M-1 layers of different age, they are precipitated from a different internal body fluid and may record vital effects not present in the outer two layers.

Thank you, the authors are glad that this figure clearly illustrates the issue at hand. The additional information has been added to the caption of Figure B2.

Figure B5 – the y-axes of the D47 plots (a, d, g, j, etc) are all different scales, making it look like some of your Eth standards are much more/less variable than others. I also still am surprised that you can possibly get d18O and d13C fractionation without too much offset in D47.

This is true. However, the aim of these plots is to show a lack of drift over time in D47/d18O/d13C values for each standard. The y-axes do not need to be compared between different standards for this purpose. We therefore have decided to keep these plots as it is. The corresponding data is also available in the supplementary materials.

Figure B6 – this offset in raw d18O is pretty large (0.5 permil). Is this corrected for anywhere? I mentioned this figure earlier as well…

See the answer to a previous comment. It is now explicitly stated in the caption of Figure B6 that the Mass 44 offset correction (Fig. B3) corrects for this offset.