We would like to thank the reviewer for their careful reading of our manuscript and their constructive comments. Below we have copied the review in full and provide our response in orange text.

Text quoted from the original manuscript is in grey and proposed changes based on the review are in blue.

We feel that thanks to these suggestions the manuscript will improve considerably and hope that our proposed revision will meet the criteria for publication in Climate of the Past.

Lukas Jonkers
On behalf of all authors.

The manuscript by Jonkers and colleagues compares multiple samples of the stable isotopes from the shells of the planktic foraminifer N. pachyderma from the same sediment trap samples. They then use a combination of nearby hydrographic records, modeling, and statistical analyses to assess the variability within a population not attributable to environmental factors, primarily temperature. They find a substantial amount of variability in multiple samples from the same cups, which is used to illustrate the inherent “excess” variability of reconstructions using very few shells. With increasing use of high resolution instrumentation making use of small samples and individual foraminifera analysis (IFA) more frequent, the implications of these findings are important.

I have a few suggestions which I hope the authors will find useful. My primary suggestion for the manuscript is to do with framing. From line 1 of the abstract, the rationale of the study is laid out to be an estimate of excess variability in individual shells measurements and therefore utility of IFA. The catch is that the methodology used here is not IFA but rather multiple pooled samples. Several assumptions are required to make the leap from environmental data and pooled measurements to an estimate of excess variability by a theoretical IFA measurement, some of which require additional justification. My comments include a few specific suggestions of where this may be helpful. However, it is also my opinion that the framing of this manuscript as a quantification of IFA excess variability may be slight overreach drawing from this particular dataset. There are certainly implications for IFA, and the rough calculation done here are useful in illustrating that. However, given the number of assumptions required and the use of pooled rather than individual shells in the analyses, overemphasis on a quantification of “noise” in IFA analyses specifically, may do a disservice to the really important findings of large excess variability.

We thank the reviewer for their constructive comments. We agree with the reviewer that our quantification of the excess variability requires more discussion and will add the following paragraph in section 3.3: “Whereas our modelling approach provides an estimate that is likely closer to reality than assuming that foraminifera reflect environmental conditions averaged over a single (calendar) month, our estimate could be evaluated by simulating other calcification trajectories. We found that our results are insensitive to the duration of chamber formation and experiments where we allowed complete shell formation within one day, equivalent to assigning all weight to the last chamber, yielded an expected 0.09 ‰ standard deviation of individual foraminifera δ18O. Therefore, the assumption of equal weight of the four chambers has little bearing on our results. Ultimately, the modelled foraminifera δ18O depends on the hydrographic data used to estimate δ18Oequilibrium. By using data from the surface and from great depth, we have obtained two end-member scenarios of vertical δ18Oequilibrium variability that implicitly encompass ontogenetic vertical migration. However, future estimates of expected individual foraminifera δ18O variability could be improved by explicitly incorporating horizontal δ18Oequilibrium variability and advection during shell growth in the modelling strategy.

Apart from being sensitive to our modelling design and data availability, our estimate of excess δ18O variability among individual shells is also sensitive to the quantification of variability among shells. To obtain a conservative estimate we excluded potential outliers. Were we to consider all measurements, the average standard deviation among groups would be 0.15 ± 0.11 ‰ (0.17 ± 0.09 ‰ during spring) and the resulting excess δ18O

variability 0.25 ± 0.19 ‰. Thus our approach yields a conservative and better constrained estimate of the excess variability."

We will also make sure to be more careful with our wording regarding the estimate of the excess noise in the abstract and in the conclusions. However, we think that our phrasing in the main text (e.g. "Assuming that our simulations are a reasonable approximation of reality, the excess variability (s.d.) that cannot be explained by variability in temperature and δ18Oseawater is therefore 0.11 ± 0.06 ‰.") is not overselling the results and we would prefer to keep the original text here.

Minor/specific points:

111: Why were outliers removed? Points that deviate farther from the mean would seem particularly valuable for this dataset, unless there is specific justification for their removal. Perhaps there is a reason for this data treatment that just needs to be better explained?
This point was also raised by reviewer 2. The only reason to apply this filtering was to ensure that our analysis is insensitive to potential outliers, without making statements about the reliability of the removed data points. One could therefore view the variability in N. pachyderma stable isotope ratios that we use as a minimum, rendering our estimate of the magnitude of the excess variability conservative. We will make this reasoning clearer, both in the method section and in the discussion (see e.g. our suggested change above).

146: The assumption of chamber formation over one day in pachyderma is a bit misleading. While initial chamber formation may occur over one day (as in the referenced studies), calcification is likely more prolonged in this species. A better model than the spinose foraminifera observed in the Spindler and Be papers, might be congener N. dutertrei, where laboratory labelling experiments affirm that much of the calcite is added over a period of several days and nights as evidenced by banding and the apparently continuous uptakes of 'spikes' added in culture (see Fehrenbacher et al., 2017).
We agree with the reviewer that our modelled chamber formation is an absolute minimum. It is, however, in agreement with the data from Spindler (1996) on *N. pachyderma*. We nevertheless checked what the effect is of longer chamber formation and reran our simulations with a four day duration of chamber formation as suggested for *N. dutertrei* (Fehrenbacher et al., 2017). The effect is negligible because of the high temporal autocorrelation in the d18Oequilibrium time series that renders the effect of smoothing insignificant. The expected standard deviation of foraminifera d18O based on our model is in both cases 0.08 permille. (Note that in our original submission we modelled chamber formation within at most one day and that yielded an expected standard deviation of 0.09 permille.) We will add the following text to section 2.3 to clarify this issue: "The assumed duration of chamber formation is based on culture studies (Bé et al., 1979; Spindler, 1996). However, culture studies in the closely related species N. dutertrei have shown that chamber formation may take up to four days (Fehrenbacher et al. 2017). Longer chamber formation could in theory reduce the variability foraminifera δ18O because of increased smoothing of the environmental signal. In practice this effect is however negligible because of strong temporal autocorrelation in the δ18Oequilibrium time series that renders the effect of smoothing of up to four days insignificant. Our approach thus yields an estimate of variability that is robust against the likely range of chamber formation duration."

281: I am struggling with this calculation, on which so much of the interpretation relevant to IFA rests. While this estimation accounts for the N term, it makes two assumptions. The first is that the sample mean would have been the same if IFA had been carried out rather that multiple pooled analyses – this is probably a reasonable assumption, if one has on minimal instrumental error and near identical calcite contribution from all shells. However, the other assumption is that the stable isotope value of an individual shell would be the same as the value of the pooled analyses. This is a less robust assumption, belied by even the conclusions of this paper. Individual shells would be expected to represent a greater range of values, and therefore overall greater deviation from the sample mean. I think the argument for calculating excess of theoretical IFA as such could benefit from a statement of these underlying assumptions.

The obvious rebuttal to the caveat(s) raised above is that these are necessary assumptions given the sample set and/or that once again the estimate of unexplained variance is highly conservative. This might be the case, but if so perhaps there is too much emphasis on the quantification of this speculative 0.19 per mill (and therefore 0.11 per mil) number as a noise threshold.

We appreciate the concerns by the reviewer and will better explain the way we performed the calculation. The reviewer is right about the first assumption that we assume an identical contribution to the total calcite mass for each shell (and hence identical mean values). We will state this more clearly. However, we do not make the second assumption. Instead, we explicitly derive the standard deviation among individual shells from the standard deviation of the pooled measurements, the former is - as the reviewer correctly notes - indeed larger (double in our case) than the latter. To clarify these issues we will change the sentence: "Since our measurements are based on groups of four shells, the standard deviation of individual shells is double (√4) the observed standard deviation." to: "Since our measurements are based on groups of four shells the observed standard deviation is an underestimate of the standard deviation among individual shells. Assuming that each shell in the group contributed equally to the total mass, the degree of underestimation of the standard deviation scales with the square root of the group size (Groeneveld et al. 2019). Thus we multiply the observed standard deviation by two (sqrt(4)) to obtain an estimate of the standard deviation of individual shells."

333-335: My reading of Livsey et al. (2020) is that lamellar and crust calcite were indistinguishable in d18O space

Good point, we accidentally mixed up Mg/Ca and d18O. This makes the likelihood that variable encrustation could explain the observed variability even smaller. We will delete the sentence and add: "However, the difference between crust and lamellar calcite δ18O of N. pachyderma intercepted in spring when the water column was well-mixed is not significant (Livsey et al. 2020). Variable encrustation can therefore not be the explanation for the excess δ18O variability observed during the isothermal conditions in spring."

Other minor points: I was curious about the lack of shell measurements here, as stable isotope values are well known to correlate with size, something that the authors discuss. I understand that this is a reanalysis and such measurements may no longer be available, but it is a point potentially worth addressing.

The reviewer rightly points out that size of individual shells would be an interesting parameter to have at our disposal. However, as the reviewer also correctly infers such

measurements are unfortunately not available. We would like to highlight though that we have analysed larger scale pattern in shell size and its influence on sedimentary stable isotope ratios in a previous paper (Jonkers et al., 2013).

References:

Fehrenbacher, J. S., Russell, A. D., Davis, C. V., Gagnon, A. C., Spero, H. J., Cliff, J. B., ... & Martin, P. (2017). Link between light-triggered Mg-banding and chamber formation in the planktic foraminifera Neogloboquadrina dutertrei. Nature communications, 8(1), 1-10.

Livsey, C. M., Kozdon, R., Bauch, D., Brummer, G. J. A., Jonkers, L., Orland, I., ... & Spero, H. J. (2020). Highâ resolution Mg/Ca and δ18O patterns in modern Neogloboquadrina pachyderma from the Fram Strait and Irminger Sea. Paleoceanography and Paleoclimatology, 35(9), e2020PA003969.

References

Fehrenbacher, J. S., Russell, A. D., Davis, C. V., Gagnon, A. C., Spero, H. J., Cliff, J. B., Zhu, Z., and Martin, P.: Link between light-triggered Mg-banding and chamber formation in the planktic foraminifera *Neogloboquadrina dutertrei*, Nature communications, 8, 15441, 2017.

Jonkers, L., van Heuven, S., Zahn, R., and Peeters, F. J. C.: Seasonal patterns of shell flux, $\delta^{18}O$ and $\delta^{13}C$ of small and large *N. pachyderma* (s) and *G. bulloides* in the subpolar North Atlantic, Paleoceanography, 28, 164–174, 2013.

Spindler, M.: On the salinity tolerance of the planktonic foraminifer *Neogloboquadrina pachyderma* from Antarctic sea ice, Proceedings of the NIPR Symposium on Polar Biology, 9, 85–91, 1996.