Thank you Constantijn for the detailed feedback and comments on the manuscript. We respond to all comments below and include a tracked-changes version of the manuscript showing the changes to the main text and figures that we have made in response. The review questions/comments are in blue and our response in black. Line number references refer to the tracked changes version of the manuscript. In the revised text we have also made some minor formatting, spelling and small text adjustments to improve readability.

Now that you show the details of the d18O-CO2 regression more clearly, and the improved comparison to CO2 records, the results of this study are near-identical to the work of Berends et al. (2021), which also included a statistical model for prediction. To ensure that this new work has novelty, please highlight in the main text whether this work extends or replicates Berends et al. (2021) and explain how the approaches may have been different, or not.

Thanks for this comment. We agree it is important to better clarify the differences between our approach and Berends et al., 2021. We make three main points in response and show further below the changes made to the main text:

1. Berends et al., 2021 uses an inverse forward modelling approach, which is quite different and more complex than our GLS approach. We add text to the introduction, as shown further below, which references Berends et al. (and the preceding van de Wal 2011 and Stap et al., 2016 studies) and gives more context on the different assumptions in the inverse modelling approach compared to our GLS approach.

2. Thanks for pointing out that Berends et al also included a statistical (ordinary least squares regression) prediction of CO_2 from the benthic stack with $r^2 = 0.7$ as a point of comparison to the main result in that paper which is the inverse model. To be fair, the statistical model is a minor part of their paper and is not plotted in the paper nor provided in the data supplement. Our statistical model uses generalised least squares (GLS), rather than ordinary least-squares (OLS). While both models assume Linearity between variables, GLS allows heteroskedasticity in the variances and autocorrelation in the error terms. We used Autocorrelation Function (ACF), and Partial Autocorrelation Functions (pACF) and determined autocorrelation was present between observations and therefore that OLS is not reliable for parameter estimation. In the main text we expand on the GLS and AR(1) methodology used on our paper.

3. Our GLS reconstruction is compared to proxy data not available at the time of the Berends et al. study. Including recent reconstructions from boron data (Guillermic et al., 2022) and leaf wax $\delta^{13}C$ (Yamamoto et al., 2022). We also include comparison of our modelled CO_2 record to two sets of blue ice CO_2 data (Higgins et al., 2015; Yan et al., 2022), whereas the Berends et al. study does not make any comparisons with blue ice CO_2. The blue ice comparisons are particularly important as they are CO_2 measurements rather than proxy reconstructions and we calculate and compare mean glacial and interglacial concentrations between our predictions and blue ice records (Fig 3b, c.), which is a novel way of handling the dating uncertainty in the blue ice data.

Changes to the main text in response to this comment:

In the Introduction we add text to clarify the difference between our statistical GLS approach and the inverse modelling approach in Berends et al. 2021. New text in italics from Line 149:

> .. we make the simple assumption that the relationships between the LR04 benthic δ^{18}O stack and CO_2 can be extended beyond 800 kya and use generalised least squares (GLS) regression modelling between benthic δ^{18}O and CO_2 to make a prediction of CO_2 spanning 800–1500 kya. The deliberately simple implicit assumption, and null hypothesis, is that there is no change to the feedback processes linking benthic δ^{18}O and CO_2 before and after the MPT.
>
> *[Our] approach differs to previous more complex model studies that have attempted to reconstruct CO_2 using the LR04 benthic δ^{18}O stack as an input variable (van de Wal, 2011; Stap et al., 2016, Berends et al., 2021b). The latter studies use an inverse forward modelling approach, in which climate and ice sheet models of various complexities are used to capture physical relations between CO_2, global temperature and ice volume. For example, in Berends et al., 2021b the offset between modelled and observed benthic δ^{18}O is used to calculate a value for atmospheric CO_2 that is iterated back to the inverse model. The CO_2 record which minimises the difference between the modelled and observed benthic stack is then taken as an estimate of how atmospheric CO_2 may have evolved to force coupled climate, deep ocean temperature and land ice volume changes that reproduce the observed benthic δ^{18}O signal. Accuracy of the reconstructions in the inverse modelling approach depends on the ability of the climate and ice sheet models used to capture the correct climate dynamics across the MPT. Our GLS method is a simpler statistical approach, designed with the specific null hypothesis in mind, that does not attempt to simulate the physics linking benthic δ18O signal, land ice volume, global temperature and CO_2. A range of approaches to reconstructing CO_2 have been called for and are of value in the context of forthcoming continuous ice core records across the MPT from oldest ice projects currently underway in Antarctica [IPICS 2020].*

We also add the Berends et al., 2021 inverse model reconstruction of CO_2 to our Fig. 4b. And add some text to the Discussion comparing the results, which despite the different approaches do lead to quite similar reconstructions.

We also note the inclusion of the OLS approach within Berends et al., 2021 around Line 308:

> *"This is similar to that reported in ordinary Linear least-squares regression (r^2=0.70) by Berends et al. (2021b)".*

*Why did you use the LR04 benthic stack rather than the more recent stack by Ahn et al. (2017)? (**Line 11**)*

Differences between Ahn et al., (2017) and the LR04 benthic stack are small for the past 1.5 Myr. We decided to use LR04 because and for consistency with previous model studies which use that version of the benthic stack as an input variable in reconstructing $CO_{2,}$ e.g. van de Wal et al. (2011), Stap et al. (2016), Berends et al. (2021b).

Lines 73-74: "Emergence of … after the MPT (Fig. 1A)" While the spectral power in Fig. 1 does indeed show a peak around 100 kyr, which coincides with the eccentricity cycle, this kind of analysis is problematic. The "skipped obliquity cycle hypothesis" posits that the post-MPT glacial cycles are mostly alternating 80/120-kyr. Such a signal shows up in a spectral analysis as a single 100-kyr peak, not as separate 80 and 120-kyr peaks (try it out if you like!), so that this result by itself is not conclusive. Please discuss this.

Good point. We add the following text to the main text at Line 60:

> *Indeed, the skipped obliquity cycle hypothesis, proposes that the 100 kyr signal seen in spectral analysis of the post-MPT global benthic $\delta^{18}O$ stack (e.g. Fig 1A) may be comprised of alternating 80 and 120-kyr signals, i.e. in which the intervening obliquity cycles are skipped.*

And further down (e.g. Line 76) we refer to the '100 kyr signal' rather than 'eccentricity signal'.

Lines 113-115: please add some additional text regarding the combined record of bottom water temperature and global ice volume. For the 0-800 ka time interval the proxy record of Elderfield et al. (2012) showed that the shape of the glacial-interglacial changes are quite different between these two components, which may also impact your premise that Southern Ocean is an important link. It's not a Linear relationship between ice volume and ocean temperature.

We add additional text to the main text as below (in italics) in response at Line 118, we clearly note the combined signals and the non-Linear relationship.

> The $\delta^{18}O$ ratios in the LR04 benthic stack are governed primarily by deep ocean temperature and global ice volume at the time the foraminifera lived, with higher values indicating both increased ice volume and a colder climate. *The relationship between the ice volume and ocean temperature components contributing to the $\delta^{18}O$ benthic stack are not Linear. Separating the two signals remains challenging and has been attempted elsewhere using a range of approaches from comparison with paired deep ocean temperature proxies (Elderfield et al., 2012), inverse modelling (Berends et al., 2021b) and spectral analysis (e.g. Huybers and Wunsch, 2009).*
>
> As explained in the main text (see from Line 141), our paper does not attempt a quantitative separation and attribution of the processes linking global ice volume, ocean temperature and atmospheric CO_2. In this way our statistical approach is different to inverse modelling or other modelling approaches that attempt to represent the complete physics. We note clearly that our approach limits us to testing a simple implicit null hypothesis through comparison with existing and forthcoming data.

Line 158 presents the reconstruction by van de Wal et al. (2011) but the more recent version by Berends et al. (2021) should be presented since it has updated ice sheet and climate models.

We have added the Berends et al., (2021b) inverse model CO_2 reconstruction to Fig. 4B and some text describing the model in the Discussion (ca. Line 370-376) as follows:

> *Our simple GLS model demonstrates a similar long-term trend and timing of glacial-interglacial signals and an atmospheric CO_2 level that sits approximately mid-way between*

the van de Wal et al. (2011), and Willeit et al. (2019) models and is remarkably similar to the Berends et al., 2021b reconstruction despite their different approaches. Notably the Berends et al. reconstruction shows greater glacial to interglacial amplitude in the CO_2 signal compared to our GLS-model. The decreasing Linear trend in CO2 in Willeit et al. (2019), which is not seen in the other reconstructions, was directly prescribed in that study to induce Northern Hemisphere glaciation at 2.6 Myr ago.

Please refer also to the abovementioned additions to the Introduction and Discussion citing Berends et al. 2021b and the inverse model approach.

***Line** 171: there needs to be some additional text to explain what this "autoregressive factor" represents.*

We revised the text around Line 189 to clarify our use of both the GLS technique and the AR(1) factor:

We use a generalised least squares (GLS) model with an auto-regressive (AR) factor 1 to predict atmospheric CO_2 from the LR04 benthic $\delta^{18}O$ stack (Fig. 3A and B). We use GLS because the assumptions of ordinary least squares (OLS) are violated by the presence of autocorrelation and heteroskedasticity in the regression errors. We selected the AR(1) correlation factor as it yielded the lowest Akaike information criterion (AIC) value from a test of multiple correlation factors. The AR(1) process assumes and accounts for dependence of error at a given point in time on the previous error term. In practise this makes the model assumptions more realistic and improves parameter estimation where, as in the climate system, observations are dependent on past values.

Line 190: "after removing 50% of data". Please provide additional explanation on how this works, and what the results are. If your data is significantly auto-correlated (which with a 3-kyr time step I think it is) then I'd think it's not very surprising that removing some of the data points (which essentially don't contain any new information) doesn't affect the least-squares fit.

Each iteration of the model removes a different, random 50% of the data. Doing this with replacement introduces variability among the bootstrap samples. Since each sample can have different combinations of the original data points (including repeated ones), this variability helps in assessing the robustness and stability of the model. We clarify the text around Line 205. We don't expect the bootstrap method to address auto-correlation but is an accepted method to gauge sensitivity to data and dating uncertainty.

To gauge the GLS model stability we took a bootstrap approach, selecting a random 50% subset of our data (with replacement) and re-running the model 1000 times to determine 95% confidence intervals for the predictions. While the GLS method itself addresses autocorrelation, the bootstrap method introduces variability such that each iteration of the model has different combinations of the original data points (including repeated ones), this variability helps in assessing the robustness and sensitivity of the model, e.g. to variable data and dating uncertainty.

Lines 261-262: "In the following section … interval" Why do you compare to some data in the Results section, and then to some other data in the Discussion section?

We decided to split into results and discussion accordingly:

> Results: Comparison of our CO_2 record to those data that contain measured atmospheric concentrations of CO_2, specifically blue ice data from Higgins et al. and Yan et al.
>
> Discussion: *Proxy (e.g. boron and leaf wax)* and model reconstructions of atmospheric CO2.

This separation is to distinguish between ice core data, which is an actual measurement of past atmospheric concentration, versus proxy or model data, which are reconstructions that are based on a range of assumptions. Once can see by the spread between overlapping proxy and model data series that the records are inconsistent with each other and therefore the proxy and model data cannot alone be used to refute our predictions (we note a similar point is made in Berends et al., 2021b). So we think these comparisons to proxy and model are suited to the Discussion. A weakness of the blue ice data is the dating uncertainty, which we address quantitatively in the Results by calculation of mean glacial and interglacial ranges for the overlapping GLS predictions and blue ice measurements.

We add a Line to the main text to make this approach more clear to the reader (Line 179):

> *In the Discussion, we also compare our predicted record to existing proxy-CO_2 reconstructions…*

Line 325: this statement is not supported by a citation, please include reference to the evidence that boron isotopes over-estimate CO2.

We have modified this statement, it no longer refers specifically to δ11B-based reconstructions specifically over-estimating CO_2 concentration. Instead we refer to the large spread in existing proxy-CO_2 reconstructions, as follows (Line 358).

> "*The strong spread between these different proxies and the large associated uncertainty of the alternative marine and leaf wax proxy-CO_2 reconstructions mean that we do not find cause from the existing CO_2 proxy data to reject our predictions nor our associated null-hypothesis."*

For information though, we do see evidence for overestimation by $\delta^{11}B$-based estimates in data from Chalk et al., 2019 overlapping the observed continuous ice core record as shown below. We do not include this in the main text since on reflection, we do not have evidence that this overestimation is systematic for pre-MPT data or other $\delta^{11}B$-based reconstructions.

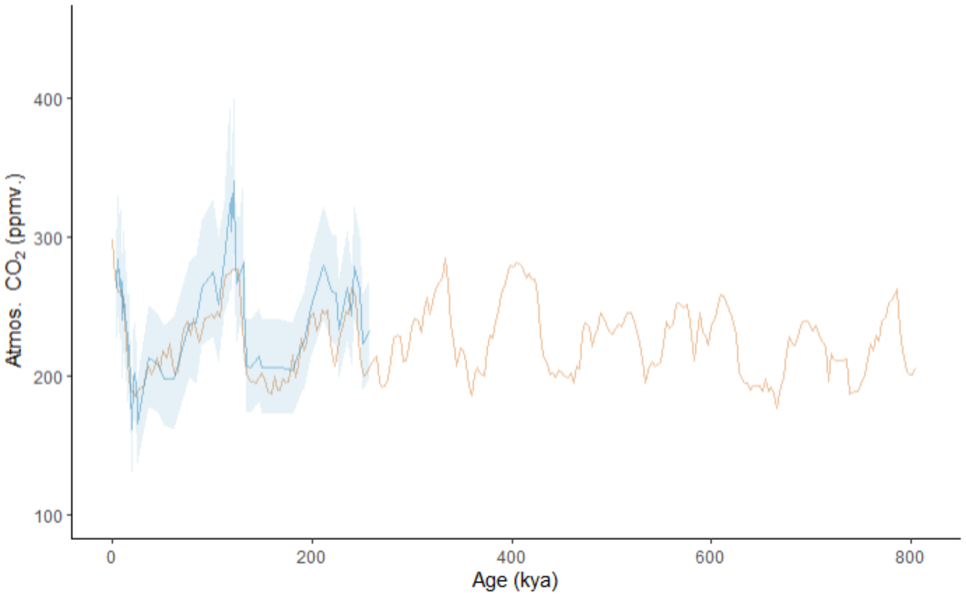

Fig R1. The continuous ice core composite atmospheric CO_2 record in orange (Bereiter at al., 2015) and reconstructed atmospheric CO_2 from boron isotopes for the interval 0–250k in blue (Chalk et al., 2017).

Lines 350-351: clarify here whether you are referring to the whole time interval shown on your figures (0-1.5 Ma), or a narrower period? Figure 3 seems to suggest that glacial CO2 was fairly constant >1 Ma, then declined through ~1000 to ~650 ka. The text currently doesn't distinguish between these scenarios, which have quite different implications for climate forcing.

Agreed, this is an important point to clarify. Here we refer to figure 3b, we see a drop in glacial CO2 across the MPT when comparing the regions before and after the grey shaded region (representing the MPT as defined by Chalk et al. (2017)).

We adjust the main text accordingly at Line 396:

> *That fact that our LR04-based prediction of CO_2 captures this same trend, with predicted glacial CO_2 fairly constant from 1.5 to ca. 1.0 Mya before declining from 1.0 to 0.6 kya, reflects that the LR04 benthic stack features an increase in the interglacial to glacial benthic $\delta^{18}O$ difference across this same interval, which is dominated by the glacial stage changes (Fig 3A.).*