Review #1

de la Vega and co-authors present two beautiful new records on paleo-CO2 reconstructions that support the standing of the boron isotope proxy as a reliable indicator for paleo-CO2. The authors assess several aspects of the proxy, including dissolution and proxy calibration. The findings from these assessments are not entirely new, but that does not diminish the value of the manuscript. I have several recommendations for the authors to further improve the manuscript:

We thank the reviewer for their feedback and respond to each comment below.

The Dlog$_{10}$ CO_2/DpH approach to estimate paleo-CO_2 is an interesting one but the authors already indicate that it is only useful for short time scales, similar to the original study of Hain et al. 2018. This is somewhat disappointing for deep-time studies, where we really do not have a good sense of a second parameter of the carbon system. It would be helpful if the authors could discuss how this approach adds to data that we can already assess from ice cores over the past 800 kyr.

The advantage of this approach is that we compare δ^{11}B-derived pH and CO_2 forcing (derived from ice core CO_2 od the same age) and to show the near-linear relationship between the two variables. Beyond 800 ky, in the absence of ice core CO_2 data, we can use the reconstructed pH from boron isotopes and apply this linear relationship to reconstruct CO_2 forcing.

We note that here only pH is needed and not the full calculation of δ^{11}B-derived CO_2 that requires the assumption on a second carbonate parameter. As shown in equation (5), the calculated pH is dependent on $\delta^{11}B_{borate}$, α_B (both known) and $\delta^{11}B_{sw}$, that has to be reconstructed beyond the last 2-3 million years due to the residence time of boron in the ocean (10-20 Myrs). However, as detailed in Hain et al (2018), the uncertainty associated with δ^{11}Bsw has a minor effect on delta ΔpH (see for example their Figure 4).

Whilst the requirement for a short time scales (i.e. <400 kyr) is still needed (to remove long-term variations in ΔDIC and ΔlogHCO_3^-), this doesn't prevent the reconstruction of ΔFCO_2 for much of the geological record, even in the face of uncertainty in δ^{11}Bsw.

The data added here to the compilation of Hain et al (2018), allows for a richer dataset that allows to compare pH to ice core CO_2 and compare the updated linear relationship with the basic formalism $\Delta F/\Delta pH$ =-12.3 W/m^2 and CaCO3, DIC, and Temperature driven relationships.

If the ΔFCO_2 /pH relationship was constant for the last 800 kyrs we have no reason to believe it should change across other orbital cycles (e.g., during the Mid Pleistocene Transition). The increase in higher resolution δ^{11}B record over the Plio-Pleistocene will allow to evaluate CO_2 climate forcing with different climate background.

We will add in the introduction the following text to clarify the point about slope comparison: "Comparing the slope of the regressed ΔF/ΔpH line from our data to theoretical endmembers (temperature, DIC, CaCO3) could allow to decipher which mechanisms is primarily at play during Glacial-Interglacial (G-IG) cycles. Drivers during these cycles are known to be a combination of changes in water mass dynamics, the soft tissue pump (nutrient availability and consumption), sea-ice cover and the $CaCO_3$ counter pump, with a major role of the Southern Ocean (see Sigman et al., 2010; Hain et al., 2010, 2014 for a full review). Providing the δ^{11}B accuracy is well constrained and the δ^{11}B uncertainty optimised, the change in ΔF/ΔpH slope over time should theoretically allow to know which mechanism was prevalent".

Site selection: The authors keep using the same sediment records that they have been using for many years but Figure 1 demonstrates that neither site is ideal. This raises questions about the utility of using either site to "calibrate" the proxy and the reconstructed CO2 shown in Fig. 3 shows more extreme deviations from glacial and interglacial CO2 extremes in the ice cores than site 999. An ideal site would be located in the vast ocean areas that are shaded green in Fig. 1. The final reconstructions are still impressive, but require corrections for a CO2 disequilibrium of which we cannot be certain that it remained constant through time. This caveat should be considered throughout the manuscript, as a source of uncertainty that affects all aspects of the study, including the calibration.

We acknowledge that we have focused much of our work at ODP 999, but this is the first study to produce $\delta^{11}B$ data for ODP 871. The reason we chose these sites is that they have sufficient *G. ruber senso stricto* for our purposes (1-2 mg), they have a priori determined age models based on $\delta^{18}O$, have sufficiently high sedimentation rates to allow for the high-resolution sampling we carry out here, they are relatively shallow (<2800m) and preservation (although explored in detail in our manuscript) has been previously described as good or better. We agree with the reviewer that a site in the ocean gyres would be ideal in terms of CO_2 disequilibria but it is important to note that such sites would be less ideal in other ways. Most importantly for this study, the sites located beneath the gyres tend to either be outside of the range of *G. ruber* ss geographical distribution and/or have very low sedimentation rates. Furthermore, the gyres tend to be situated above the deepest parts of the ocean such that the carbonate sediments below them are typically below the regional lysocline/CCD. Sites ODP 999 and ODP 871 are therefore a necessary compromise with respect to their disequilibria. Importantly both sites are far from upwelling locations today, and are oligotrophic with CO_2 disequilibria less than 20 ppm. Our assumption of constant levels of disequilibria (as is commonly done) is, however of fundamental importance as the reviewer notes. As a result, we discussed this specific point in section 4.4. where we provide recommendations for future studies. We have also attempted to address this point directly in section 4.2.2 by using records of surface $\delta^{13}C$ and $\delta^{18}O$ in *G. ruber*. These two variables should be impacted if upwelling is happening at either core sites (lighter $\delta^{13}C$ and heavier $\delta^{18}O$, for colder nutrient rich waters) . In the absence of any visible deviation from the expected change during G-IG variations, we infer that upwelling had a minimal impact on our sites over the study interval. We however acknowledge that this interpretation could be further constrained by having records of productivity at each site (e.g., opal fluxes, alkenone concentration, Ba excess). In the absence of such records, we infer our conclusions based on planktonic $\delta^{13}C$ and $\delta^{18}O$ only. Also see reply below (about SST in Figure S9) about upwelling at site 871.

In order to better reflect the uncertainty in disequilibrium, in the revised manuscript we will fully propagate the uncertainty in this term using a conservative +/- 10 ppm (1SD) uncertainty on the disequilibrium correction. We will also refer to model outputs (Gray and Evans, 2019, suggested by reviewer 2) that shows that relative delta-pH does not change very much between the LGM and the pre-industrial at these two core sites.

Stable isotope record: The authors used only 10 planktic foraminifer shells for each sample in this record, which is a small number given the geochemical variability from shell to shell and bioturbation. Even laser ablation studies in laboratory culture, where specimens experience well controlled, constant environmental conditions, use at least 12-25 shells to overcome interspecimen variability (see e.g., Holland et al. 2020). It would have been better if the authors had picked larger samples for boron isotope analyses, crushed and homogenized them and then taken a small split for stable isotope analyses. While it would be asked too much to replicate the record with a larger shell number per sample at this time, the authors should mention that this sample size is not ideal, so that other researchers do not use it as a guideline. This also reflects on the genotype comparison (Fig. S5), which might have shown more significant results with a larger, more suitable sample size.

We appreciate the reviewer concerns on this point. It is normal procedure in our lab to measure a mixture from the ~200 or so foraminifera specimens to measure for oxygen isotopes alongside the boron isotopes". However in this case (where we ran $\delta^{18}O$ after $\delta^{11}B$) we decided to concentrate on precise and well defined morpho-types and thus limited the samples to 10 individuals of each. We recognise that this will lead to more variability in the record, but this is exactly what we are looking for to identify diagenetic alteration or preservation bias, hence the noisier record may actually be more informative in this instance. This will now be made clear in the text.

Age models: The authors generate new data and display them in Fig. 2E,F but do not really discuss them. The figure caption describes a species correction but it is not clear how that correction has been determined and if it has already been applied to the displayed data or was applied afterwards. LR04 provides no guideline on this, as far as I can tell. Figure 2 shows site 999 data are too low compared to LR04 but site 871 data fall on LR04. This means that at least one of these records deviates from LR04, and the cause for this deviation (and the choice of species offset) should be discussed.

Given the benthic foraminiferal isotope records are only used for age correlation purposes (to provide a constraint independent from planktic $\delta^{18}O$ and $\delta^{11}B$ records) the precise offset used is unimportant, as long as it is consistent. Of course we recognise the importance of stating the correction applied so that others can go on to reuse and reinterpret the data. A correction of +0.47 was applied to *Cibicidoides wuellestorfi* at site 999 following Marchitto et al. (2014). The ODP 871 samples are measured on *Uvigerina peregrina* and thus required no correction. This is now clearly stated in the figure caption and the supplementary tables and we thank the reviewer for bring this important point forward.

Dissolution experiment: This is the weakest part of the entire manuscript. The authors do not describe which sediment samples they used for the experiment. What makes those samples ideal for such an experiment? The lack of shell weight data is detrimental and essentially prevents any confidence in the data that have been collected. How large was the volume of acidified fluid? Is it possible that it got saturated right away, was there any dissolution at all? Here again the authors should say more clearly that their experiment falls short on several fronts. They do, but still interpret the data, which does not seem justified. Dissolution in acid and deionized water is likely very different from dissolution in corrosive seawater, so there really is little value in the experiment and associated data. The discussion draws mostly from earlier, much better dissolution studies, which serve the authors' purpose well, so the dissolution experiment could just as well be removed from this study without any impact on the discussion or results. The concern about including such an experiment is that it may lead others to follow the example, which would be unfortunate.

The samples used are from two intervals of ODP site 871 (age 182 ky and 160.81 ky), and were chosen because of the high abundance of forams at that site and intervals that allowed multiple repeats of the same material of the same age. We very regretfully could not measure elemental data for *T. sacculifer* due to machine down time for a significant period of time, after samples were dissolved. However weight and elemental data were obtained for the repeat leaching experiment on *G. ruber*. No change in weight and elemental composition were observed after 10 hours of leaching (in 1ml of acidified fluid, like all other leaching experiments) for *G. ruber*. Whilst this shows that indeed, no dissolution has occurred for this species, it shows some relative robustness to dissolution compared to *T. sacculifer* that showed a significant $\delta^{11}B$ fractionation after only 6 hours of leaching. We however acknowledge weight and trace element data for *T. sacculifer* is a crucially needed addition. In the absence of these data, we agree with the reviewer and will remove this auxiliary data until further constraints can be obtained.

Temperature estimates: The authors use both pH-corrected and uncorrected calibrations to translate Mg/Ca to SST. In the end they use the uncorrected estimates for calculating CO2 but it is not discussed why this should be the better choice. This is particularly striking after multiple studies have highlighted the pH dependence of Mg/Ca in *G. ruber*. The authors should discuss why they think this is the better approach following line 461. This choice affects the downcore calibration for d11B and deserves more attention.

We thank the reviewer for their comment. This point was also raised by reviewer 2. We have conducted sensitivity tests with seven treatments of Mg/Ca or temperature: (1,2) Gray et al., 2018 with and without a pH correction, (3,4) Gray et al., 2018 using Mg/Ca corrected for depth-dependent dissolution, with and without a pH correction, (5,6) Anand et al., (2003) with and without a depth correction, and (7) constant temperature. The results are displayed bellow (Figure R1).

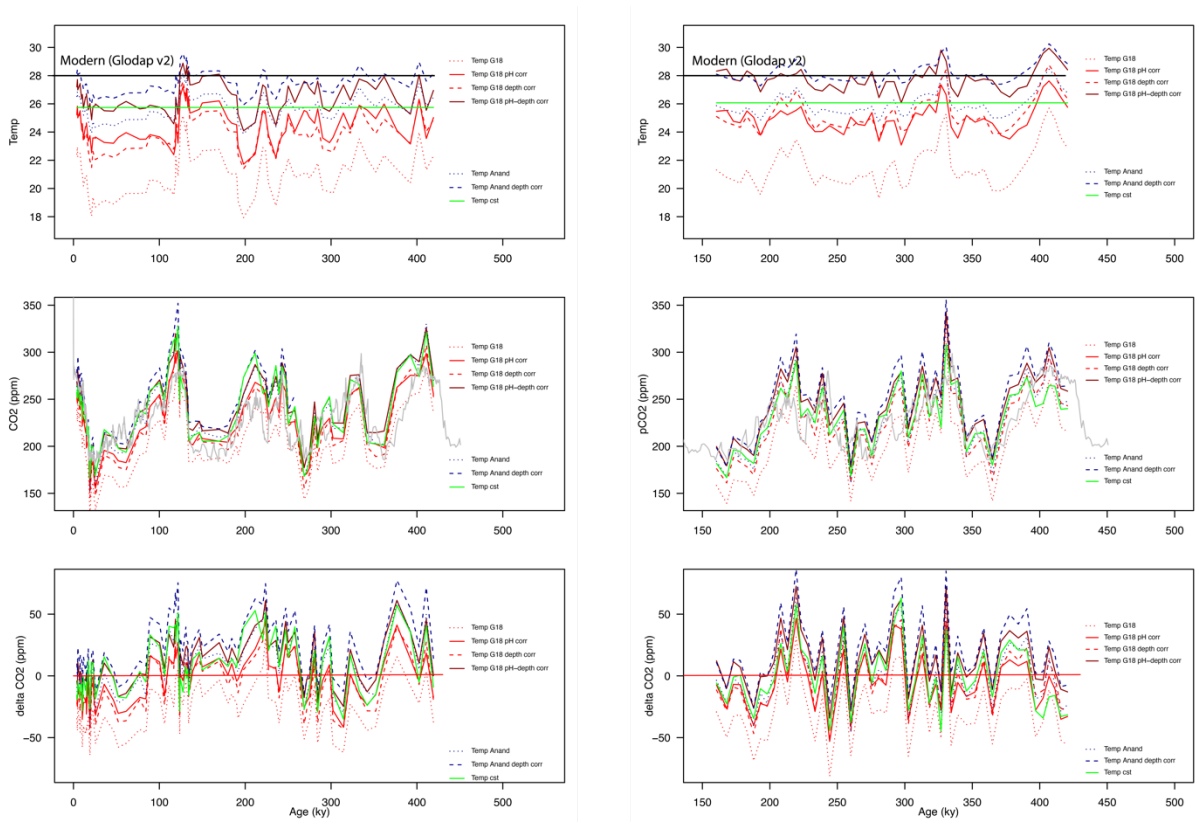


Figure R1. Effect of various temperature treatments (top), on $\delta^{11}B$-derived CO_2 (middle) and CO_2 offset to the ice cores (bottom). Left: site 999, right: site 871. Temperature treatments are: G18 (Gray et al., 2018 no correction), G18 pH corr (Gray et al., 2018 with pH correction), G18 depth corr (Gray et al., 2018 with Mg/Ca corrected depth), G18 pH-depth corr (Gray et al., 2018 with depth and pH correction), Anand (Anand et al., 2003, no correction), Anand depth corr (Anand et al., 2003, Mg/Ca corrected for depth), Temp cst (constant temperature of 26°C).

We have chosen the Mg/Ca treatments that accounts for pH effect on Mg/Ca and yields the closest agreement between coretop and modern temperature from Glodap v2 at both sites (Note coretop at 871 is not displayed and the most recent Mg/Ca from Dyez and Ravelo, 2013 was used). This treatment is with a pH correction (Gray et al., 2018) and Mg/Ca corrected for depth/dissolution.

Following suggestions from both reviewers we will update all data in the revised manuscript including $\delta^{11}B$-derived CO_2 to reflect this change in Mg/Ca treatment. The resulting average CO_2 offsets for each Mg/Ca treatment are displayed below (Table R1).

core	Mg/Ca treatment	DpCO2	2sd	core	Mg/Ca treatment	DpCO2	2sd	Average DpCO2 for both records
999	T Gray18	-27.03	40	871	T Gray18	-28.21	52	-27.62
999	T Gray18 pH corr	-3.87	39	871	T Gray18 pH corr	-6.40	47	-5.14
999	T Gray18 depth corr	-6.77	43	871	T Gray18 depth corr	-3.30	54	-5.03
999	T Gray18 pH corr, depth corr	12.06	41	871	T Gray18 pH corr, depth corr	13.89	51	12.98
999	T Anand03	6.64	45	871	T Anand03	1.37	54	4.00
999	T Anand03 depth corr	23.28	48	871	T Anand03 depth corr	20.88	57	22.08
999	T constant	6.33	46	871	T constant	0.23	53	3.28

Table R1. Effect of various Mg/Ca-derived temperature calibrations on CO2 offset (DpCO2). The line in green is the chosen updated calibration.

-Figure S9 also discusses "anomalous" temperature estimates but no discussion is provided as to what constitutes such an anomalous deviation. What is the point of reference? How does the SST record compare to sites of similar latitude but outside of potential upwelling areas? (e.g., gyre sites). How do we know that SST was not cooler than expected?

We thank the reviewer for this relevant comment. A comparison between site 871 SST with an Mg/Ca-derived SST record from the Western Pacific warm pool site MD97-2140 (de Garidel-Thoron et al., 2005, figure R2) a location outside of the upwelling from the Pacific cold tongue, shows that the periods of high CO_2 offset (230, 290, 380 ky) are not associated with relatively cold periods at ODP site 871, which suggests that they are not related to upwelling (since upwelled water should be cold and have high CO_2). We note that this comparison needs to be caveated by the different treatments of Mg/Ca in de Garidel-Thoron et al. (2005). Site 871 was also chosen as it is a previously studied site (Dyez and Ravelo, 2013, 2014). These studies compared temperature records of the on-equatorial ODP site 806 and the off-equatorial site 871, and concluded site 871 was unlikely to be impacted by upwelling due to its deep thermocline and the stronger cooling observed at 806 possibly linked to upwelled thermocline waters at this site forced by Northern hemisphere summer insolation. Similar to the reviewer's previous comments about disequilibrium, we also acknowledge that this interpretation could be further constrained by having records of productivity at each site (e.g., opal fluxes, alkenone concentration, Ba excess).

Upwelling at site 999 is thought to happen seasonally in the modern with associated CO_2 flux of +20 ppm (Olsen et al., 2004) that we correct for in downcore CO_2 reconstruction. Changes in upwelling at Site 999 may have occurred in the past in relation to the position of the ITCZ (see discussion in Foster, 2008). Foster and Sexton (2014) have also reconstructed CO_2 zonally across the equatorial Atlantic and the Caribbean and show that for the last 30 ky at least, Site 999 has remained in equilibrium with the atmosphere. Whilst SST is a first order constraint on upwelling, future studies need to focus on paired proxies of temperature and productivity to evaluate change in local CO_2 fluxes. As mention above in the reviewer's comment about disequilibrium. we will fully propagate the uncertainty in this term using a conservative +/- 10 ppm (1SD) uncertainty on the disequilibrium correction.

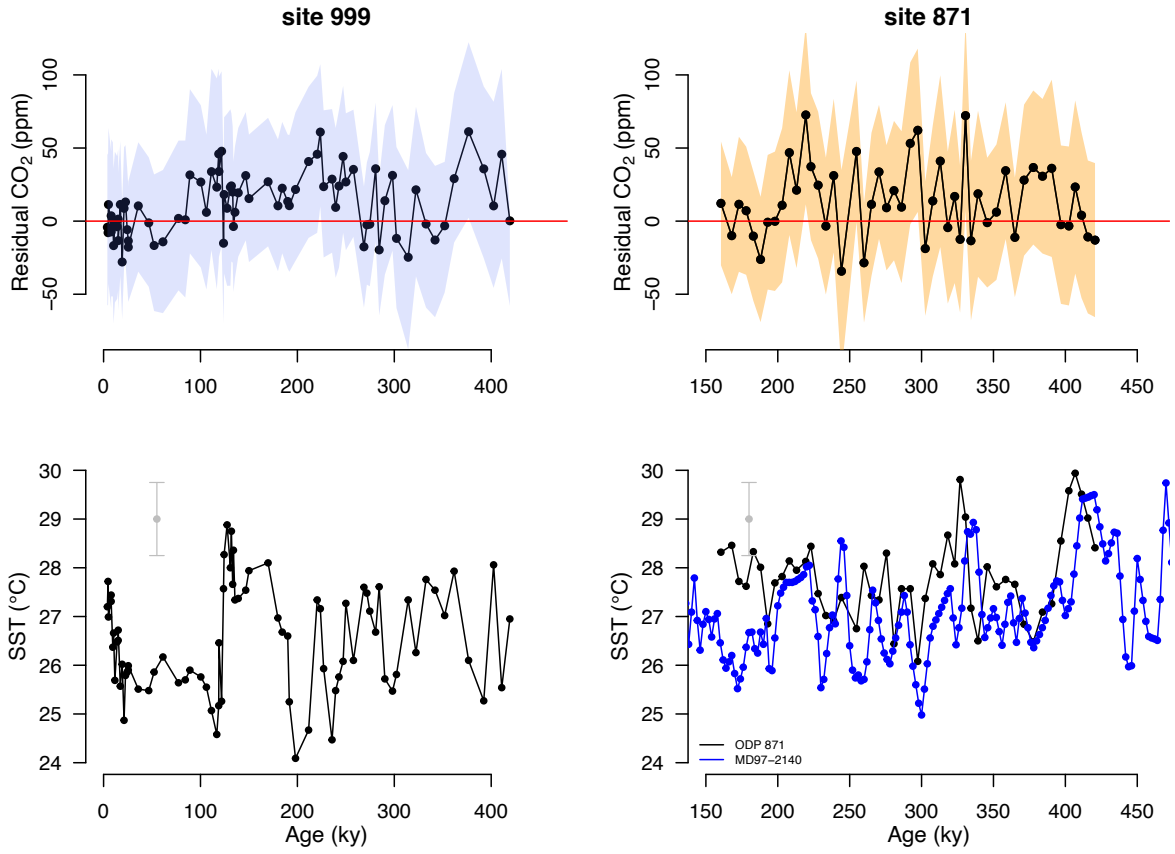


Figure R2. CO_2 offsets (or residual) and updated SST (pH and depth correction on Mg/Ca), and comparison with SST record from the Western Pacific warm pool site MD97-2140 (De Garidel-Thoron et al., 2005)

-CO2 forcing: Figure 4 shows a nice correlation between DFCO2 and pH but the data deviate at least 5-6 times from the regression lines and their uncertainty, if that is displayed by the grey shading, it does not capture the true data variability. Based on the scatter around the lines, how many d11B data do the authors suggest are needed to provide a single reasonable estimate of DFCO2 for a given point in time? Fig. S7 suggests an uncertainty of +/- 0.3-0.8 W/m2, which is clearly an underestimate given the data scatter and requires an assessment of the number of data needed to provide such a minimal uncertainty.

The key variable to look at is the goodness of fit MSWD. The higher the uncertainty in the data, the better the MSWD is (i.e. close to 1) since points furthest away are then more consistent with the regressed line. As demonstrated in Figure S7 and updated below (Figure R3) with new temperature estimates, a lower uncertainty assigned to pH gives a poorer MSWD: the data is over dispersed, given their low uncertainty. This variable is more informative about the data dispersion than the envelope around the York regression. The number of $\delta^{11}B$ data points helps to assess how accurate the basic formalism ($\Delta F/\Delta pH = -12.3$ W/m^2) is. Providing a minimum number of points to estimate ΔFCO_2 would be rather arbitrary. With the current data set, the average ΔFCO_2 deviation from the best fit line is -0.13 W/m^2. The main idea of the approach is (1) to validate the basic formalism and (2) to evaluate where the regressed slope derived from $\delta^{11}B$-pH data falls relative to the temperature, $CaCO_3$ and DIC driven slopes. The richer the dataset the more accurate the regressed slope is. Improvement in high resolution records as well as analytical uncertainties of $\delta^{11}B$ (which accounts for the main uncertainty in pH), will refine the accuracy of the data-derived ΔF-ΔpH relationship and how much it deviates from the basic formalism. The original study of Hain et al (2018) reconstructed ΔFCO_2 with an uncertainty empirically determined by the ΔFCO_2 range between the lowest and highest endmember slope of the ΔpH-to-ΔF relationship (i.e. $CaCO_3$ and DIC change), this accounts for both uncertainty in $\delta^{11}B$-derived pH as well as the conversion from ΔpH to ΔF. With the benefit of more data and a more thorough consideration of the pH reconstruction uncertainty we empirically determine the pH/logCO_2 relationship to be closer to the steep DIC endmember. We think it is a conservative approach to consider the full endmember range, which allows to confidently reconstruct ΔF in the past, but our data suggests this range can be reduced with further work.

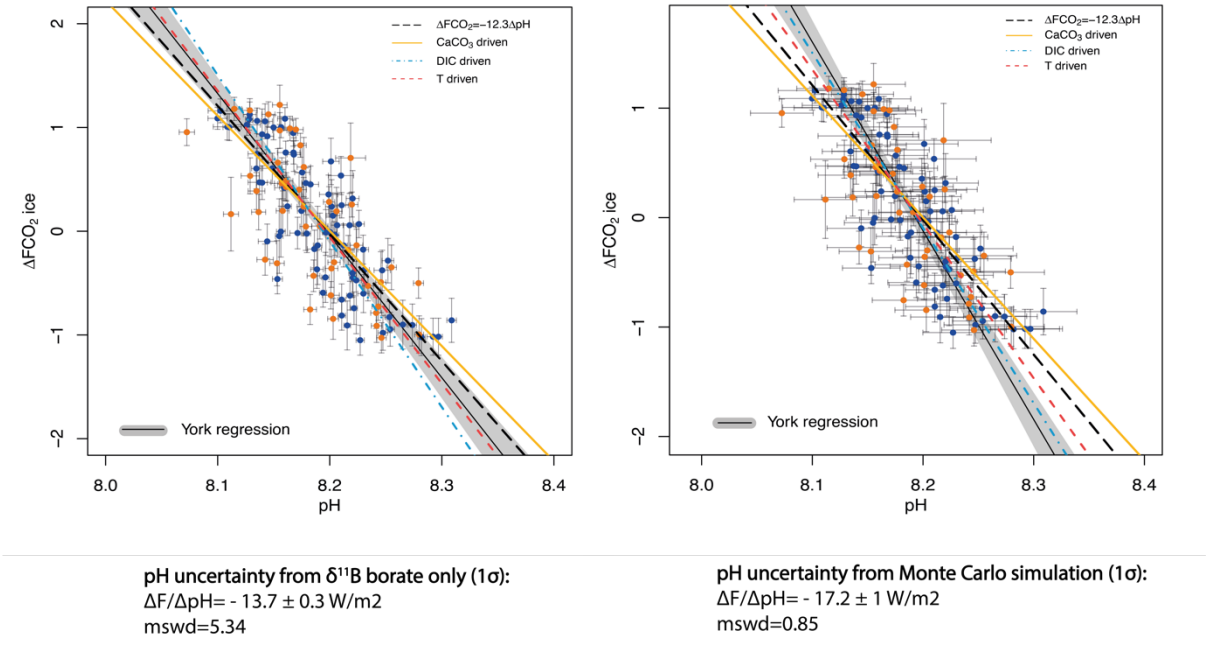


Figure R3. Updated ΔFCO_2-ΔpH relationship when using Mg/Ca pH and depth corrected.

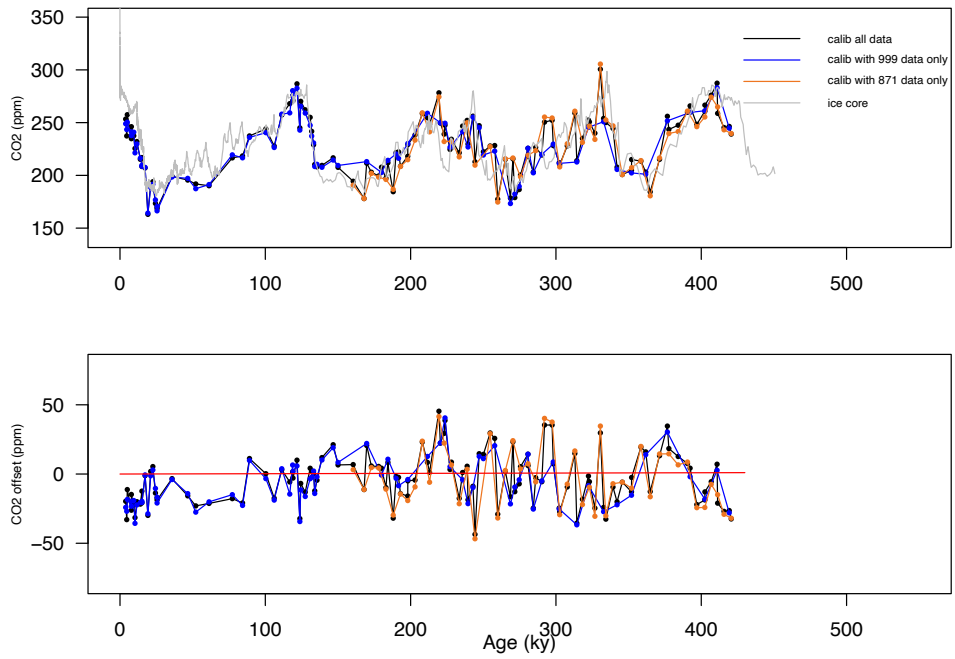
-Downcore calibration: This is an interesting approach that could be applied to species that have not yet been calibrated in culture, and it is an approach that could rival coretop calibrations because the modern pH range in

surface seawaters is generally too small to allow for a high-quality calibration. However, here again I wonder how many cores and data should be included in a calibration exercise, and whether the downcore calibration is really stronger than the existing culture calibration for *G. ruber*. To do so, I would recommend that the authors generate a new calibration for each core site and then apply that calibration to the respective other core site. How different are the calibrations from each other and from the culture calibration, and do both calibrations improve the match of the CO2 estimates to the ice cores?

We thank the reviewer for their comment and while we agree that this is a useful approach we did not want to give the impression this is the best way to deal with $\delta^{11}B$ data going forward. As section 4.3 demonstrates there is considerable power in having an independent record of ocean pH and boron is unique among the CO_2 proxies in that it is not tied to the ice core CO_2 records in any way. We agree that for species without a culture or core-top calibration this approach would be useful, we have used the combined record here principally for the sake of illustration.

However, since the two records have CO_2 offset that occur at similar periods (Figure S3), it is unlikely that a downcore calibration conducted on each site greatly differs from the combined records. To showcase this, we include here the same exercise performed on each separate record and display all the slopes and derived CO_2 offset (figure R4 and table R2 below). These results show that each separate calibration do not show significant deviation in calculated CO_2 (the average CO_2 offset for 871 and 999 calculated with their respective downcore calibration is -5 ppm) from the calibration obtained from the combined two record (+4 ppm average offset).

We agree that this approach is dependent on a given record, since cores from different locations can differ from one another by having different oceanographic setting or dissolution history, and we discuss this in section 4.4.



Downcore calibration	Slope	Intercept	Average CO2 offset (ppm)
All data	0.713	6.492	4
999 only	0.724	6.326	-7
871 only	0.710	6.532	-3

Figure R4. $\delta^{11}B$-derived CO_2 and resulting CO_2 offset calculated using optimised calibration using combined data set from both cores and each separate data set from ODP 999 and 871.

Table R2. Slope, intercept and average CO2 offset for optimised calibration using combined data set from both cores and each separate data set from ODP 999 and 871.

-Finally, the authors should check spelling and grammar throughout, including names of authors whose work they cite. There are several typos throughout the manuscript, and in some cases incomplete sentences. Please check spelling in lines 42, 45, 74/75, 149, 150, 186, 215, 413, 433, 526, 570, 691, 693. The sentences in line 607-610 should be rephrased entirely.

This typos have been updated and the lines 607-610 rephrased as follow (including updated numbers):

"Intervals of high fragments occur 5% and 33% of the time, at site 999 and 871, respectively, during positive CO_2 offsets (and 59 and 77% of the time during negative or no offset to the ice cores)."

-In summary, this study adds valuable confirmation to an already strong proxy. There is still room for improvement in this manuscript, mostly by clarifying certain choices, but also by assessing the paleo-calibration from different angles.

We thank the reviewer for their thorough constructive feedback.