Dear reviewers, dear editor,

We are grateful for the detailed suggestions provided by both anonymous reviewers, and believe that incorporating these suggestions will considerably improve our study. Below, we address the comments provided by each reviewer individually. Reviewer comments are included in italics.

On behalf of all co-authors,

Cordially,

Jai Chowdhry Beeman

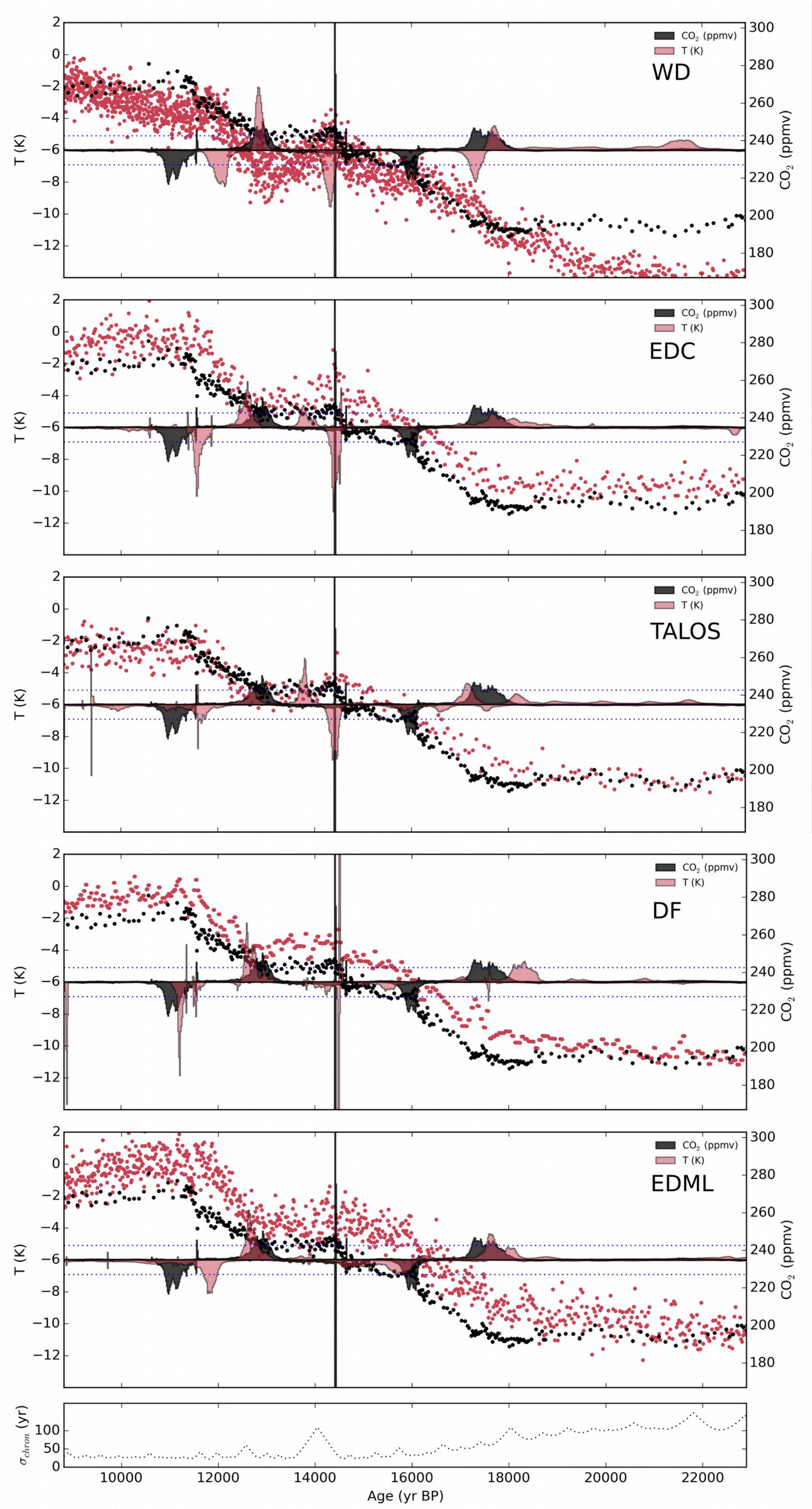
Reviewer 1.

Summary: Beeman et al. investigate the phase relationship between atmospheric CO2 concentration and Antarctic temperature during the last deglaciation. This question has been investigated various times over the last decades as it can illuminate on the role of CO2 forcing and carbon cycle changes in the deglaciation. In comparison to earlier studies, Beeman et al. use a CO2 dataset (WAIS) that is better resolved in time, and better constrained in terms of its delta age uncertainty. In addition, they include new water isotope data from the WAIS divide ice core. Furthermore, they do not just estimate a mean lag between CO2 and temperature over the entire study period (e.g., Pedro et C1 al. 2012) but investigate the relative timing of change-points in both timeseries, similar to (Parrenin et al. 2013). Their results indicate that CO2 and temperature change synchronously except at the end of the Antarctic cold reversal (CO2 leads) and the onset of the Holocene (CO2 lags). The availability of the new high resolution CO2 data from WAIS and the approach to investigate change-points rather than a mean lag make this paper interesting and timely. The results will be a valuable contribution to the discussion of causes and effects of deglacial climate changes. However, I have some reservations about aspects of the manuscript and method that warrant further testing and/or justification.

General Comments: i) The authors use a stack of five Antarctic ice core records as their temperature estimate. Their argument is that this reduces local noise. While this may be true, it also removes local differences in the temperature sensitivity to CO2 and/or sea ice changes. For recent climate change, Jones et al. (2016) showed that the local temperature trends vary strongly across Antarctica, including cooling in some regions. If we use figure 1 in Jones et al. as a potential analogue, than 4 out of the 5 cores used by Beeman et al. fall into regions where temperature hasn't warmed or is cooling since 1979. Thus, from recent observations one can expect the response of Antarctic temperature to CO2 forcing to be spatially heterogeneous. Hence, instead of stacking the isotope records, I think it is more informative to determine the phase relationship for each ice core. Because: If we want to determine cause and effect between CO2 and Antarctic climate, it is the first robust temperature rise at any ice core that is informative and not their mean. Obviously, this will increase the noise in the temperature dataset, but ideally a change-point detection method can handle noise (see also specific comments). If all ice cores lead to internally consistent estimates of temperature change-points, their likelihood estimates can be combined to reduce the errors – but this consistency would first need to be shown. And in case of inconsistencies this needs to be discussed. The stacking may also influence the detection of fast temperature rises (such as around 16k) where already minor synchronization uncertainties and differences in resolution can lead to smoothing.

We have calculated timings for each ice core for the revised manuscript (See figures 1-5). We greatly appreciate this suggestion, which we think significantly improves our study.

The timings for the temperature records are generally consistent. The features analyzed for the stack in the original versions of the manuscript are not absent from any of the individual records, though they are assigned different probabilities. However, there are several differences between the temperature records, perhaps indicating the regional character of certain events.



Figures 1-5. Histograms made for all five cores included in the stack.

These new results merit discussion, and we propose to include the following text:

In the introduction:

"Previous work on the relative timing of CO2 and Antarctic Temperature did not take into account the possible regional differences in climate. Differences between West Antarctic and East Antarctic temperature during the last deglaciation have been noted (Cuffey et al., 2016). On much shorter timescales, ~~(~~Jones et al.~~,~~, 2016) note differing temperature trends at the drilling sites of the five cores used in this study over the period for which direct temperature observations exist (beginning in the mid-20th century). On the other hand, the interpretation of individual isotopic records can prove complicated, as local effects, including those of ice sheet elevation change and sea ice extent, are difficult to correct.

We provide change point timings and lead-lag estimates for the five individual isotope-derived records used in our stack as well."

In the results section:

"The change points for the five individual isotopic records are largely coherent with those identified for ATS2. However, some major differences do appear, and these differences merit discussion.

In the WAIS divide record, a small probability peak appears around 22ka. This peak does not surpass the 95% probability threshold (but does surpass the 90% threshold). We find it worth mentioning because ~~it represents~~ a much earlier beginning of temperature rise in the WAIS record. A stabilization after the 18ka temperature acceleration surpasses the 95% probability threshold at WAIS divide as well.

~~The temperature stabilization after the 14ka event is not significant in the WAIS record. Indeed, this stabilization is only a significant event in the EDC and Talos Dome records.~~

In the Dome Fuji record, the beginning of the temperature acceleration at the T1 onset appears to occur considerably before 18ka. We have no clear explanation for the subsequent temperature excursion. In the other four cores, this rise appears to occur more concurrently with the rise in CO2. It should be noted that the temporal resolution of the Dome Fuji temperature record is lower during the T1 onset than the other records.

Also of note in the Dome Fuji record is the abrupt jump in temperature around 14.7 ka, concurrent with a similar rise in CO2. No similar event is detected in any of the other regional temperature series.

Finally, a stabilization around 16ka is detected in the EDML core, concurrent with a similar stabilization in CO2. This stabilization is not detected in any of the other series. The probability of change around the ACR onset, on the other hand, does not pass the significance threshold in the EDML core."

In the conclusions:

"Some of the major millennial-scale changes in isotopic temperature records do not occur concurrently across all five cores. ~~The~~ T1 onset, ACR end, and Holocene onset do appear as significant changes in all five isotopic records. The ACR onset appears as a significant change in four out of five records. ~~The temperature stabilization following the ACR onset appears in two~~

~~cores. Finally, the 14ka jump only appears in the Dome Fuji record, and the 16ka stabilization only in EDML.~~

The inter-core differences in detected events, particularly those that appear to correspond with CO2 variations, pose challenging questions. Could these events indicate a relationship between regional isotopic signals and the mechanisms of CO2 release and uptake? These questions will remain to be answered by future studies."

The authors state throughout the manuscript that they detect abrupt temperature rises corresponding to the rapid CO2 rises discussed by Marcott et al. (2014). However, their method actually only detects the stabilization afterwards and not rises themselves (at least in temperature).

This comment is correct, and any phrasing which implies that the beginning of a rise is detected where it does not surpass the 95 % threshold (see below) in the revised manuscript.

In fact, this hints at another problem: The applied change-point detection method requires user input regarding how many change-points it is allowing for. The authors chose a number of 8 (and tested 7 for comparison). However, just by visual inspection of the CO2 record one can identify more change-points: i) the onset of the deglaciation, ii) the 16k rise, iii) the levelling off after 16k, iv) the rise at 14.6k, v) the levelling off after 14.3k, vi) the end of the ACR, vii) the jump into the Holocene and viii) the levelling off at the onset of the Holocene. Including the start and end point in the series which have to be marked as change points in the method, this yields 10 change-points. In fact, all of these points are discussed in the manuscript. This creates the inconsistency, that the method only allows for 8 (minus 2, i.e., start end) change points, but 10 (minus 2) are discussed – a solution that cannot be true in any of the realizations of the method: Some of these points must exclude each other if the maximum is set to 8. I think the authors should test their method allowing for more change points (e.g. 10).

We may not have been clear enough about a subtle point of our method, which we think likely makes it less sensitive to the number of change points. When we analyze a time series with *n* points, these points may be proposed anywhere in the time interval of the series, as long as the x-values increase monotonically. Then, all the accepted points are considered in the calculation of one probability distribution (rather than one distribution per point). Because a fit need not be perfect to be accepted in the Markov Chain Monte Carlo simulation, we may estimate more peaks of high probability than the number of points used in the linear representation. Thus, 10 peaks of probability for an 8-point simulation is not inconsistent. This will be emphasized in the revised manuscript.

We have performed a test using 10 change points, to be included in the supplementary materials for revised manuscript (Figure 6, below). This test shows convergence to approximately the same distribution as the 8-point test.

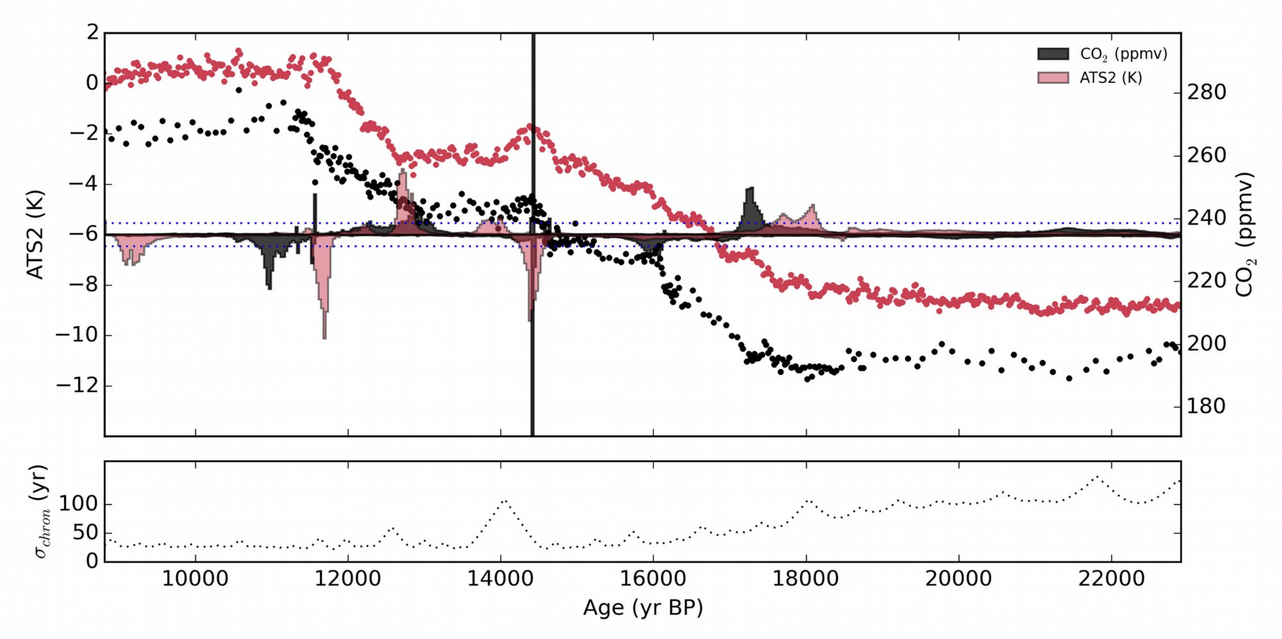


Figure 6. 10-point test.

It is unclear, how the change-points that are actually discussed are chosen. The authors discuss the CO2 rise at 16k, which has a very low probability peak, but do not discuss the large probability peak of a CO2 increase at the onset of the ACR. How is this justified? I agree that the rise at 16k is the relevant one, but the method obviously implies a higher likelihood of a CO2 increase at 14.3k? What does this imply about the methods ability to detect the correct change points and infer their timing? The authors should discuss more clearly, how they evaluate the likelihood of a given point to be a change-point at all, before discussing its timing.

We implement a probability threshold to discuss the significance of change points, and propose to include the following text on page 6, line 4, in addition to a figure :

"To estimate the significance of change points, a probability threshold is implemented. For each of the histograms, more than 95\% of the bins have (normalized) values below 0.0004—the value we select for the threshold. This threshold does not evaluate significance in the sense of comparison with a null hypothesis. We perform such a test by creating linear fits using 8 random samples of the linear interpolations between data points of each series, and do not apply a Metropolis-Hastings type criterion, but rather accept all fits. The bin values of the resulting normalized histograms do not surpass 0.0002."

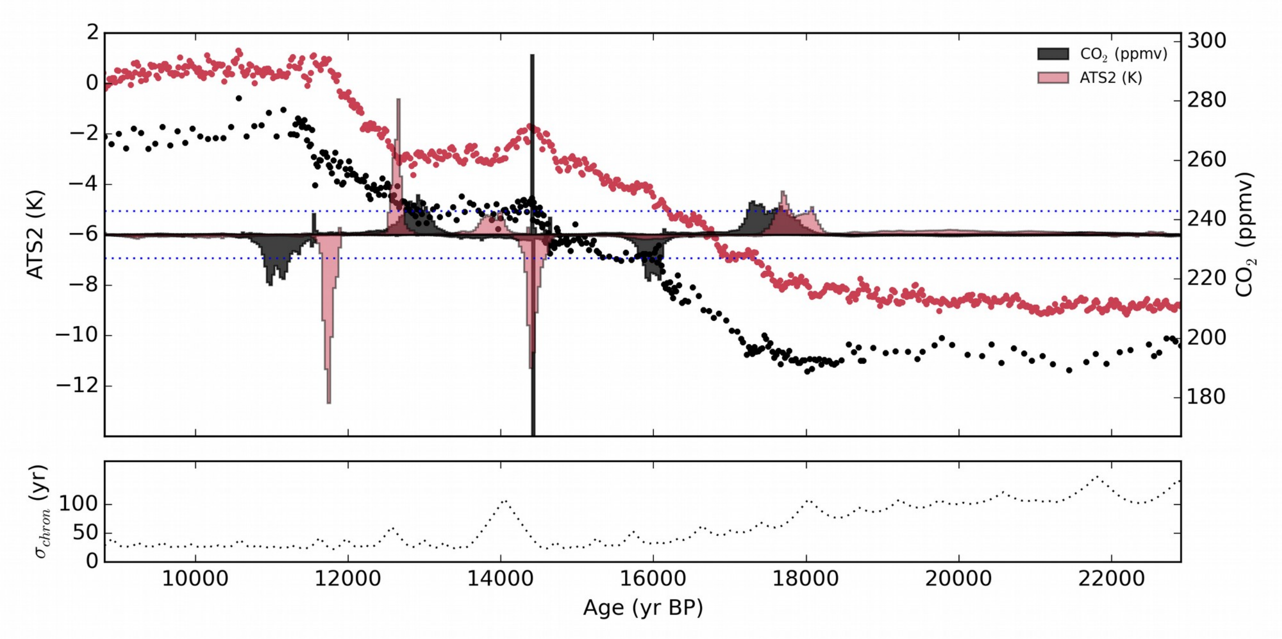


Figure 7. 95 % confidence threshold at 0.0004 (dashed blue lines)

See our discussion about the autocorrelation of residuals for more explanation of the large peak around 14.3 ka...

Specific comments:

PP1,L13: "Multimodal timings" – what do you mean? Multimodal probability distributions of change point estimates? I suppose the real change cannot be multimodal.

Omitted from the abstract.

PP2, L6: "Consistently": Replace with "on average", since Shakun et al. do not discuss whether this is consistent over the entire time.

Accepted.

PP2, L8: Also global T did not increase continuously according to Shakun et al.

This line has been changed to read:

"But Antarctic temperature and CO2 concentrations changed much more coherently as T1 progressed."

PP2, L20-21: Please rephrase the colloquial "thanks to the so-called isotopic paleothermometer". Possibly: ". . . due to temperature dependent fractionation of water isotopes during condensation"?

This line has been rephrased to "...due to the temperature dependent fractionation of water isotopes..."

PP2, L24-25: Please replace the CO2 lock in depth with the corresponding estimate from WAIS and *a reference to (Buizert et al. 2015). Generally it is not clear throughout the manuscript, whether EDC CO2 is used at all, and if so, how it is spliced together with WAIS CO2. If it is not used, please shorten the methodological discussion of EDC CO2 as it is misleading the reader to believe EDC CO2 data was used too.*

Replaced with:

"However, the age of the air bubbles is younger than the age of the surrounding ice, since air is locked in at the base of the firn (on the order of 70 m below the surface on the West Antarctic Ice Sheet (WAIS) Divide) at the Lock-In Depth (LID) (Buizert and Severinghaus, 2016)."

EDC CO2 data were not used at all. The later line on P4:

"At EDC, Delta-depth, the depth shift between synchronous air and ice levels, is calculated using an estimate of the LID based on nitrogen-15 data (Parrenin et al., 2013) that assumes negligible convective zone height."

has been omitted.

PP3, L10: d15N of N2 as a proxy for DZ height: Refer to (Buizert and Severinghaus 2016) who propose some uncertainty to this assumption?

Buizert and Severinghaus (2016), do not propose a quantitative uncertainty, we propose to include the following line.

"However, the assumption that d15N reflects DZ height is imperfect, as it may underestimate the DZ height for sites with strong barometric pumping and layering (Buizert and Severinghaus, 2016), generally those closer to the coast."

PP4, L14: "stack": Please describe how this is generated, so that other people can reproduce this. Is this an average over all records? Are they resampled to equal resolution beforehand? Are they standardized or kept in degC? If kept in degC, which slope (per mille/degC) is used? Do all cores have similar amplitudes across the deglaciation (in degC or per mille) or can the stack be biased by single records with exceptional amplitude? Later on, the fitting procedure requires an error in K for each data-point: How is this derived? Does it include uncertainty from the isotope temperature conversion and other (e.g., circulation) influences on isotopes?

We now include:

"The individual isotopic records are converted to temperature (C) and are corrected for source temperature (Bintanja et al., 2013), resampled to a timestep of 20 years, and averaged. The standard deviation of the records at each timestep is assumed to be representative of the uncertainty concerning the conversion from isotopes to temperature, and of the uncertainty rooted in the geographic distribution of the stack."

The spreadsheet used to calculate the stack will be made publicly available.

PP4, L17: "previously published ties": The list in (Parrenin et al. 2013) includes a lot of isotopic tie-points between the ice cores as well. Are these used here too? They introduce some circularity in the approach as they will reinforce the structure and timing of isotope (temperature) changes in the *ice core that is used as a target. Please clearly state, whether all tie-points are volcanic or not.*

We only use the volcanic tie points from Parrenin et al. (2013) + new EDC-DF and EDC-WD volcanic tie points. The tie-points themselves will be made available.

The line now reads

"We use previously published volcanic ties..."

PP4, L17-18: Synchronization: Please provide a list of tie-points as well as the ECM data for both cores in the supplementary, so that people can reproduce the analysis.

We will make these available in the supplementary materials and the appropriate paleoclimate databases.

PP4, L23: "1,030 measurements": Earlier (PP3, L14) it is stated that there are only 320 points? Are these replicates? If so, please state this as it is slightly confusing.

Rephrased to "1,030 measurements at 320 depths..."

PP4, L27-28: "At EDC. . .": Again – is this used at all? If not, remove. If yes, elaborate how CO2 records of such different resolution are stacked.

EDC was not used at all, this has been removed.

PP5, Figure2 caption: "Ratio of the age difference between two consecutive tie points.." . . . and what? Not clear what is plotted here.

Will rephrase to "Ratio of the age differences between two consecutive **pairs of** tie points..."

Section 2.3 – 2.5: I encourage the authors to have a look at these sections again, and try to rewrite them more clearly. As it is now, it is near impossible to really understand what's going on. The authors elaborate on how the MH sampler is working. This may be nice for saving computing time, but hopefully doesn't affect the results. A reference could be enough?

Only section 2.4 treats the MH sampler. Lines 15-23 are moved to the supplement.

The other two sections are important for reproducibility. Section 2.3 indicates how the goodness of individual fits to the time series is assessed, and section 2.5 details the formal calculation of leads and lags.

At the same time the authors do not discuss more relevant aspects of the method: How does the method deal with irregular sampling resolution in the records. They just say it becomes less precise (PP6, L3), but is that really so, or can it become biased? Similarly, it is not discussed, how the uncertainty for CO2 or ATS is derived. Are the residuals/uncertainties treated as independent or correlated? Since the method only detects linear trends, any other internal climate variability would basically be a correlated uncertainty (i.e., red noise, as opposed to white measurement noise)?

The residuals are corrected using the inverse of an autocorrelation matrix (we assume the residuals to be correlated), which is estimated after an initial sampling run. This is treated on P6, lines 0-5.

We add, on P6, line 5:

“Some limitations of this method should be made clear: first, where data are more sparse (i.e. in the CO2 series after the Holocene onset) change point identification becomes less precise and may be biased by the lack of data. Second, the method is not designed to treat variations that depart significantly from a linear shape, and accelerating trends can lead to noise in the histogram representations.”

PP7, L13: “. . .and corresponding rise in temperature around 16ka”: Is that true? Looking at figure 3, there is no detection of a ATS increase around 16ka. Only the stabilization. In principle I agree, that there appears to be an ATS increase. However, there is a similar increase around 17kaBP in ATS without a corresponding change in CO2, implying that this may just be internal variability/noise. And in any case: The method does not detect either of these increases in ATS.

Eliminated.

PP7, L17: see previous comment. The ATS rise at 16k is not actually detected by the method.

Eliminated.

PP7, L18-21: Multimodality: This paragraph is written as if the CO2/ATS increase was multimodal. However, each realization of the method probably only picks one or the other mode as a possible change point, and never both. Hence, the multimodality reflects an uncertainty in the change-point identification, not the identification of two separate change-points. Is this correct? If so, please rephrase.

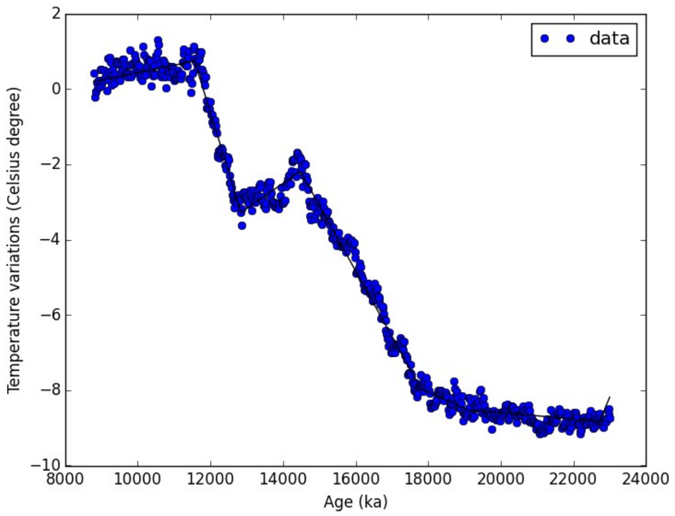


Figure 8. Two change points in a small region, 18-19 ka.

Two change points can be detected, see the figure for reference. However, fits may also only contain one point.

The text is changed to:

“The deglaciation onset begins with a large, postive change point mode for Antarctic temperature, centered around 18.08 ka. A second mode follows, at 17.70 ka. Similarly, there are two modes for the CO_2 series, centered at 17.63 ka and 17.30 ka. In both series, the two modes are upward-oriented. Investigation of the ensembles of fits indicates that the modes can represent an initial, more gradual increase in the rate of change, followed by a sharper acceleration in both series. However, since the probability peak here is continuous, we take it to be representative of a single change point."

PP7, L22: "a small positive probability peak". The probability of this point being a change-point is very low, much less than for example the positive peak for CO_2 around 14.4ka which is not discussed. How do the authors choose which peak to discuss? How trustworthy is a change-point that is apparently only used in a small number of iterations?

This is resolved by the implementation of the 95 % probability threshold. Reworded to:

"CO_2 began to change more quickly at around 16.15 ka, though we do not calculate a significant change point here. This rise abruptly peaked at 16.07 ka, and finally stabilized at 15.9 ka. These events are both identified by downward-oriented probability peaks."

PP7, L24: See comment above. The ATS probability peak is very low. How reliable is the inference of a change point there (and its timing)? PP7, L 24-25: This section illustrates my concerns about how well the method deals with noise, and how subjectively some probability peaks are discussed while others are not. Why is the positive probability peak in CO_2 at 14.64ka discussed, while the bigger positive peak at 14.42 isn't? I agree that the rise at 14.64ka and the stabilization around 14.42ka are the relevant change points, but the statistical method doesn't. To me this highlights, that the method may underestimate noise in the CO_2 and ATS data, and hence, depict potentially erroneous change-points, which also have a seemingly high degree of certainty (in terms of timing). Please comment.

The apparent rise at 14.64 ka indicates that the method attempts to treat the rapid rise as a step-change, which is probably not correct. Sensitivity tests to the autocorrelation matrix taken into account when estimating the residuals indicate that if the values of this matrix are artificially lowered, the step-change can be removed.

However, perhaps counter-intuitively, lowering the values of the autocorrelation matrix (which is normally estimated empirically) amounts to allowing the model to over-fit with respect to noise. The baseline values of the histograms become higher in the regions between peaks, and the peaks themselves lower, indicating that change points are allowed to represent what is essentially white noise.

Reworded to:

"A second abrupt CO_2 rise preceded the Antarctic Cold Reversal. During this rapid rise, CO_2 experienced major sub-centennial scale variations, and the corresponding probability peaks are noisy and large. Two narrow spikes in probability, one at 14.64 ka, and one at 14.42 ka, mark its beginning and end. The second peak is significant in both directions; investigation of the time series shows that a CO_2 jump of around 9 ppm occurs here in the data series."

PP8, "leads and lags": Generally, I think the results should be more explicitly compared to (Parrenin et al. 2013), who applied a largely similar method to a different CO_2 (and slightly different ATS) dataset. Their estimates could be shown in figure 4 for comparison. Are the results *consistent for each change point?*

A new version of figure 4 will be included, and is shown here:

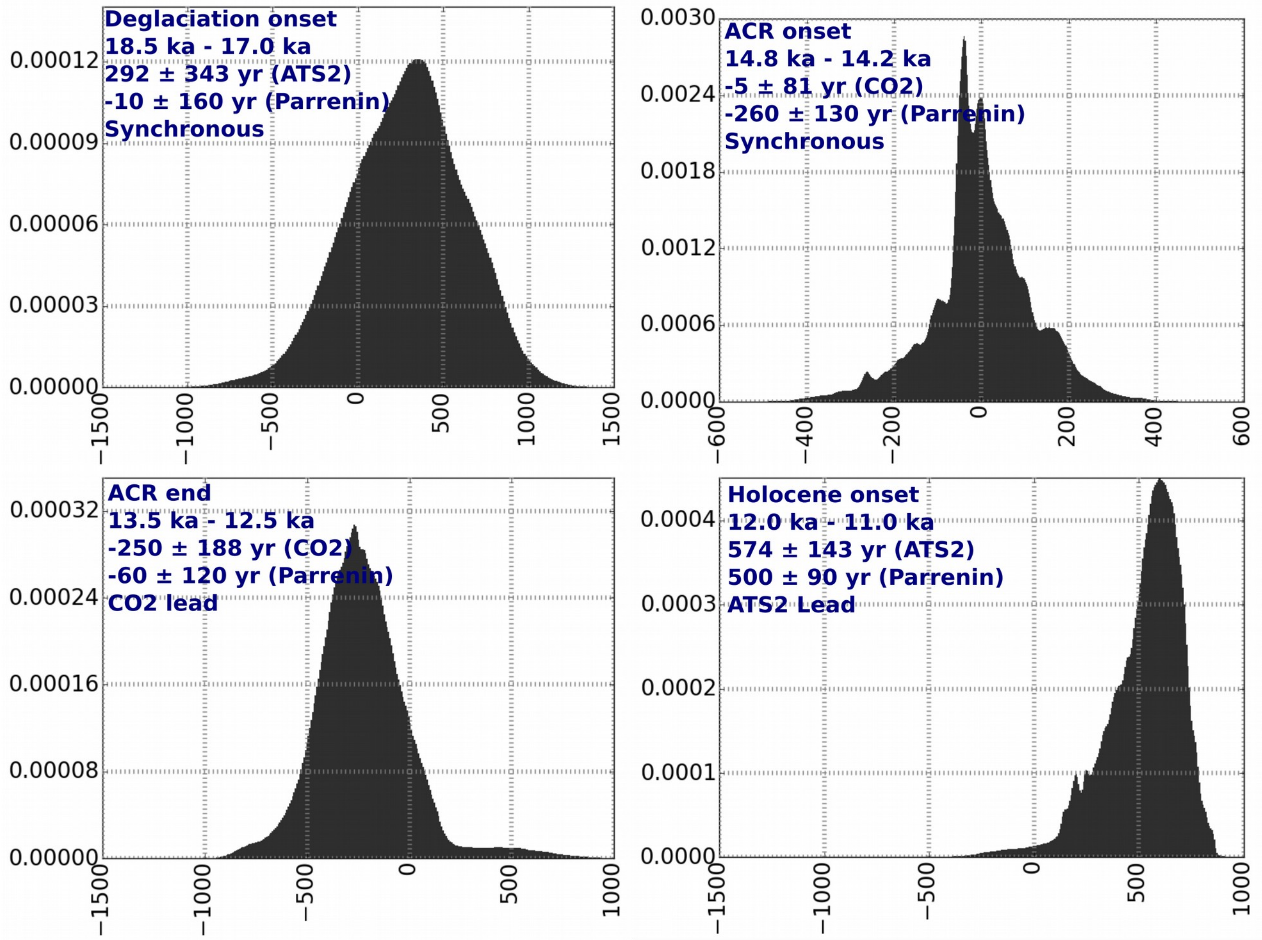


Figure 9. Revised leads and lags

PP8, L2: ". . .it is not obvious." Change to: "our method doesn't detect it."

Accepted.

PP8, L7: "either the histogram peak around 12.9 or . . .". I don't think that you can interpret single peaks in the histogram like this. The timing of the rise in CO2 at the end of the ACR is detected as a broad probability distribution and not just by two minor peaks in the histogram. All values within the probability distribution are the possible "true" value with a given likelihood. Even if the peaks are the most likely single values, the cumulative probability of the true value not being these peaks is higher (cf. it is very unlikely to actually draw the exact mean value of a normal distribution).

Accepted, these peaks are likely not distinct. Changed to:

"The ACR terminated with an increase in CO2, beginning at the peak occuring around 12.9-12.8 ka. ATS2 most likely began to increase at 12.66 ka."

PP8, L6: "which loses data resolution.." See earlier comments: How does the method deal with this? PP8, L6-8: Again: Can all these modes really be interpreted in other terms than uncertainty? The authors mention "later peaks, where data resolution is lower, to be likely indicative of higher *frequency variability or noise". If the method cannot handle those, how good are the timing estimates for the other change-points? Please elaborate.*

The modes are still interpreted as method-related uncertainty, and are still included in the lead-lag calculation. The line "later peaks, where data resolution is lower, to be likely indicative of higher frequency variability or noise" is confusing, as it implies we do not include this period in the calculation. We omit this line.

PP9, L1-2: "Applying the cross correlation operator. . .". How is it handled when the method indicates near equal probabilities of a CO_2 rise and fall like around the ACR onset?

The cross-correlation operator is applied only to the probability of a fall. We pick only the coherent direction.

The equality around the ACR onset is an error of scale, this will be changed in the figure.

Text changed to:

"The probability that one variable leads the other over a given time period can be calculated by applying the cross-correlation operator to the histograms of the two variables over a given time period and given direction (we make calculations only for the coherent change direction)"

PP9, L13-14: See earlier comments on the reliability of the method and the handling of noise and resolution.

PP9, L14: "Calculating the phasing between 12-11.5. . ." How is this done? Are certain values excluded from the histogram for CO_2? How?

Rephrased to:

At the Holocene onset, a CO_2 lag is certain. Calculating the phasing between 12.0 ka and 11.0 ka, we obtain an ATS2 lead of 574 +- 143 years. However, the decrease in data resolution here may bias the estimate.
~~We could consider the initial stabilization in both series to be coherent, and the following CO2 modes to be representative of bias resulting from the change in data resolution.~~ Calculating the phasing between 12.0 ka and 11.5 ka, we obtain an ATS2 lead of 195 +- 62 years. Note that this timing appears as a minor mode of the phasing calculated between 12.0 and 11.0 ka. However, this calculation can only be performed making considerable assumptions about the data, and the estimate of 574 +- 143 is statistically more appropriate.

PP10, L5-6: "We identify a coherent ATS2 change-point. . ." See general comments. I don't think this is the case.

Accepted. Changed to:

We identify a CO_2 change point not treated in these studies at 16 ka, associated with the centennial-scale rapid rise...

PP 10, L11-12: "minor modes". See earlier comments. Can these really be interpreted?

Changed to:

Additionally, many of the changes in CO_2 are overlayed with centennial-scale substructures. The CO_2 modes around the ACR onset are major and statistically significant.

PP10, L27-29: "Mt. Takahe" I do not understand why the cumulative probability of the ATS2 change-point is relevant here. The method does not detect whether there are multiple change points there, but only a single one with a given uncertainty in timing. Mt. Takahe falls into the uncertainty range of the detected change point at that time. Correct?

The cumulative probability is important when assessing whether one event occurred before or after another. Reviewer 2 comments:

"P10 line 22 to 29. McConnell et al suggested that accelerated warming was triggered by the Mt Takahe eruption. The finding here, that accelerated warming begins *before* the Mt. Takahe eruption, contradicts the McConnell hypothesis. The spin about "additional forcings beginning to accelerate warming before Takahe" is very unconvincing."

We agree that the eruption falls within the histogram, but it falls far to one side! (For a gaussian-shaped distribution, which this peak is not necessarily, we could calculate whether or not it falls outside the 1-sigma range).

PP11, L1: "Here we confirm". See general comments. The method does not detect ATS rises coinciding with the rapid CO_2 rises.

Changed to: Here, we confirm that the ends of two of these events correspond with stabilizations in the Antarctic temperature record.

PP12, L 3: "we identify change points". See earlier comments.

Accepted. (Deleted)

Technical Comments: PP2, L14: ". . . is thought to have. . ." (not "haved")

Accepted.

PP2, L19: ". . . and atmospheric composition. . ." (not "atmosphere")

Accepted.

PP4, L18: "Sigl et al. 2015": Change to 2016.

Accepted.

PP4, L17-18: "the offset in between ice and the air trapped much later at a given depth": Convoluted, please rephrase. Possibly: "The age difference between trapped gas and the surrounding ice matrix, delta age, . . ."

Accepted.

PP7, L22: "a small positive probability peak". Probability is always positive. Please rephrase.

This peak does not meet our probability threshold, will be eliminated.

We could generally refer to downward-oriented and upward-oriented changes.

PP9, L4: "2 sigma": The numbers in the following paragraph (PP9, L7-9) match the numbers in figure 4. However, the caption of figure 4 says this was 1 sigma? Please check.

PP9, L8: "At the peak of the 16ka rise". Since you do not detect the 16k rise in ATS a better formulation could be: "CO2 and ATS stop rising synchronously at the onset of the ACR" or similar.

Accepted. Changed to:

"At the peak of the 16 ka CO2 rise, a rise is not detected with statistical certainty in the temperature series."

PP10, L4: replace "in AMOC" with "from AMOC"

"originate from" seems less correct than "originate in". Decision left to the editor.

Reviewer 2.

Chowdhry Beeman et al. investigate the time relationship between Antarctic temperature and atmospheric CO2 during the last deglaciation. The question is of importance for our understanding of climate-carbon cycle feedbacks and has been tackled in multiple previous papers, most recently Parrenin et al 2013 and Pedro et al 2012. Chowdhry-Beeman et al. is distinguished from these previous studies by its use of the high time resolution and low delta-age WDC ice core CO2 record and a new regional Antarctic temperature stack. The consensus emerging from previous studies is that there is little-to-no significant time delay between CO2 rise and Antarctic temperature rise throughout most of the deglaciation. Chowdhry-Beeman et also find synchronous changes within uncertainties (excepting the Holocene onset), however they place more attention than others on centennial-scale signals in the temperature and CO2 series and their purported relationship. The techniques used to analyse of the phase relationship are more complex than used in previous studies. Although I have no reason to doubt the techniques, I do have some concern about their rather qualitative and selective interpretation. In particular the conclusion that there is a significant change in Antarctic temperature corresponding to the abrupt CO2 change around 16ka is not convincing. My overall impression is that the complex technique used to generate the PD histograms is out of proportion to their qualitative interpretation. Missing is some clear hypothesis testing and a more sceptical view of whether the centennial scale signals detected in ATS and CO2 are really meaningfully related. In general the approach appears very promising but the interpretation is still requiring quite some work. I would support publication after revisions to address the concerns below.

Major Comments Section 3.1: *The technique applied to study the time relationship between ATS2 and WDC CO2 is more complex than previous studies and difficult (at least for this reviewer) to follow. My concern is that despite the complex approach the interpretation of phasing, in the end rests on a rather qualitative assessment of the change point histograms presented in Figure 3. Adding to my concern is that there is never a clear description of what makes a mode distinct and worthy of discussion or of the precise criteria for defining a significant change point.*

We propose to address this criticism as proposed by the reviewer, below.

The caption of Fig 3 says that the y-axis range for the probability histograms is 0 to 0.0024. It's not clear to me precisely what is meant here.

Probabilities are normalized, and thus the integrals of the histograms over the entire study period should sum to one. We will make this more clear in the caption.

Can you show horizontal lines marking e.g. 0.05 probability cut-offs, which could then be used to judge which modes are significant?

For each of the histograms (concave-up and concave-down change points for ATS2 and CO2), more than 95% of the bins have (normalized) values below 0.0004—the value we select for the threshold.

Since defining a threshold does not evaluate significance in the sense of comparison with a null hypothesis, we perform a second test—we create linear fits by using 8 random samples of the linear interpolations between data points of each series, and do not apply a Metropolis-Hastings type criterion, but rather accept all fits. These fits should still approximate the shape of the series, but the points will not converge to major changes. The bin values of the resulting normalized histograms do not surpass 0.0002.

Of the events we discuss, this threshold notably eliminates the discussion of a possible temperature increase around 16ka. (See Figure 7, response to reviewer 1).

The following text is included:

" To estimate the significance of change points, a probability threshold is implemented. For each of the histograms, more than 95\% of the bins have (normalized) values below 0.0004—the value we select for the threshold. This threshold does not evaluate significance in the sense of comparison with a null hypothesis. We perform such a test by creating linear fits using 8 random samples of the linear interpolations between data points of each series, and do not apply a Metropolis-Hastings type criterion, but rather accept all fits. The bin values of the resulting normalized histograms do not surpass 0.0002."

We are told (pp7 line 18) that the degalacial CO2 rise features 'two modes' (17.63 ka and 17.30 ka) separated by a 'distinct anti-mode'. On what basis is the anti-mode distinct? Could this be over-interpretation of noisy data?

The description of the antimode is removed from the text.

Further down (line 24) the authors describe 'a broad low probability peak' in ATS2 at 15.96 ka. It's not explained how 15.96 is selected as the centre of this peak or why this peak is considered significant, given there are similar amplitude peaks elsewhere in the deglacial CO2 record that are not discussed at all. The same can be said for the 'small upward probability peak' in CO2 at 16.15ka. This ambiguity about what is a significant feature and what is not continues throughout the section.

This ambiguity is removed by the introduction of a probability threshold. This peak will no longer be discussed.

To give another example (line pp8 line 7), the author's describe 'two larger modes in CO2 at 11.12 and 11.01 ka.' as being 'indicative of higher-frequency variability or noise.' It's not clear on what basis these peaks are considered noise whereas the (smaller) peaks around 16 ka are considered *meaningful and related to ATS2. My overall impression is that the complex technique used to generate the PDs is out of proportion to their qualitative interpretation.*

The large modes in CO_2 at 11.12 and 11.02 ka are still included in the calculation of leads and lags. However, we cannot be certain if they are representative of individual changes, or rather of the severe reduction in the resolution of the CO_2 series at the onset of the Holocene; we thus refrain from further commentary. This section is now reworded to:

"At the Holocene onset, a CO_2 lag is certain. Calculating the phasing between 12.0 ka and 11.0 ka, we obtain an ATS2 lead of 574 +-143 years. However, the decrease in data resolution here may bias the estimate.
We could consider the initial stabilization in both series to be coherent, and the following CO_2 modes to be representative of bias resulting from from the change in data resolution. Calculating the phasing between 12.0 ka and 11.5 ka, we obtain an ATS2 lead of 195 +-62 years. Note that this timing appears as a minor mode of the phasing calculated between 12.0 and 11.0 ka. However, this calculation can only be performed making considerable assumptions about the data, and the estimate of 574 +- 143 is statistically more appropriate."

I suggest the authors revise Section 3.1 to be shorter and more quantitative. I think part of the reason the some of the interpretation is unconvincing is that the authors do not appear to use their results to test any specific hypotheses. Instead we get a rather post-hoc interpretation of the leads and lags. Reframing the introduction to set up some specific hypotheses for testing could make the discussion more convincing.

We accept this suggestion for the revised manuscript. The main working hypothesis is that CO_2 and ATS2 are synchronous and coherent. An important hypothesis that can be tested is included in the comments of reviewer 1, with respect to regional differences in temperature series.

Section 3.2: The methodology here appears good, however the section rests on the selection in Section 3.1 of five 'common' change points in ATS2 and CO2. As above its not clear by what criteria these 5 are selected. Please clarify.

These points are now using a probability threshold, as mentioned above.

p10 line 5. The significance of the ATS2 change point at 16k is not convincing. Please be more clear about the criteria for its selection over and above other peaks in the PDs that are not discussed at all.

We are not convinced by this point either-it does not meet the probability threshold. This point will no longer be included.

P10 line 11. " during the complex, centennial-scale changes associated with the 16 ka rapid rise and the ACR onset, ATS was most likely synchronous with CO2 ". This is not convincing given the +- 340 yr uncertainty and the ambiguity of the 16ka peak in ATS2. I'd suggest a more cautious interpretation: centennial scale variability in both series (possibly physically related, possibly not) restricts ability to make any clear statement on significant leads or lags during this interval.

We agree with this interpretation, reworded to:

"However, during the complex, centennial-scale change at the ACR onset, ~~ATS was most likely~~

~~synchronous with CO2...”~~

*P10 line 22 to 29. McConnell et al suggested that accelerated warming was triggered by the Mt Takahe eruption. The finding here, that accelerated warming begins *before* the Mt. Takahe eruption, contradicts the McConnell hypothesis. The spin about "additional forcings beginning to accelerate warming before Takahe" is very unconvincing.*

We accept this clearer rephrasing of our findings. However, it is important to note that our findings do not fully contradict the McConnell hypothesis (see reviewer 1's comment as well), but rather find it improbable to a certain degree. The line about additional forcings is eliminated, and the paragraph now reads:

“However, the cumulative probability of the ATS2 change point is much greater before 17.7 ka than after.”

*p11 line 1. The authors claim here to ’confirm’ an imprint on Antarctic temperature of ice berg discharge to the Sth Ocn *and* Nth Atlantic around 16ka". This is not convincing at all. First, as above, the ATS2 signal around16k is questionable given other similar sized peaks in the PDs that are not discussed. Second, as the authors well know, correlation in timing does not prove of a casual relationship. Third, what is the imprint supposed to be (warming, cooling, stabilization?) and how did the icebergs drive it? Revise.*

This paragraph is removed.

p11 line 5. The ’reversal in phasing’ between T1 and the ACR end is not convincing. The phasing at T1 is 292+-343 yrs (1 sigma!), thus spanning from CO2 lag to CO2 lead. How can the phasing reverse if it is not distinct at T1? A simpler interpretation is that the ATS2 and CO2 are roughly synchronous with the exact lead-lag varying between change points due to centennial scale variability in both series.

We accept this interpretation. This now reads

“Though the T1 onset and the ACR end are both roughly synchronous, ~~the most likely~~ directionality of phasing ~~is~~ reversed.. These two events are structurally similar, and it has been postulated that both originate in AMOC reductions (Marcott et al., 2014). It appears that centennial-scale variability in the two series can modulate the timing of the effect of an AMOC change on CO2, Antarctic temperature, or both.”

p 11 line 9: "Centennial-scale variability may have been superimposed on coherent millenial scale trends, for example". Performing a similar analysis on band-pass filtered versions of the two series could be used to test this idea and would add a substantial new result.

We propose to include the following in the results section:

“We apply a Savitsky-Golay filter designed to have a cutoff periodicity of approximately 500 years to the two series, resample the series to a 200 year timestep, and perform the fits again, assuming residuals to be uncorrelated because of the lower resolution. We identify change points in the same regions, though the distributions of these points are much broader and smoother. ~~This~~ indicates sensitivity to the centennial-scale variability superimposed on major millennial-scale changes ~~(but not, importantly, to periodic centennial-scale variability in general).~~

Figure 10, which represents this fit, will be included in the supplement.

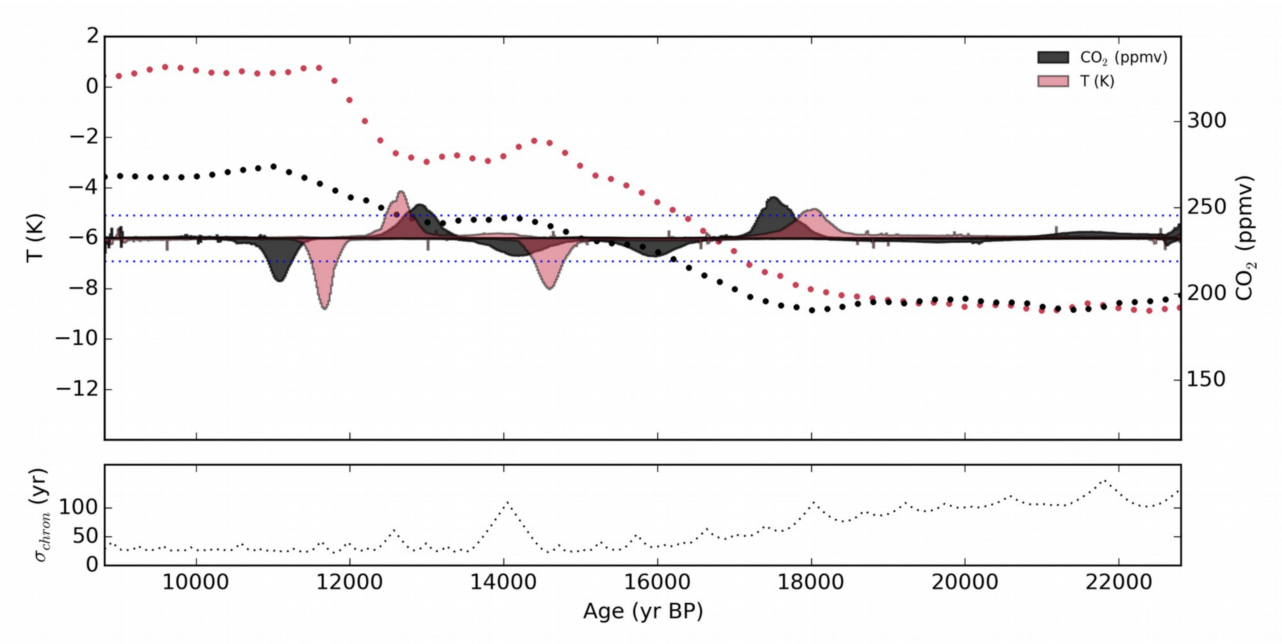


Figure 10. Fits to Savitsky-Golay filtered series.

p12 line 18. Comparison between east and west Antarctic temperature and CO2 could already be done by making an east Antarctic and west Antarctic stack. The authors might consider doing this in revisions, it would add a substantial new result to the lead and lag discussion.

As suggessted by Reviewer 1, we have applied our method to the isotopic records from each ice core. See Figures 2-6 above.

p10 line 5. It should be mentioned that within uncertainties the results are consistent with Parrenin et al and Pedro et al.

Accepted. Now reads

"Within the range of uncertainty, our lead-lag estimates are consistent with those of Pedro et al. (2012) and Parrenin et al (2013). The addition of the WD paleotemperature record and removal of the Vostok record from ATS2, the updated atmospheric CO$_2$ dataset, and our more generalized methodology are all, in part, responsible for the differences in computed time delays ~~(SI)~~."

*Figure 4. Important typo. I think the phasing at the ACR end should read *-*250 +- 188.*

Accepted, likewise for the ACR onset.

P 11 line 11. It's very difficult to follow this sentence. Please revise

Reword to :

Bauska et al. (2016), for example, hypothesize that an earlier rise 10 in CO2 at 12.9 ka, driven by land carbon loss or SH westerly winds, might have been superimposed on the millenial-scale trend.

P12 line 3. " Notably, we identify change points in ATS2 that are associated with rapid rises in CO2." Which change points exactly? The previous paragraph comments that rapid change in CO2 and ATS2 around the ACR are not clearly in common. And my concerns remain about the significance of any signal in ATS2 around the rapid 16ka signal in CO2. Without further evidence this conclusion of related abrupt changes in not convincing and not justified to include as a major conclusion here or in the abstract.

Accepted.

P12 line 13-15: "This variability suggests complex mechanisms of coupling. Indeed, perhaps different mechanisms of ATS2 and CO2 rises, some coupled, others decoupled, were activated and deactivated (Bauska et al., 2016) throughout the deglaciation." This statement so encompasses all possibilities that it is almost meaningless

The second sentence (highlighted) will be omitted.

Please advise where the new Antarctic temperature stack will be made publicly accessible upon publication.

The stack is already available on the linked github page (https://github.com/Jai-Chowdhry/LinearFit-2.0-beta/tree/v0.0) as ATS2-new-sigma2.txt. It will be made available on Pangaea/NOAA Paleoclimate upon publication.