Response to review 2 (Nat Lifton) of 'Reversible ice sheet thinning in the Amundsen Sea embayment during the late Holocene', by Balco, Brown, Nichols, Venturelli et al.

Nov 22, 2022

To begin, we would like to thank reviewer Nat Lifton for his close attention to the technical details of the C-14 measurements. As the review recognizes, the C-14 data set reported in this paper is more extensive, and involves measurements at lower concentrations, than nearly any other published work. For this reason, we describe in this work a larger data set of measurement blanks than described elsewhere, and we thought very carefully about blank and background corrections and the associated uncertainty analysis. We very much welcome the opportunity to air these issues in the online discussion.

The bulk of Lifton's comments focus on these background corrections, specifically on three main questions: one, whether time-dependence of measurement blanks should be taken into account in blank corrections; two, how detection limits are quantified; and three, the importance of scatter in replicate measurements, and how it should be taken into account in data interpretation. Analysis and discussion of these issues was, in part, addressed in various aspects of the supplementary data provided with the paper. For example, several of the figures shown below are taken from a MATLAB notebook that is included as part of the supplementary data. However, for convenience for readers of the online discussion, in the first section of this response we will summarize and augment this analysis.

A secondary aspect of Lifton's review focused on production rate estimation and interlaboratory comparison via the CRONUS-A measurement standard. We will address this in the second section of the response. A third section of the review had some queries about the form of the measured concentration-depth profiles, which we will address in a third section of the response. Finally, the review pointed out several minor errors and omissions (e.g., 5730 vs 5700 +/- 30) that we will correct in a revised manuscript.

Overall, we will argue that (i) our analysis of the measurement background data and the blank correction scheme derived from this analysis are justified by the properties of the data set, and (ii) the issues the review brings up relating to CRONUS-A reproducibility are useful as context for the measurements, but fundamentally not relevant to the conclusions of the paper. However, we will not propose alterations to the main text of the paper based on these issues. Our reasoning here is that these technical aspects, although in many ways the most novel aspects of this work and potentially the most important for future similar applications of cosmogenic C-14, are highly technical in nature and, sadly, of interest only to a small readership. Because Cryosphere is an open review journal and this review response will therefore be publicly accessible with a DOI in future, the minority of readers that are interested in the technical discussion of this issue can find it here, without the necessity of adding extensive technical material to the main text of the paper itself that would potentially distract from the main point of the work. In addition, we will update the supplementary data to include the data and code needed to generate all the analysis and figures shown below.

I. C-14 blank corrections.

I.1. Time-dependence of C-14 measurement blanks.

Lifton proposes in his review comment that instead of using an aggregated distribution of measurement blanks over a long period (henceforth, the 'aggregate approach'), as we have done, it might be preferable

to use only blanks measured close in time to samples for blank corrections. The premise of the approach proposed by Lifton (henceforth, the 'close-in-time approach') is, basically, that the measurement background, as quantified by process blanks, varies systematically in time such that the time-dependent variation is larger than any non-time-dependent variation (of course, it is impossible to measure a non-time-dependent variation because blanks must be measured sequentially, but this is a good way to think about it).

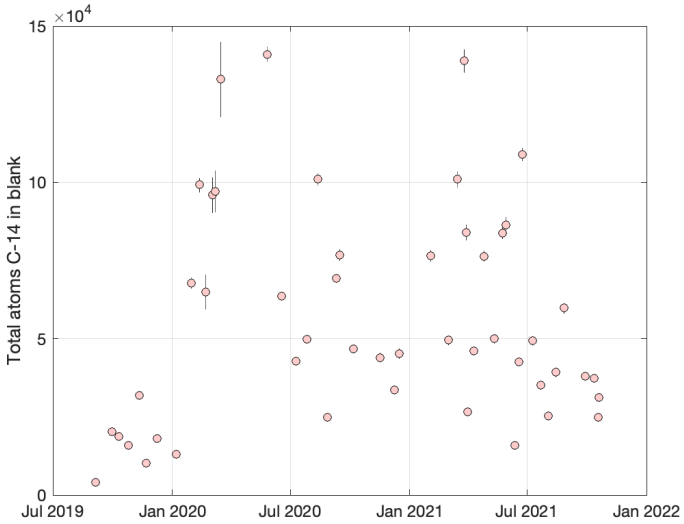

Figure 1. C-14 measurement blanks run at Tulane in 2019-2021. Error bars are 1-standard-error nominal measurement uncertainty, which primarily comes from the AMS measurement.

Note: the data and code to make all the figures in this response are included in MATLAB workbooks in the supplementary information:

C14_blank_time_dependence_202211.mlx
C14_uncertainty_experiments.mlx

Figure 1 shows the sequence of measurement blanks run at Tulane over a 2-year period. This two-year measurement period is chosen because it begins at the time of the last significant change to the extraction system and procedure (a change from alumina to platinum boats for sample heating) and also encompasses all of the measurements (surface and core) included in this study. As a first point, it is immediately evident that the variability in the blank on any time scale greatly exceeds the internal measurement uncertainty in each blank inferred from the AMS and CO_2 measurements. Thus, the internal uncertainty estimate is not relevant to computing the true uncertainty in a blank subtraction scheme. Henceforth, we ignore the internal measurement uncertainty of the blanks and accept that the uncertainty distribution in any blank correction scheme must be derived from the distribution of a number of measurement blanks.

From visual inspection, some time-dependence appears to be present in these data, but the time-dependence appears to be manifested mainly as changes in the variability among blanks, expressed as an increased or decreased likelihood of observing high values. For example, blanks measured after July 2021 were mostly < 50,000 atoms. Between Jan 2021-July 2021, there were also many blanks in the same range, but they were interspersed with many other blanks approaching and exceeding 100,000 atoms. Thus, the variation in measures of the dispersion (e.g., the standard deviation) is more striking than that in central measures (e.g., median or mode). This is not like the situation for a typical application of a time-dependent blank correction in which different time periods display similar scatter but different central values. Remember, the condition that would justify a close-in-time approach is that the time-dependent variability is larger than the non-time-dependent variability. At first glance, it does not appear that this is the case.

It is possible to quantify the relative importance of time-dependent and non-time-dependent variability by considering the autocorrelation of the blank series. If time-dependent variability is important, the value of each blank should be highly correlated with the value of the previous (or next) blank.

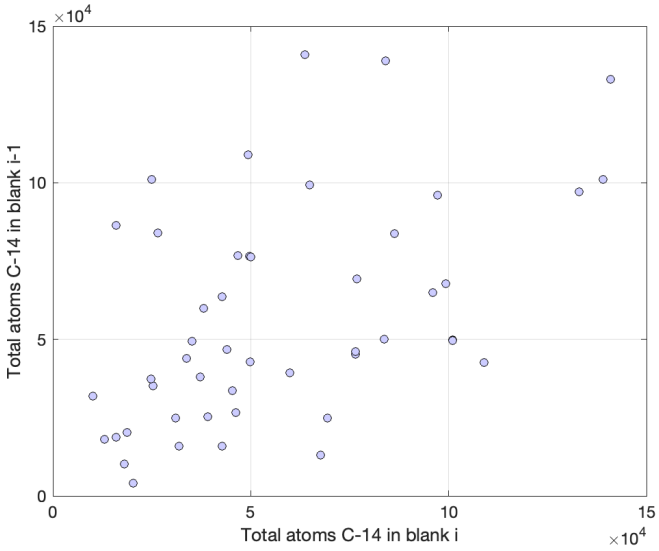

Figure 2. Autocorrelation in series of measurement blanks.

Figure 2 shows the autocorrelation analysis. Although the correlation of blank *i* with blank *i-1* is significant ($p < 0.05$), it is weak (correlation coefficient 0.49). The r^2 of this relationship is 0.24, which indicates that only 24% of the variability in the blank is time-dependent. The majority of blank variability is therefore non-time-dependent.

If we use this observation as a guide to how to proceed with a blank correction scheme, it is evident that a strict close-in-time approach, for example using the average of two bracketing blanks for each sample or similar approaches, would lead to underestimation of the uncertainty in a blank subtraction scheme. We therefore reject this type of approach. This leaves two possible approaches. The more complex of the approaches would be to try to incorporate both time-dependent and non-time-dependent variability by a moving-window or moving-kernel type of scheme in which we sought to characterize, for example, the mean and standard deviation of measurement blanks within a certain time period around a sample measurement, and use those values for blank correction. We believe that this type of an approach is not strongly justified because it implicitly includes assumptions about the physical processes leading to variability in the blanks. For example, if measurement blank results were related to a "memory" of the sequence of non-blank samples that were run in the extraction system (which is very possible), then a moving-window or moving-kernel approach that was not aware of the samples could be very misleading. If we knew what the process leading to blank variability was, we could tailor such a scheme to avoid gross errors (for example, if we KNEW that blank variability was related to sample history we could use a backward-looking window). However, we don't know this, so we don't know whether a moving-window scheme would or would not result in serious errors.

The less complex of the two possibilities is to disregard possible time-dependence, and use an aggregate approach that uses blank data collected over a long period of time in which the extraction line configuration was not modified (as is the case for the data set shown here). The advantages of the aggregate approach are that (i) it is consistent with the observation that the majority of blank variability is non-time-dependent; (ii) it uses the maximum amount of data so is more likely to correctly detect and quantify a long-tailed distribution (which is a suspected issue based on previously published C-14 data); and (iii) it does not require assumptions about the physical process leading to blank variability. The disadvantage is that we would be ignoring the fact that even though the time-dependent component of the

blank variability is small, it is not zero. As Lifton proposes in his review, if there is a period of time during which the blank distribution is different from the aggregate distribution, the blank correction will be incorrect. As pointed out in the review, the core sample measurements were made in late 2021, when the distribution of blanks appears to be lower and less scattered than the aggregate distribution. In this case, the risk is that the aggregate approach may lead to an overcorrection for blank and thus an underestimate of the true C-14 concentration in core samples. This, in turn, could lead to a spurious conclusion that C-14 was not present above background in core samples when in fact it was present, that is, a false nondetection. However, it could not lead to the opposite spurious conclusion, that is, a false detection. In the context of this study, a false detection is a more significant risk than a false nondetection.

Another way of expressing this is that, as stated in the review, if we used a close-in-time approach in which only blanks measured in late 2021 were used as a basis for blank-correction of the core data, it would strengthen the conclusion that we have measured C-14 concentrations significantly above background in these samples. However, as we discuss below in section I.2, this conclusion is already true at very high confidence when the aggregate approach is used. Thus, the choice of a close-in-time vs. aggregate blank correction approach does not have a significant effect on the main conclusion of the work.

Taking these considerations into account, we concluded that blank correction using a lognormal distribution derived from the aggregate approach was more consistent with (i) the analysis of time-dependent vs. non-time-dependent variability showing that non-time-dependent variability is more important, (ii) suggestions in previous studies that C-14 measurement background is long-tailed, and (iii) the importance of minimizing assumptions about the unknown physical processes leading to blank variability. In addition, as we discuss later in this response, we found that the blank distribution inferred from an aggregate approach was consistent with scatter in replicate measurements, whereas a narrower distribution that would be inferred from a close-in-time approach would not be consistent. Finally, we noted that the main risk of the aggregate approach is a false non-detection of C-14 in our subglacial bedrock cores, so from this perspective the use of the aggregate approach can be considered a stricter or more conservative test of the hypothesis that there was a middle Holocene ice thickness minimum. Thus, we proceeded with the aggregate approach.

I.2. Detection limits

The issue of detection limits is complicated in this study because the concentration of cosmogenic C-14 must decrease with depth in each core. Thus, depending on the exposure time and core length, one might commonly expect to encounter a situation where the C-14 concentration in a core top sample was above some defined detection limit at high confidence, but the C-14 concentration in a core bottom sample was not. The question of whether C-14 is present in a core is different from the question of whether it is above some detection limit in one sample from that core. Thus, in this section we will approach the question of whether measured C-14 concentrations are distinct from background by looking at the entire data set at once, rather than individual samples.

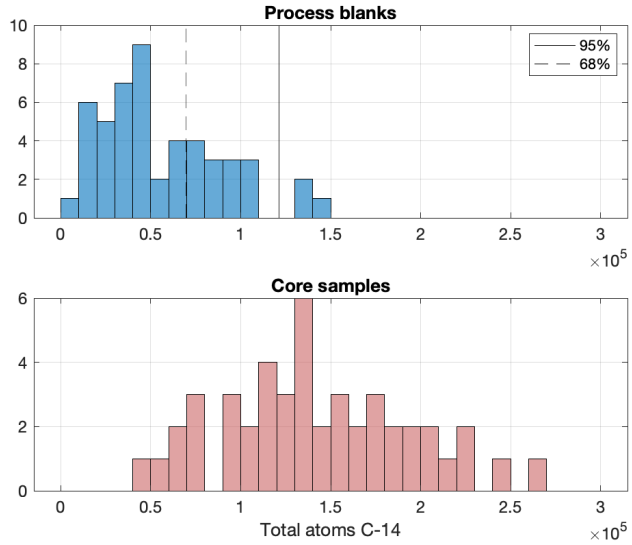

Figure 3. Aggregate distribution of total C-14 atoms observed in all measurement blanks (top panel) compared with distribution of total C-14 atoms observed in all core samples (bottom panel).

First, Figure 3 compares the aggregate distribution of the total amount of C-14 observed in the 2-year series of blanks with the distribution of the total amount of C-14 (total atoms, not atoms/gram) measured in quartz from core samples. Because the core samples are not all expected to have the same amount of C-14, the core sample distribution is not expected to have any particular form. However, it is evident that these are distinct distributions. Using only the observed distribution of blank concentrations and making no assumptions about the form of the distribution, 91% of the sample measurements are above blank at 68% confidence (see Fig. 3), or, 59% of the sample measurements are above blank at 95% confidence. The probability that one could obtain this result by randomly sampling the blank distribution is effectively zero. Therefore, C-14 is present in the core samples in excess of measurement background.

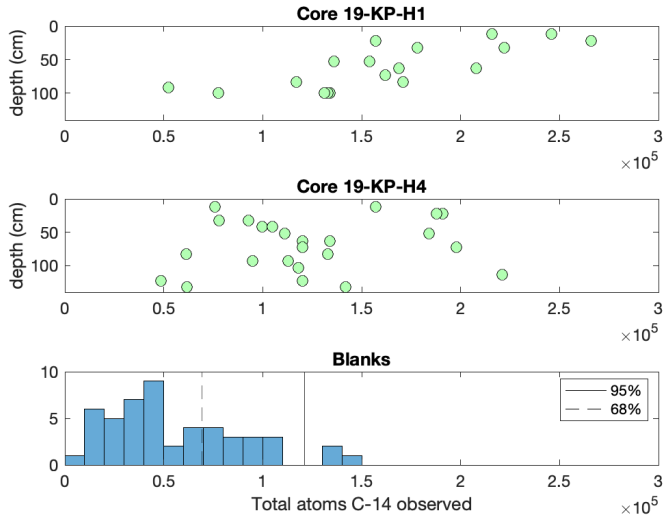

Figure 4. Aggregate distribution of total C-14 atoms observed in all measurement blanks compared with distribution of total C-14 atoms observed in core samples from H1 and H4. These data are the same as in Figure 3, but this presentation highlights that samples from near the top of core H1 are well above background.

Figure 4 shows the same data that are in Figure 3, except by core and in stratigraphic order, which highlights the issue of how to separate the detection limit for C-14 in a sample considered by itself from the detection limit for C-14 in a core. Consider core 19-KP-H1. All replicates of core top samples exceed the 95th percentile of the blank distribution, but some replicates of samples from the bottom of the core are close to the mode of the blank distribution. *However, if cosmogenic C-14 is present in the core top, it*

is physically required that it be present throughout the core, so even if some replicate analyses would be indistinguishable from blank if considered by themselves, the data from core H1 considered together requires that we have measured C-14 above blank. This relationship is not as clear for 19-KP-H4 because there is less variation with depth in this core, but it is still the case that 88% of the measurements from H4 exceed the 67th percentile, and 38% exceed the 95th percentile, of the blank measurements, so it is nearly impossible that we could obtain these results if C-14 was not, in fact, present above measurement background. We conclude that C-14 is present above measurement background.

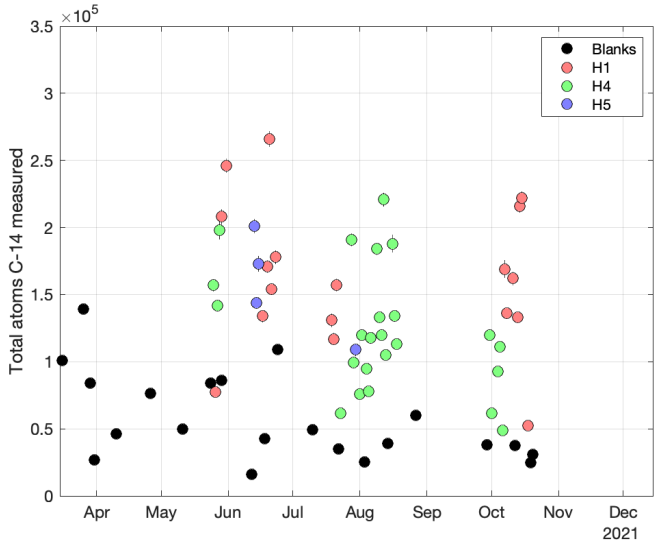

Figure 5. Total C-14 atoms measured in blanks and core samples during core sample measurement in 2021.

Finally, Figure 5 shows total C-14 atoms measured in blanks and core samples during the period of core sample measurement in 2021. If, as the review suggests, a close-in-time blank correction were used, the distribution of concentrations observed in the core samples would be higher in relation to the blank distribution used for correction. Thus, a close-in-time blank correction would increase, not decrease, confidence in the conclusion that C-14 concentrations in core samples are above measurement background.

I.3. Replicate scatter

The final point in the section of the review dealing with C-14 blank handling has to do with the scatter in replicate measurements. For the core samples, scatter among replicate measurements is extremely large in relative terms, exceeding a factor of 2 for some samples. This observation, taken out of context by itself, would tend to lower confidence in the results. However, we investigated this and found that the replicate scatter in the core samples is exactly as we expect from the measured distribution of measurement blanks.

One way to quantify scatter in the core samples is to observe that because all the samples are closely-spaced sections of the same core, their true C-14 concentrations must vary along a smooth curve (which can be reasonably well approximated over a short depth range by an exponential). Thus, all samples from a single core can in effect be considered a large set of replicates, and the expected distribution of a large set of replicate analyses can be inferred from the distribution of the residuals with respect to a best-fitting exponential.

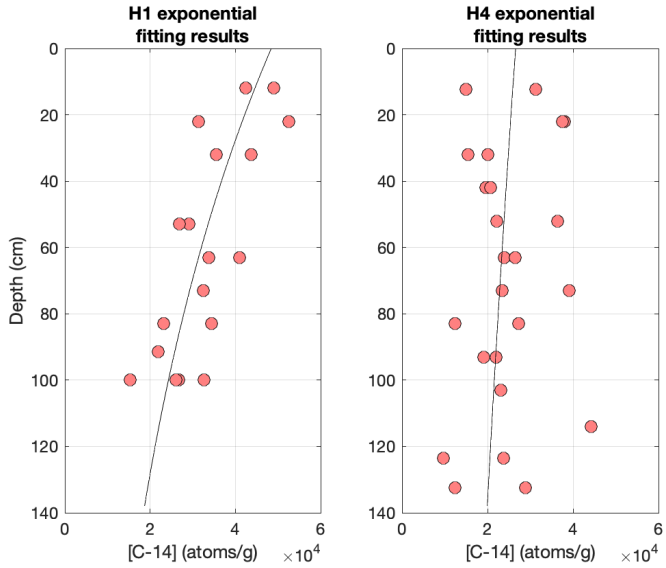

Figure 6. C-14 "concentrations" from cores H1 and H4, with best-fitting exponentials. We quote "concentrations" because these values are the total number of atoms present divided by the mass analyzed, so have units of concentration, but have not been blank-corrected, so are not the true concentrations in the samples. The purpose of converting to concentration units is to normalize for small variations in sample mass before fitting an exponential curve.

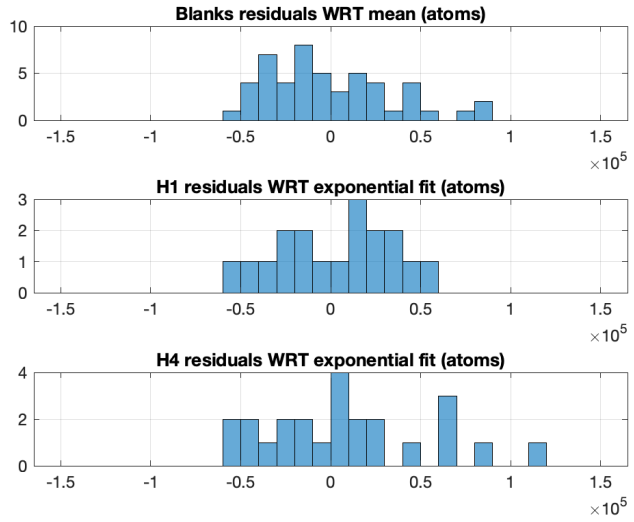

Figure 7. C-14 residuals (atoms) with respect to mean for the aggregate distribution of blanks compared with residuals with respect to a best-fitting exponential for uncorrected concentrations for cores H1 and H4. Residuals computed from Figure 6 have units of atoms/g, but have been transformed back to units of atoms for comparison with the blank distribution. This highlights that the scatter is similar for all three data sets.

Figures 6 and 7 show the results of this exercise. The distributions of the residuals around a smooth curve for both cores H1 and H2 have essentially the same range as the distribution of blanks around their mean. Although there is no reason to believe that any of these distributions are normal in nature (in fact, we propose that a lognormal distribution is most appropriate for the blank correction and use it throughout the paper), one can compare the distributions in a simple way by comparing standard deviations, which are 35000, 32000, and 46000 atoms for the blanks, H1, and H4 respectively. These have similar magnitude (the higher standard deviation for the H4 data is mostly explained by one outlier), which shows that the scatter among replicate analyses of core samples is as expected if scatter in measurement blanks is the dominant contributor.

It is also possible to look at replicate scatter based on replicate analyses of individual core samples alone, without assuming a relation between adjacent core samples. For this experiment we generate a

distribution of replicate differences by compiling all the differences between combinations of replicate analyses of each sample to define a distribution for replicate scatter. We compare this to a synthetic distribution developed from the measured distribution of blanks by choosing a large number of random blank pairs and recording the difference between them.

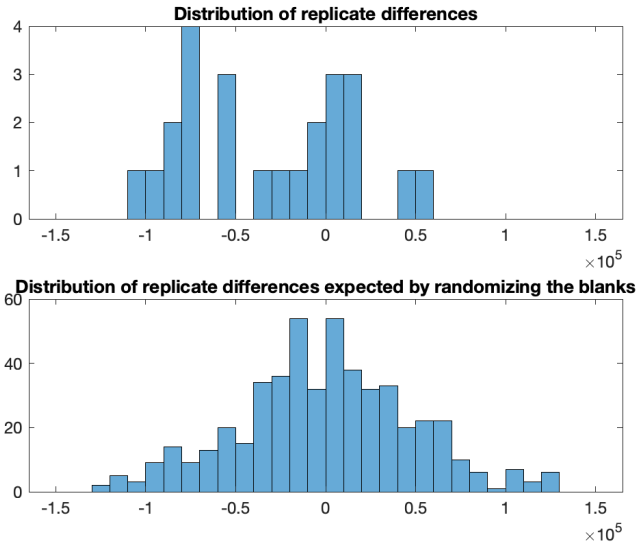

Figure 8. Compiled differences between measured replicates compared with synthetic distribution of expected replicate differences generated by random draws from the distribution of blanks.

Figure 8 shows this experiment. Although the distribution of observed replicate differences uses fewer data than in Figs. 6-7 so is sparser, the conclusion is the same in that the range of the two distributions shown in Fig. 8 is similar. If we approximate these as normal distributions, they would have standard deviations of 46000 and 48000 atoms, respectively, which are the same. Again, the conclusion is that the scatter in the replicate analyses of core samples is exactly as expected if scatter in the measurement blanks is the dominant contributor.

We conclude that replicate scatter in the core samples is indistinguishable from scatter in the aggregate distribution of measurement blanks. *This is exactly as expected, because at low concentrations, measurement uncertainty should be dominated by the uncertainty in the measurement blank.*

This last point brings up another issue mentioned in the review, which is the apparent scatter observed in C-14 concentrations from erratic and bedrock samples above the present ice surface. By "scatter" in this context, we mean the fact that these results do not lie along a smooth age/elevation curve. It is important to note that, in contrast to the sets of samples from bedrock cores, these samples are not physically required to lie along a smooth curve in age-elevation space. This would only be the case given several fairly restrictive assumptions, including (i) a monotonic and smooth ice thinning history, (ii) a common thinning rate of the last several meters of ice overlying each sample during thinning, and (iii) zero shielding or cover by ice, snow, sediment, or other rocks following initial exposure by ice retreat. There is no reason to believe a priori that these assumptions are strictly true. For these data, as pointed out in the review and also in our manuscript, and as is evident by inspection of the uncertainty bars in Figs. 4 and 7 (Figs. 4 and 7 in the paper, not in this response), the scatter around a smooth age-elevation curve is about twice as large as can be accounted for by scatter in measurement blanks. Although there are other sources of measurement error that could potentially contribute (see discussion of CRONUS-A below), we conclude that the samples do not, in fact, lie along a smooth age-elevation curve. As there is no physical expectation that they should, this is not concerning or surprising, and is not evidence that our blank

correction scheme is correct or incorrect. As described in the paper, we attribute this variation around a smooth age-elevation curve to unsteady ice thinning and variations in the size and shape of ice-marginal snow- and icefields during deglaciation.

To summarize this section, we believe we have thoroughly justified our approach to blank handling with the reasoning above. As we suggest above, we think the best place for this highly technical discussion is in the online discussion, not the text of the paper. Thus, we have not proposed any revisions to the paper here. As noted, we have amended the supplementary data so that all the code and data needed to replicate the analysis and figures above is included in MATLAB workbooks.

II. CRONUS-A.

The answer to the main question in this part of the review is yes, the production rate calibration in this study is unchanged from that in previous papers, including among others Goehring et al. (2019) as mentioned in the review. Although additional CRONUS-A measurements have been made at Tulane since that time, they are not significantly different than the data previously used for production rate calibration.

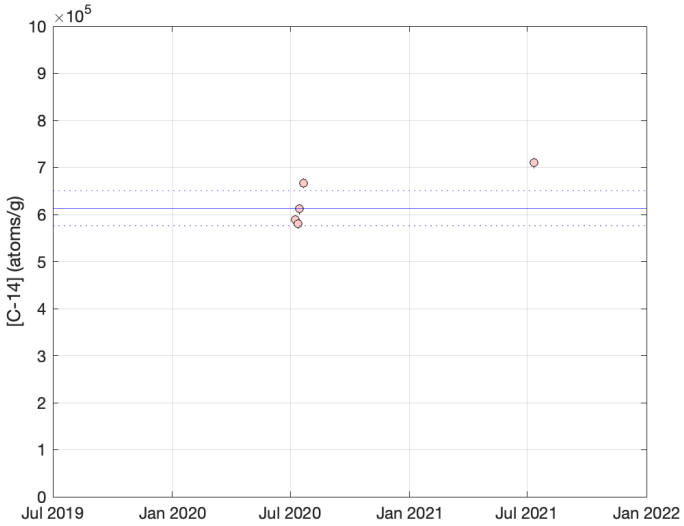

Figure 9. CRONUS-A measurements during the period in which measurements for this study were made. The x-axis in this figure is the same as in Figure 1. The horizontal line with error bounds is the mean and standard deviation of the set of Tulane CRONUS-A measurements made prior to 2019 and used for production rate calibration in this and previous work. Error bars reflecting AMS and blank measurement uncertainty are not visible at this scale.

Figure 9 shows these data. As with all other measurements we have discussed here, scatter in excess of nominal analytical uncertainty is present, and because this sample has a much higher C-14 concentration than the samples measured in this study, this scatter cannot be explained by the distribution of process blanks and is therefore unexplained. However, when these measurements are considered together there is no significant difference from the value that was used for production rate calibration in previous work and in this paper. We did not update the production rate calibration for two reasons. First, on general principles, maintaining continuity in production rate estimates to the extent possible makes it easier for readers, so there is no strong reason to continually update the production rate with new data if the changes implied are not significant. Second, the production rate calibration is minimally to negligibly important in this paper. The reason for that is that the paper focuses on samples from subsurface cores having low C-14 concentrations, so the uncertainty in these measurements is dominated by the uncertainty in the blank subtraction. As discussed above, the uncertainty derived from blank subtraction and expressed as reproducibility of replicates and samples from the same core can be up to a factor of 2. However, the entire range of CRONUS-A concentrations observed in multiple laboratories is about 15%, and the total range in CRONUS-A results from Tulane across several years is about 10%. Furthermore, tf

the production rate used in our calculations varied by 10%, it could result in no more than a 10% difference in the duration of the ice thickness lowstand inferred from model fitting. As the length of the lowstand is only constrained by the model fitting to be in the range of 3000-8000 years, an additional 10% uncertainty is insignificant.

Again, we agree that this is important contextual information for the performance of the Tulane lab in a general sense, but as we have discussed above, this issue is not relevant to any of the conclusions in the paper. Thus, we think this discussion is also sufficient without changes to the main text of the paper. We will amend the supplementary data to include the CRONUS-A measurements made in 2019-20.

III. Comments on model fitting.

In this section of the review, Lifton asks for our opinion on why there appear to be some systematic misfits between model predictions and Be-10 data in the lower portions of cores H4 and H5. Basically, models that fit the entire data set well appear to underpredict Be-10 data in these areas. Although the well-fitting model predictions are not significantly outside the measurement uncertainty for most of the samples considered individually, predictions are systematically below measurements for several samples in the bottom of both cores. We have, in fact, thought about this, and we see two possible reasons. One possible reason is that, for some reason, the model is predicting that ice is too thin during the lowstand period. This could arise from inaccurate estimates of C-14 production by muons. If this were the case, the effective attenuation length of predicted Be-10 concentrations would be too short. However, this is not a great explanation because if this were true, it would imply that the lowstand was longer than predicted by well-fitting models: if we increase the ice thickness to reconcile predicted and observed attenuation lengths, we have to also increase the duration of the lowstand so that we can continue to match the observed concentrations. As the lowstand in the well-fitting models already takes up all the available time permitted by the constraints at each end of the period during which a lowstand could have taken place, it is not actually possible to increase the lowstand duration and the ice thickness to better fit the Be-10 data. Thus, this explanation fails.

A more likely explanation is that our assumption that the model should start with a zero Be-10 concentration is oversimplified. Even in a situation where subglacial erosion takes place during glacial maximum conditions and removes nearly all Be-10 associated with any hypothetical surface exposure in Pleistocene interglacials (as we believe is the case here), the predicted Be-10 concentration is not zero. If we assume that for the past few million years, this site has, on average, been covered by several hundred meters of ice, and also is subglacially eroded by a few meters during every glaciation, millions of years of exposure at a depth of hundreds of meters would still be expected to lead to hundreds to thousands of atoms per gram of Be-10 in production-decay equilibrium with subsurface production by fast muons. For example, Stone et al. (2019) observed 3500 atoms/g Be-10 in a sample 8 m below the surface of bedrock that is covered by 150 m of ice in present interglacial conditions. It is likely that a nonzero amount of such background subsurface Be-10 is present in our cores and not accounted for in our model.

As the presence of background Be-10 is likely from first principles, but the amount is largely unconstrained, it would be possible to add a background Be-10 concentration that was constant throughout the core depth as an additional free parameter in our model. If produced at hundreds of meters depth, it would be essentially constant over a 1.5-m depth range. We experimented with this, and found that a background Be-10 concentration near 1000 atoms/g slightly improved the fit to Be-10 data in cores H4 and H5. For example, for core H4, the best attainable value of M is about 15% lower if a background Be-10 concentration of 1000 atoms/g is included. This experiment can be made with the MATLAB code in the supplemental data by changing the value in line 66 of 'random_search_wrapper.m'.

However, the improvement in the fit was not large, and there is no significant effect on the best-fitting lowstand duration and ice thickness (without background Be-10, models that fit H4 best have lowstands 5000-8000 years long with 5-8 m of ice, whereas with background Be-10, best-fitting models have 6000-8000 years with 6-9 m of ice). Thus, we took this parameter out of our model on the grounds that it added additional complexity but did not change the results or increase the explanatory power of the model. However, we believe that unaccounted-for background Be-10 is probably the most likely explanation for the small misfit.

As in the previous sections of this review, we think this discussion is important for readers who might wish to duplicate the model-fitting calculations using our MATLAB code, but not highly relevant for most readers of the paper. Thus, we again propose that the public online discussion is the appropriate place for this, and propose no revisions to the main text.

4. Minor comments.

This review made 4 minor corrections and suggestions (the 5th replicates material covered above), which we will correct in the revised version.