Response to R2 on:
'*Antarctic ice sheet model comparison with uncurated geological constraints shows that higher spatial resolution improves deglacial reconstructions'*

We thank the reviewer for the supportive comments and helpful suggestions towards increasing the clarity of this technical manuscript to help us reach a wide audience. Below we respond to reviewer suggestions in detail; for typographical errors or small corrections, the * symbol indicates our intention to enact these corrections in a revised manuscript. We appreciate the reviewer's time and insights.

This study by Halberstadt & Balco presents a series of new metrics for paleo data-model comparisons, and applies them to an ensemble of ice sheet model simulations to show that (i) higher resolution nested domains (in their case, 2 km) can improve data-model mismatches, and (ii) there is no "best metric" for data-model comparison, as each metric captures different aspects of ice sheet model behaviour. This often results in a tradeoff between matching the magnitude and the timing of thinning. Furthermore, they show, using their metrics, that the use of an uncurated dataset for model scoring is as constraining as the use of curated data, meaning that it is much easier to include new geologic data as they are published (and also has no subjective/interpretive bias), further improving our ability to constrain ice sheet model ensembles.

This study is relevant to the paleo ice sheet modelling community, both due to the new misfit metrics presented, and to further show within a non-idealised context that higher resolution is needed for ice sheet models to better match constraints of ice sheet geometry and time evolution. The small number of ensemble members and explored parameters is not a problem, as it is justified by their different goals from previous data-model comparisons: not necessarily try to obtain statistically significant distributions of different metrics for AIS behaviour, but rather showcase their new tools and that high-resolution simulations improve data-model mismatches.

Overall, the findings are well worth being published, but the manuscript needs a more thorough, careful revision of the text before I can recommend it for publication. There is a lot of redundant information spread across different sections which read more like a repeat of the text rather than a helpful recap. In other instances, the explanation of the scoring metric feels a bit convoluted, or the justification behind certain ways to circumvent a problem is not super clear. Several figures are not cited anywhere in the text (and I suspect one is missing entirely), and some subsections could be reordered (with slight rephrasing/rearrangement) for a more fluid read. Below I offer some more general suggestions, as well as line-by-line and figure-specific comments.

Thank you for your suggestions to improve the readability (and thus utility) of this work.

General comments

Section 2.3:

- How often is information carried to the nested domains, i.e., what is the coupling time step? Which variables are passed on? How is the downscaling done at the beginning of the high-resolution simulations? Is it just a linear interpolation from the coarser domain?

Added to Sect 2.3: "At the beginning of each high-resolution nested simulation, initial conditions are provided by linearly downscaling the continental simulation across the preceding 2,000 years (32-30ka) before branching off the 30ka-0 ensemble experiments. The nested domains receive 3D ice thickness and ice velocity boundary conditions at the domain edges, provided by the nest-driving continental simulations, and updated every 500 years."

Section 3:

- Section 3.1.1 could be made shorter, and perhaps part (if not all of it) should go to the introduction, which is more fitting for this kind of information. As it stands, it feels like a massive break in reading flow to provide adequate background, which should have been presented earlier. For example, there's overlap on the discussion of coarse model resolution being unable to resolve glaciers where data is collected, which feels more like redundant information than a helpful recap.
Partially moved to Introduction.

- Section 3.2.1: the first paragraph could be more concise, it replicates a lot of information already repeated 2-3 times...
Corrected.

- The description of the metrics could be made clearer:
Also note added Table 1 at the end of Sect 3 that summarizes the various kinds of misfits (from sample misfit to site misfit to model misfit).

- In "float scoring", does it mean that the changes in ice thickness are being compared, rather than in ice surface elevation? Fig. 5 shows "elevation above modern ice", but I it's not completely clear how the modelled change in thickness is translated to surface elevation (is it modern+offset? if so, please explicitly state that)
We simply add a range of vertical offsets (so, yes, modern surface elevation + offset) to find the best-fitting profile. This will be clarified in text, and the Fig. 5 (and subsequent) axis label changed from 'elevation above modern ice' to 'ice surface elevation' to encompass both the original ice surface as well as the vertically-offset ice surface profiles.

- In "best-time-offset metric", it is not stated at any point that the curve you are trying to minimise is the best-fitting curve obtained through the float scoring metric. This is only made clear way further down in the discussion (L865-866). This is important to be explicitly stated here as well.
Corrected.

- ice thickness exceedance: is that simply how much the modelled ice thinned beyond what the cosmo data shows for sites where no Holocene (re)thickening happened?
The ice thickness exceedance metric quantifies how much thicker the LGM ice sheet is modeled, compared to the sites where we can estimate LGM ice thickness from the data.

- I honestly struggle with the fact that the way each scoring is minimised is discussed before the scoring formula is presented. After reading the entire manuscript a second time, I can better see the reason, but only because I had already read the latter sections. Please consider if there is an easier way to introduce all concepts needed in this section, e.g., first introducing the different misfit metrics (3.3.1), then how they are minimised (float scoring, best-time-offset; 3.2.3), then what happens once they fall out of this range (sect 3.3.2). Naturally some of the discussion text around limitations and justifications for the scoring in section 3.3 (e.g., L562-567, Sect. 3.3.2) could be kept where it is, as it is about what was discussed in current Sect 3.2.
Reorganized accordingly – all within Section 3.3 Assessing site misfits – where we now we first introduce the equations for minimizing misfit (3.3.1 Site misfit calculation) and then describe the two metrics (3.3.2 Metrics for model data scoring) and finally what happens once they fall out of this range (3.3.3. Scoring samples outside of the range…). This should also be further clarified by the additional Table 1 that summarizes the different metrics and associated variables, prior to these more detailed sections.

- Section 3.5: part of the information in L641-650 feels more repetition than recapping (as in Sect 3.1.1). I suggest the half-life discussion here and in Sect. 3.1.1 to be put together, maybe in the introduction. This way this section can be made much more concise. The paragraph in L651-663 could also be significantly shortened, considering it is about types of data that were not used.
This section has been shortened.

Section 4:

- The explanation behind variations in OCFACMULT and CSHELF (two ice-shelf related parameters) having a much stronger impact in the model-data fit is quite brief and in my opinion could be better highlighted, as it is an important result. Similarly, I miss a discussion on why TLAPSEPRECIP does not affect the results as much. I would expect this factor to have a stronger impact, considering it directly affects how the ice surface is resolved in the grid cells where the data is present, and potentially by how much the ice sheet interior thickens (or not) during the LGM.

The discussion section briefly mentions the impact of these parameters on overall ice sheet behavior and the corresponding impact model-data fit. We have now added more details: "Models with the best (lowest) float misfit scores tend to produce a greater amplitude of ice thickness change than indicated by the geologic record (e.g., OCFACMULT=5 and CSHELF=7, top left in Fig. 12). Visual inspection of model results reveals that these large amplitudes of model ice thickness change often improve model-data fit by producing delayed and more rapid deglaciation, but these simulations tend to 'over-thicken' relative to the maximum ice thickness constraints (i.e., have higher exceedance misfit scores). This modeled ice sheet behavior can be attributed to the impact of stiffer sliding parameter on the continental shelf (CSHELF=7), supporting larger LGM ice thicknesses, along with heightened sensitivity to ocean melt (OCFACMULT=5), driving rapid ocean-driven grounding-line retreat. Conversely, models with low exceedance misfit scores … The third varied parameter TLAPSEPRECIP influences upstream ice thicknesses but had an inconsistent impact on model-data fit. While constraining paleo accumulation patterns is key for robustly reconstructing deglacial ice sheet dynamics, the TLAPSEPRECIP parameter only scales the precipitation field input to the ice sheet model relative to temperature; future work will explore whether alternate time-evolving precipitation histories can generate a more consistent impact on model-data fit." (Section 5.1)

Line by Line comments

L22: I believe it was supposed to be "simulations" in plural *

L32: It should be more specific what "proximal geologic records" means. It comes up rather early in the manuscript, without providing enough background. Considering this article should be of interest to a wider audience than paleo ice sheet modellers that compare their models with geological constraints, it's worth giving better background to what that refers. *

L133: there's an extra ")" *

L168-170: How is the oxygen isotope curve translated to temperature anomalies? Similarly, are the time slices of Liu et al. linearly interpolated within the model, is it a step-change, or how exactly is it prescribed? I appreciate a lot of technical information on the model might deviate from the manuscript's focus, but it is good to provide a good explanation on how the climate is prescribed to help the discussion about the model limitations leveraged by the time offsets.
This further information has been added (Sect 2.1 Model description)

L172-174: I would assume that the way the atmospheric forcing is prescribed is the same as in Sect 2.1 (see comment above), but that cannot be the case for the ocean. Please make it clearer how the initial conditions were obtained.
This further information has been added (Sect 2.1 Model description)

L178: It would be helpful to mention Fig. 1a here for the reader to be promptly pointed to where these domains are *

L205: It should be made clearer that CSHELF is needed because in paleo simulations the grounding line will likely advance beyond the areas where information can be obtained by inversion/nudging. This is a problem that paleo ice sheet modellers are familiar with, but other ice sheet modellers might not be. *

L230: Please clarify which type of interpolation was used. Was it just a simple linear interpolation? *

L245-246: It would be useful to state the resulting ranges of this approach so it can be more easily compared with the other methodologies described. *

L264-269: I am confused here. Is amplitude of thinning the thickness difference or magnitude of thinning rates? Are the "thinning patterns" the magnitude of thinning rates or the shape of the thickness-evolution curve? Is the dataset for both metrics totally different, or is just the same (Ice-D), but processed in a different way to allow computing the desired metric?
Rephrased here: "In addition to model-data fit with respect to ice thinning patterns at a site, we also investigate model-data fit with respect to the maximum amount of ice sheet thinning during the last deglaciation (i.e., LGM ice surface relative to modern ice surface)."

L344-349: I am not sure I follow the procedure described. I don't think a "steepest descent scheme" is a familiar term to all target readers of this manuscript, and hence more details should be added. *

L382-387: When comparing the text with Fig. 2d, the smaller error bars obtained from the MC approach make 1 or 2 accepted samples to lie outside of the derived interval. Is that correct? Does that mean that they are then also discarded? It is not clear to me how the obtained error/uncertainty is used.
Yes, that is correct, but these samples are not discarded. We use the MC scheme in two ways: First, *any* samples that lie along the steepest-descent path during *any* MC iteration are used in our model-data comparison. Second, for each sample, we identify the 1-sigma error *at that elevation* based on the MC envelope. So, for example, the sample at about 850m in Fig 2d (third from the top) is included as a youngest-bounding sample going forward, because it was selected by at least one MC iteration of steepest-descent path, and its MC-based error is still calculated based on the width of the MC 1-sigma envelope (distance between the two red dashed lines) at that elevation, even though the sample itself doesn't fall within the MC 1-sigma envelope.

L391: I assume this a different vertical window than discussed before in L245-246. Is there another term or an adjective that could be used for either to avoid confusion?
We now rephrase this as 'tolerance range' instead of 'vertical window' to avoid confusion.

L401-402: Is that the age spread in the 'MC cloud' at each sample elevation? It is not quite clear how the number "500" was obtained, as the way it reads it feels rather subjective.
It was previously rounded up; we now specify '495 years'

L437-442: At which time step/slice is that search done? From the explanation of the reasoning behind the search, it could either be at every step (since the base of the hillside would move as the ice retreats) or at present (the 'lowest base possible').
The maximum velocity location is selected across all possible timesteps, so that the grid cell in question is most likely to always reflect the adjacent glacier.

L455-456: Wouldn't a site, by definition, not span different model grid cells? I am confused what this sentence is trying to say.
Rephrased: "If the transect of samples comprising a single site spans different model grid cells, we use the model thickness history corresponding to each sample."

L485-486: Do I understand correctly that the thickness-evolution curve is added on top of the sample-derived modern ice surface, and then offsets are applied until the minimum misfit is found? This needs to be clarified.
Yes, that is correct. This sentence is reworded.

L575-579: Were there any sensitivity tests performed to show to which extent this happens, and to support the choice for 10kyr? I would assume so, but it would be good to explicitly say/show the numbers.
This uniform time gap is inherently arbitrary; it reflects users' judgments regarding the importance of a model capturing the full extent of thinning. Since this user preference will probably vary, based on the motivating question that underlies each model-data comparison endeavor, we suggest a 'default' value of 10kyr (which reflects our own preference that a model should span most of the deglacial thinning at a site, but it's not critical for a good model to span the entirety of the site's thinning amplitude at the cost of rewarding 'overthickening'). This time offset value is intended to be adjusted by a user, and is therefore a required input within our user interface GHub tool.

L605-608: Please provide the formula for S_w, since it is explicitly used in Eq. (5). *

L615: I assume "all site misfits" is the n in Eq. (5)? Please make it clear. *

L674-676: does distinguishable/indistinguishable refers to whether their nuclide concentration lying below the steady-state value considering the standard deviation? I honestly appreciate that there is a lot of care not to use overly technical jargon, but in some situations it might obscure the real meaning of a sentence.
We agree that 'indistinguishable' here is vague. Specifically, we consider a sample saturated if the measured C-14 concentration is higher than the steady-state value, without considering the uncertainty. This is effectively equivalent to saying that the sample is saturated at 50% confidence. Arguments could be made for either a less restrictive criterion (e.g., up to one standard deviation above the steady-state value, or equivalently not possible to prove that it is saturated at 84% confidence) or a more restrictive one (e.g., less than one standard deviation below the steady-state value, or equivalently not possible

to prove that it is below saturation at 84% confidence). However, we have not been able to come up with any clear reasoning in favor of either of these, so we used the simple 50%-confidence approach. However, the algorithm for using these data to estimate the LGM ice surface elevation is slightly more sophisticated in that it only accepts a saturated sample at a higher elevation than a sample with a post-LGM exposure age as LGM ice thickness constraints. That is, a sample with a finite-at-face-value exposure age older than 20 ka, which would commonly be within uncertainty of the steady-state concentration at typical measurement uncertainties, is not counted as either a saturated sample or a post-LGM sample. We have clarified this section of the text.

L699-700: Wouldn't these events be less likely to deposit erratics at all? Why "post-LGM-age erratics" only?
Yes, 'post-LGM' is redundant. The point here is only that the likelihood of an erratic being deposited is probably proportional to the duration of ice cover, so short periods of cover have low likelihood of being recorded. If the ice cover was at the LGM , then the exposure age of any erratic emplaced would be 'post-LGM' by definition. We have corrected this.

L702: Can you be more specific regarding how Slgm is calculated between 0 and 1?
Added and adjusted this sentence "…we therefore assign each site a weight (Slgm, calculated from the number of samples at the site, and normalized across all sites in the dataset to range from 0 to 1)."

L708: I am not sure I follow the argument, please rephrase/expand. *

L712-715: Is the "present reference" in the modelled thickness above present based on the model's state at the end of the simulation, or based on the present-day dataset? I would assume this is a similar problem than the "float scoring" tried to circumvent? Or are elevation and thickness changes being used interchangeably here? Or is "thickness" not the full thickness of the ice sheet, but the difference in thickness/elevation? This part is getting me confused!!
We agree this is confusing as written. What we are comparing here is the *ice thickness change* inferred from the exposure-age data (which is the elevation of the sample less the elevation of the modern ice margin near the sample) with ice thickness change in the model. The absolute elevation is not involved, and using 'elevation' is in fact misleading. We will clarify this in the revision.

751-756: This information has already been given previously, so the paragraph could be condensed to 1-2 sentences for conciseness. *

L763: This writing is odd, I'd suggest something like "Fig. 11 also shows the impact of their tier designations on our scoring". *

L776-784: Most of this information was already given in previous sections, so this paragraph could be significantly shortened. *

L785-788: I like the way that the technical part is summarised in less technical terms: "domains are less wrong" and "are wrong more consistently". I think, however, that you do not need to explain it straight away within parentheses what it means, because the following paragraphs do pretty much that, using almost the exact same words. Recapping information is good, but when sentences are so close to each other it feels too repetitive...
We agree, but Reviewer 2 remained confused about this particular phrasing, so we err on the side of over (rather than under) explaining.

L806: The mention of panels d,e in Fig. 13 comes before panels a,b,c. I'd recommend reordering the figure panels (i.e., panels d,e should be a,b and panels a,b,c should be c,d,e) *

L806-813: Unless I misunderstood the point here, I am not sure I fully agree with it. This might be true if the different sites are relatively close to each other, but, in line with the discussions in Section 5.2, there's no guarantee that asynchronous changes in the atmosphere and the ocean did not happen around the ice sheet, meaning that different regions would have experienced an earlier/later deglaciation, or that certain basins would have responded faster/slower to a given forcing.
Agreed; rephrased this sentence to emphasize that we are considering time-offset consistency by catchment, rather than across the continent, for precisely this reason; we would expect a systematic time lag in our model's ocean or climate forcing appear at the catchment scale, but not necessarily at the continent scale.

L829: It would be easier to follow what a best-time-offset outlier is if it was first defined in the previous paragraph (suggestion, around L 822), so you wouldn't need to add this "i.e." parenthesis. *
Also, try to be consistent with best-time-offset or best-offset, as it is not clear whether they were supposed to be the same or not (if not, please make sure they are both clearly defined). *

L882: simulations -> simulation *

L928: The use of "modern model timestep" here can be ambiguous, I'd suggest something like "modern time slice" or just "modern time". *

L930-931: This information could come in the introduction already, as it is not quite dependent of the results, and yet might appease the skeptical reader sooner rather than too late.
Thanks for the suggestion; we will move this to the introduction.

L961-964: Although the argument makes a lot of sense, I can only see the mentioned trend on East Antarctica (EAIS+TAM), as opposed to being "most prominent". It is perhaps because of the clustering of points, but this trend is not visible at all for WAIS or Weddell. There are both too-early and too-late points with just as low low exceedance scores in both regions. If the clustering does not allow for this relationship to be properly observed in the figure, then the axis needs to be expanded.
Rephrased to better describe our observations of this figure: "…we observe an overall

trend, mostly in the Transantarctic and EAIS regions, where the model simulations that tend to deglaciate too early (reds in Fig. 15) also have better (lower) exceedance misfit scores."

L969-971: Some punctuation and rephrasing is needed for clarity. Suggestion: "model member 7-2-35 overthickens (...), simulating a delayed timing of thinning compared to the 5-0.3-35 model, which shows a smaller amplitude change and an earlier onset of thinning". *

L978: Unclear what a "a lingering disconnect" refers to, especially because, without being sure what that means, the following sentence seems to contradict what was just said.
Rephrased in an attempt to simplify and clarify: "In some Antarctic regions (e.g., WAIS, and Weddell, although there are limited data in these areas), model simulations can approach a perfect best-time-offset float misfit score without over-thickening; but in the Transantarctic Mountains, no one model simulation is able to produce both a low float misfit score and a low exceedance misfit score (even after eliminating any impact of a timing mismatch)."

Figures

I miss a figure that shows the name and location of sites that are mentioned in the manuscript, as this information does not exist anywhere. I even suspect it was accidentally left out, considering Fig. 16 mentions this information exists in "Fig. 0".
This will be added to Figure 1.

It would also be useful to have a figure showing how wide the range of modelled deglacial behaviours actually is (e.g., changes in grounded ice volume through time), so it can be compared with the other examples cited by the authors.
Agreed; we will consider adding a panel to Fig 12 (showing the modeled range in total ice volume change across the ensemble), rather than adding a new figure to our existing 16 figures.

Fig 2: Caption of panel (c) says 'red', but at least to my eyes it is orange. *

Fig. 4: The difference in ice surface elevation between the 2 and 20km domains before the main thinning event (panel a) is quite striking. Considering the 2km domain uses the 20km as initial conditions, is there a "relaxation period" at some point between 30 ka and ~19ka, a previous thinning event, or why is the discrepancy between the two curves so large?
The nested domains are already 'relaxed' from 32-20ka before branching off the ensemble simulations; here; the big difference between the thinning curves is that the 20km grid cell that corresponds to this site location has a different bed elevation (that 20km grid is sampling across glacier trough and mountain peak), whereas the 2km grid cell is sampling just the glacier hillside location (or, just the glacier trough, with the velocity search-around methodology). So the total amplitude of ice thickness change is different, since the ice sheet

surface elevation is similar between the continental and nested grid domain.
In a revised manuscript, we will take out the continental curve, since this figure is intended to highlight our new methodology of identifying the appropriate location for extracting the model ice thinning profile (i.e., the glacier adjacent to the site).

Fig. 6: Is the zero-offset curve in Fig. 6a the same as the black line in Fig. 5a, as the caption implies? They look very different to me before 15 ka. Please double check/clarify.
Yes; we had cropped out the 'thickening' portion of the curve (i.e., ice growth prior to the ice thickness max at about 10ka) but only for Fig. 6 and not for Fig. 5a. We will add this cropped bit back to Fig. 6 so that they are consistent.

Fig. 8: The axes labels are too small. Please enlarge *

Fig. 13: The meaning of the red highlight on the bars in panels d,e is not described in the caption. Do these bars represent the points shown in panel (a)? Also, is there a more helpful name to refer to than "best offset"? It is not clear whether that is the "smallest needed offset", or something else. It would be helpful to more clearly relate to the scoring metric that was used.
Figure caption now describes that the red highlighted bars indicate time-offset outliers (now more explicitly described in the preceding main text).
The new table at the end of Section 3 that describes the different metrics and associated variables (along with a consistent terminology of 'best time offsets') should clarify this.

Fig. 16: Caption refers to "Fig. 0", and I believe this is a figure that is missing in the manuscript because I'd suggest adding a figure that shows the site names discussed in different figures. Also, the grounding-line contour is not visible.
We will ensure that site names mentioned in text are all denoted in Fig 1, and sites plotted elsewhere (as in Fig. 16) are either denoted spatially (we will add these locations to Fig 16g panels) or also denoted in Fig. 1.

Figures 3, 8, and 9 are never referenced in the text. Fig. 10 is only referenced in the caption of Fig. 16.
We will ensure that figures are referenced appropriately and sequentially in the main text