Response to R1 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. 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 towards improving the clarity and utility of our work.

The authors present a new automated workflow for model-data comparison using the Penn State University ice sheet model and the ICE-D surface exposure age database. The paper goes in depth about different approaches and metrics for comparing model simulations to spatially and temporally sparse data, with a thoughtful treatment of uncertainties (model, analytical, and geological). This work is impressive and well suited for publication in The Cryosphere after some revision.

My largest concern with this manuscript is also the easiest to fix: the original sources of the data shown in figures need to be cited, not just the ICE-D database. The way this database is currently used means that Balco (2020) gets cited instead of the original data sources. For instance, the six sites shown in Fig 16 a–f span four publications, but only two of those publications are cited, and those not in relation to the data. Surely there is a way to easily compile the necessary citations when pulling data from ICE-D.

We completely agree; this same concern also came up 'offline' of the review process within our community. We will include the relevant citations in figure captions in a revised manuscript.

Because the paper is very long and dense, it would benefit from some significant revision for clarity. For instance, as you'll see from my specific comments below, I was confused for a long time about the different metrics being described, how they were used, and how they did or did not interact with each other. It became more clear in the Results and Discussion sections, but it made the Methods section very difficult to get through. I suggest adding a more clear and thorough roadmap of that section before getting into the details.

Following this suggestion, we will flesh out our summary paragraph in the Introduction section into a more complete 'roadmap'. This paragraph will now read:

"In this paper, we describe the development and deployment of our model-data comparison toolkit. We first discuss the modeling techniques that we apply towards this goal, including the underlying choices that shape the simulation and extraction of model deglacial history for comparison with geologic data (Section 2). In Section 3, we describe the geologic dataset

used here – cosmogenic nuclide measurements – and the acquisition and processing of raw data. We then introduce our automated methodology for identifying youngest age-elevation bounding samples and discuss our treatment of sample uncertainty. Having extracted analogous thinning profiles from both geologic datasets and model simulations, we develop and present two key metrics that capture different aspects of model-data fit (Section 4.1). We investigate the impacts of our methodological choices, and describe our formulation of model data fit (Section 4.2). We also compile a secondary set of cosmogenic nuclide model constraints on the maximum ice thickness achieved during the last glacial cycle, comprised of sites where exposure age measurements bracket the local last-glacial-cycle ice thickness change; we correspondingly develop an independent model-data metric to evaluate the modeled amplitude of ice thickness changes (Section 4.3).

With these scoring techniques in hand, we apply our model-data framework to the small ensemble of numerical ice-sheet model simulations (Section 5). Model scores using our uncurated and automated sample selection methods are compared to model scores using a recent comprehensive curated dataset (Section 5.1) to ensure that we have not introduced any systematic failures in our approach by eliminating manual curation. We also investigate the impact of model grid resolution on model-data fit by comparing results from continental simulations to nested high-resolution model domains (Section 5.2), showing that these high-resolution nested domains indeed improve model representation of ice thinning patterns across mountainous terrestrial regions where exposure-age data are often collected. In Section 6, we synthesize and interpret our multiple metrics for model-data fit across the deglacial model ensemble, providing scaffolding to leverage this suite of model-data evaluation tools in various ways to address different questions about deglacial ice sheet behavior."

We will also make clarifying edits to many of the 'roadmap' paragraphs that can be found the start of each manuscript sections, that are intended to help signpost and summarize each section and guide the reader.

We also make some adjustments to the headings and numbering organization, for additional clarity.

I have also pointed out some places in the detailed comments below where aspects of the methodology seem arbitrary or not well supported by the data, or where such complete automation could be undesirable. Obviously it would be absurd to rely on careful manual curation of an ever-growing dataset forever, but the pitfalls of automation need to be clearly addressed. The manuscript does a fairly good job at this in various places, but I think it would benefit from a dedicated sub-section of the Discussion that goes further into these considerations beyond the comparison in Fig 11.

We will add to the paragraph explicitly addressing the pitfalls of automation to the 'Comparison with a curated dataset' section:

"Our model-data evaluation approach is designed to incorporate all available geologic data, with two main aims: to easily assimilate new datasets as they are collected; and to avoid potential interpretation bias. As many studies rely on expert assessment to interpret individual samples and their geologic context, removing manual curation will also entail the loss of this expert judgment. It is likely, therefore, this uncurated approach entrains a number of spurious or erroneous data that might have been removed by manual curation but not by automated processing, or misses key elements at a complex site. (For example, the exposure-age record at Diamond Hill is complex, with multiple nuclides and sample types, and our automated methodology generates a greatly simplified thinning history (Fig. 16a) compared to the in-depth analysis by Hillebrand et al., 2021)). Acknowledging the loss of site-specific geologic context, but balancing the previously described benefits of an uncurated approach, we endeavour to assess whether there is any significant effect of curation on model evaluation. Specifically, to assess the impact of data curation, we compare model-data misfit scores using our uncurated and inclusive ICE-D collection compared to a curated dataset of geologic constraints (Lecavalier et al., 2023)."

Specific comments:

In general, locations of sample sites shown in figures need to be marked on a map somewhere. For instance, site KRING in Fig 8 is not shown on any location map.
Figure 1a now contains additional labels to denote sites discussed or shown in the manuscript.

While automation and the use of uncurated data certainly has obvious benefits, the cosmogenic nuclide record is often ambiguous and frequently relies on expert assessment of individual samples and their geologic context. I worry that automating the model-data comparison process will come with a loss of expert judgment. For instance, the exposure-age record at Diamond Hill is quite complex and uses multiple nuclides (Be-10 and C-14) and sample types (erratics and bedrock), but Fig 16a is missing a key in-situ C-14-saturated sample and shows a greatly simplified version of the history that does not agree with the preferred ice history suggested by flow-band modeling (Hillebrand et al., 2021). I'm not insisting that the preferred history must be correct but instead pointing out how much nuance is lost in the automated approach.
Included in the added paragraph above ('Comparison with a curated dataset' section)

L88–100. You could use the modern ice surface elevation at each site as a constraint, rather than a complete spatial map. This is worth considering adding to your analysis, especially given that many of the later figures in the paper show poor fit to present day ice thickness at the end of the model runs and this is essentially a free data point. And if one reason to compare model results with paleo-constraints is to calibrate a model for future projections, getting the modern state right is very important.
Yes, we considered this approach, but decided that the uncertainties with respect to the

measured/modeled modern ice surface were too intractable for our methodology – specifically, uncertainties associated with (1) establishing a modern surface elevation relative to exposure ages (the ice surface is heterogeneous so where do you register a mountain peak transect to as a baseline? The adjacent glacier surface? How far upstream?) and (2) the modern model snapshot (prone to various confounding resolution issues, discussed further in Sect. 3.2.2.), as well as (3) aligning the two quantities. This decision is outlined in Section 3.2.2. Model-data alignment with respect to a common ice surface baseline

L102: Clarify that this ~10–20km resolution only holds for paleo simulations. Plenty of models use much higher resolution than this for shorter-term simulations. *

L 163: Is "marine ice *shelf* instabilities" intentional phrasing? **Yes; referring to MICI: Marine Ice Shelf Instability feedback.**

L 167–170: All of these data sets (with the possible exception of Liu et al., 2009) are pretty seriously outdated at this point. Has any attempt been made to update these? Is there evidence that the model is insensitive to the choice of these datasets?
Inaccuracies in model input climatology and forcing datasets do contribute significant uncertainty to any model simulation, especially, as noted, for older generation data products. However, we generally assume that any inaccuracies in older-generation model inputs are likely subsumed by much larger uncertainties associated with extrapolating modern datasets back in time, although future work will explore the impact of tweaking model input datasets on simulated deglacial behavior. However, that planned future endeavor is out of the scope of this particular work, which focuses on frameworks for model/data comparison.

Section 2.2 needs more information:

What data sets are used to define present day ice thickness and bed topography? *

How does the basal friction inversion work? *

Is there also an optimized ice stiffness field?

How is the present-day temperature state achieved and what geothermal flux product is used?

What spatial resolution? Has any mesh convergence testing been done?

What sub-shelf melt and calving schemes are used? *

This '2.2 Initialization' section has been expanded to include these details as relevant to the work presented in this manuscript (for example, the basal friction inversion treatment, bed topography, and sub-shelf melt scheme do impact our results and ought to be mentioned; but our model does not have an optimized ice stiffness field, and the input GHF flux dataset

doesn't impact deglaciation significantly; DeConto & Pollard 2012). Nor does present-day temperature / ice thickness impact our results, since we conduct a full-glacial-cycle paleo simulation to provide last-glacial-maximum (30ka) initial conditions.

L 182: "The model basal inversion slipperiness input is downscaled at a correspondingly high resolution for each nested domain." Does this just mean it is interpolated from the coarse resolution, or is the inversion done on the nested domain or some other high resolution continental domain? For most of the outlet glaciers, the coarse resolution inversion is probably meaningless, since even Byrd Glacier is at best going to be a few grid cells wide at best.
"*For most of the outlet glaciers, the coarse resolution inversion is probably meaningless"*
Exactly! Yes, we re-did the inversion across the nested domain; this has been re-phrased accordingly.

"Nested simulations have been shown to be resolution-independent." Numerically speaking, this cannot be true in general. The figure referenced from DeConto et al (2021) shows that the model is more or less converged with respect to resolution at 10km grid spacing *for that particular domain and scenario.* This will not necessarily hold for other applications.
Rephrased to "DeConto et al. (2021) demonstrated resolution-independence of nested simulations (Extended Data 5g); here we additionally test different resolutions …"

Is there some relaxation period used to let the model adjust to the nesting resolution? Going to higher resolution usually means the model needs time to adjust, which can take thousands of years.
Yes; "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…"

Fig 16 caption references Fig 0, which does not exist. *

L261: Maybe this is made clear later on, but from this text it sounds like only ice thickness change (and not the actual value of ice thickness or surface elevation) is evaluated here. I understand that it's often very tricky to get models to match absolute surface elevation, but you could have an excellent model agreement to observed ice thickness change and still be extremely far off in terms of the ice surface elevation. For many cases, this would make it seem like the model is doing a good job of explaining the data, when in fact it is biased extremely high or low. Update: Okay, I see that you also have this exceedance metric later on. It would be good to

mention it here.
Yes, this section introduces the exceedance metric.

L330: What do you define as LGM age here? Depending on the production rate scaling scheme, even using 30ka would prevent C-14-saturated samples from being included. This also brings up another symptom of automation, which is that pre-exposed erratics can still provide a constraint on LGM surface elevation even if they do not provide a good constraint on the timing, but those are discounted here..
Yes, these pre-exposed erratics cannot constrain the timing of thinning so to construct our primary dataset (to reconstruct thinning profiles), our pre-processing step eliminates these old ages (greater than 40ka).
However, as you mention here, we seek out these pre-exposed erratics for building our exceedance metric (in our secondary dataset). So they are not discounted, but leveraged for their strength (i.e., constraining LGM surface elevation, even though they cannot help us construct timing of thinning).

L335: What data product do you use to define the modern ice surface? Also, do the sites have to be strictly within a model cell? Due to the low model resolution, I can imagine cases where the sample sites are in a cell that is always ice free, but could still provide constraints on an adjacent cell that contains ice. This could happen even at the 2km resolution for some of the smaller TAM outlets.
A modern ice surface elevation for each sample is recorded in the ICE-D database, although is somewhat subjective based on the sample collection field campaign.

What are the numbers of lumped sites for the different resolutions?
This is included in the next section: "Note that these 58 sites are further reduced to 50 (44) locations for evaluating 2-km (40-km) model simulations because sites that fall into the same model grid cell are lumped together, as described above."

L375: Needs some example citations *

L 400: It is starting to occur to me that there should be a summary table or list at the beginning of section 3 that briefly covers all the different metrics used.
From this and subsequent comments, we add a table (Table 1) at the end of Section 3 (in a new section named Description of terrestrial model-data comparison techniques):

Table 1. Description of terms and misfits.

Variable		Description	'Float scoring' approach: Model thinning curve at each site is 'floated' vertically (in elevation) to minimize site misfit	'Time offset' approach Model thinning curve at each site is 'floated' horizontally (in time) to minimize site misfit

m_{samp}	Eq (1)	Sample misfit (mean squared error)	Recalculated for every model thinning curve with an applied vertical (elevation) shift	Recalculated for every model thinning curve with an applied horizontal (time) shift
m_{site}	Eq (4)	Site misfit (sum of sample misfits)	The 'float misfit' site score is the minimum m_{site} that can be produced by 'floating' the model thinning curve	The minimum m_{site} is identified by applying horizontal shifts (time offsets) to the model thinning curve
		Best time offset	n/a	The 'best time offset' is the value, in kyr, of the horizontal shift (time offset) that minimizes m_{site}
M_{model}	Eq (5)	Model misfit with respect to thinning curve	The 'float misfit' model score is computed by summing each 'float misfit' site score (minimum m_{site})	n/a
M_{exceed}	Eq (7)	Model misfit with respect to maximum ice thickness change (the 'exceedance score')	n/a	n/a

S 3.1.6 Surely there is a more rigorous way to estimate the minimum geologic error? It seems strange to use this relatively sophisticated Monte Carlo analysis and then eyeball the dividing line for the tail of the distribution. Perhaps using the 95th percentile value as the cutoff? I know it's unlikely to change the number much, but this would be preferable in terms of reducing interpretation bias and leaving the workflow flexible as new samples are added to the database.
Yes; we originally used the 95th percentile value of 495, but rounded up to 500; however, we agree that maintaining precision would better fit with our goal of transparency and flexibility. Corrected.

Also, is there a strong justification for making this a uniform value, rather than varying it spatially by sector, for instance?
Our justification is primarily simplicity, to avoid adding yet another complicated workflow to an already extremely dense and technical workflow. Additionally, making it spatially variable would invite issues associated with large differences in data density across regions, and would necessitate a geological or glaciological reason to explain that regional variability. This geologic error is likely related to local topography or snow conditions.

L423: needs references for " Previous work has relied on continental-scale ice sheet models at 20-40km". In general, all of section 3 would benefit from more references. *

Fig 4 legend is missing the grey curve for the first 2km entry. The box on the continental map showing the location of b–e is hard to see. I suggest making it a bright color that contrasts better with the ice thickness map (or simply removing the ice thickness from this plot and just showing the grounding line or continental outline). *

L472–474: Possibly as (or maybe more) important than the ice-flow perturbation is the formation of blue ice areas and wind scoops on the down-wind side. See figure 4 in Bintanja (1999), for instance. *

L476: I'm not quite sure what is meant by "due to parameter variation" here. There are many factors aside from parameter uncertainty that play into this: forcing uncertainty, initial condition uncertainty, structural uncertainty, etc. You could probably just delete that phrase, as it's fairly obvious why it is hard to match the modern state at the end of a simulation lasting tens of thousands of years. I would also change "may not" to "is very unlikely to" or even "will not". *

L487: Does the float-scoring metric really account for time? If so, it seems that time is double counted when also applying the time-offset metric.
Yes, a model will have a poor float score metric if it deglaciates at the wrong time, since no matter how much the model thinning curve 'floats' up or down in elevation space, it will still be far away from the data samples. The time offset metric is another way of characterizing model-data fit, by allowing the model thinning profile to shift in time and thus scoring the model only based on its thinning shape.

L500: I don't understand this sentence: "The timing of model thinning should be generally insensitive to model-data alignment issues." Can you reword for clarity? By "alignment", are you referring to the elevation misfit? It could also be interpreted as general alignment along space and/or time dimensions, which makes the sentence rather ambiguous.
Sentence was removed for clarity and flow.

L501: The parenthetical "horizontal" here is a bit strange, since horizontal by definition refers to space, not time. Specify that this is along the horizontal axis in the age-elevation space shown in Fig 6, for instance. *

L506: There could also be random errors in the forcing datasets accounting for this. Not all forcing uncertainty is going to be systematic. Conversely, model resolution issues could certainly impart a systematic bias. For instance, low resolution model simulations tend to exhibit marine-ice-sheet-instability-style collapse more readily than high resolution simulations. If the model is not converged with respect to resolution (as I would bet is the case for 40km simulations), there would be a systematic bias across the entire West Antarctic Ice Sheet during a deglaciation event.
Agreed; sentence rephased to allow for broader implications.

Based on Figure 6, it looks like you apply a dozen or so equally spaced offset values and then score them to find the best of those user-defined values. Presumably there is a straightforward way to solve this as a minimization problem, which would be more accurate and robust and

make the workflow more automated.
Agreed; however, this would add another layer of workflow complexity. Technically we are already treating this as a minimization problem, just a fairly basic one.

Also, panel b would make more sense if the axes were flipped, since the offset is the independent variable here. *

S3.2.3: Are these metrics weighted equally? Can you do some kind of L-curve analysis to determine the optimal weights?
We don't combine these metrics together into one total score, because our metrics are aimed towards answering different questions about model data fit. The 'optimal' weights for combining metrics into one total score would differ based on the user's interest.

L529: If I understand this correctly, using "the closest time of exposure of the data point" means that the interpretation of the exposure age is dependent on the model prediction, which doesn't seem appropriate here. It also suggests that any age within the uncertainty bounds is equally likely, which is not the case.
If the modeled age falls anywhere within the sample uncertainty range, we consider that model prediction correct.

L~565: By this point, I'm fairly confused. What was the point of the metrics in 3.2.3 if you instead use this misfit metric in equation 1? I think this section needs to start with a more thorough roadmap that explains which metrics are needed and what they are used for.
The new Table 1 (at the end of Sect 3) should clarify how the misfits all fit together. Basically, this site misfit *is* the value that the float scoring metric and time offset metric seek to minimize, by shifting the model thinning curve relative to the data thinning curve (which correspondingly changes the site misfit score).

3.3.2 title: Presumably you want to score the model, not the samples? *

L583: Why not use a similar 10kyr value for this case? The way it is done here makes the two youngest samples in Fig 7 have significantly less weight than the oldest sample, even though the model fits them all very poorly.
True; but conversely, a 10kyr 'penalty' would give these samples an oversize impact on the site misfit. So either choice propagates some interpretative bias into this workflow, despite our efforts to eliminate this bias through automation. However, we do allow users to adapt these 'preset' values in our user interface Ghub tool.

Also, it seems like the choice of the 10% ice-thickness-change vertical window could have a very large impact on scoring. If that window was just a very tiny bit wider, both the oldest and the second-youngest samples in Fig 7 would be interpreted by the algorithm as having relatively small Δt values. How was this 10% window determined? Instead of using a window of uniform

height for each site, what if the algorithm could solve for the necessary window height and then account for that in the misfit calculation?
Yes, this is another choice that could impact results, although the algorithm does individually solve for the vertical window height at every site (based on 10% of the modeled maximum ice thickness change, which varies site-by-site).

Also, should this reference Fig 7 instead of Fig 5? **Fig 7 correctly references Fig. 5's best-fit curve.**

Fig 8: The top panel shows a misfit of 0 even though it underestimates modern ice thickness by 50m (out of only 350m total thickness change) and is still thinning at a significant rate at the end of the simulation. See my comment above about including the modern ice thickness as a constraint.
The motivating goal of the float-score approach is to avoid issues with the modern ice thickness constraint (see response above)

This misfit cap of 50 seems quite arbitrary as well. How was this determined? Can you show at least two other sites to show that this value is representative? This figure doesn't necessarily support the cap of 50. For instance, the difference between the misfits of 46 and 55 is not all that significant: they both thin too little, too late, and the difference is just a matter of degree. And more importantly, the run with misfit 55 thins at about the right time and is in fact significantly better than the run with misfit 120, which bears almost no resemblance to the data whatsoever. So I think the limit of 50 is actually a bit too restrictive here. One could also argue that the runs with misfits of 46 and 55 both explain the data better in some senses than the run with misfit of 34, since that thins thousands of years too early (assuming that the assertion in L 487 that the float misfit score accounts for timing of thinning is true).
We established a cutoff of 50 simply by interpreting many different plots like Fig 8 for a series of different sites, in the manner employed by the reviewer above. So, yes, this is another place where we were forced to establish some ad-hoc cutoff points.

L607: How sensitive are the conclusions to this spatial weighting scheme? It seems like you could defensibly define this in multiple ways — for instance by using ice-flow catchment boundaries (either time-dependent or modern) — and those could possibly give different answers without a clear way to choose between them.
We did not test different spatial weighting schemes; we simply follow the guidance of Briggs & Tarasov (2013) and Lecavalier & Tarasov (2025) who address the issue of variable data density by developing an inverse areal density weighting scheme which we adapt here. We plan to use (paleo)ice flow catchment boundaries in future work that will consider both terrestrial and marine geologic datasets; but for this project, we feel that the areal weighting scheme is appropriate.

Eq 5: I'm still a bit lost at this point. Reading from the top of section 3 from top to here, I don't understand how all the different metrics come together. You first define a float-scoring (i.e.,

thickness change) metric and a best-time-offset metric in S3.2.3. Then in 3.3.1 you define this sample misfit metric that appears to be more or less independent of either of those. That metric is used to calculate site misfit, which is then used to calculate model misfit. So where do the float-scoring and best-time-offset metrics come in? As I've mentioned above, this section would greatly benefit from a very clear roadmap at the beginning that gives a very clear summary of the procedure before you dive into the details. The text at the beginning of Section 3 seems to attempt this, but doesn't really help clarify the approach for me.

From this and subsequent comments, we add a Table 1 to the end of Sect 3, to accompany our introduction of our three metrics (float scoring metric, time offset metric, and exceedance metric).
This will clarify that the sample misfits are used to construct a site misfit. The site misfit statistic is what is *minimized* by either moving the model thinning profile up and down (floating in elevation) or side-to-side (allowed to shift in time, to identify the best-fit time offset).

L 628–630: need example references *

L705: Technically, saturated C-14 concentrations provide robust evidence for *lack* of LGM ice cover.
Corrected.
The scaling factor of 10 seems arbitrary as well. How sensitive are results to this assumption?
Yes, this quantification of our confidence in this C-14 constraint, relative to our other methods of inferring LGM thicknesses, is indeed arbitrary. We don't have any non-arbitrary way of quantifying confidence between two different ways of indirectly inferring LGM elevations; we simply said, 'We are an order of magnitude more confident that C14 elevations constraints are robust" and designed our weight accordingly. However, we don't think this is impacting results significantly since we only have three C14-based constraints currently; but with this scoring 'weight' we wanted to convey the increased utility of this type of measurement.

L731: Here the float misfit score has come back. Is this in fact the same as the misfit defined in Eq 5? Why is the time-misfit score not mentioned here?
Yes; the float misfit *model* score is simply the sum of the site scores that have been calculated by allowing the model curve at each site to 'float' in elevation; this is the 'float misfit score' approach. This should be clarified by our new Table 1 at the end of Sect 3.

L788: "wrong more consistently" is rather ambiguous phrasing. Reword for clarity. Something like "exhibit a more consistent time-offset bias". Refer to Fig 13 here to help illustrate. *

Fig 13: Median and interquartile range might be more meaningful than standard deviation for distributions like these.
Prior to computing the mean and standard deviation, we remove the 'pathologic' outliers (e.g., sites where models do such a poor job of fitting to the dataset that an extreme best-time-offset value of +/- 15kyr indicates that model/data alignment is meaningless at these sites). So, either mean/s.d. or median/IQR would, in our opinion, adequately represent the degree of clustering of our time offset values.

L835: Are the 20- and 10-km results shown anywhere? If not, a figure should be added.
Not at present; in a revised manuscript, we will weigh either adding a supplement with these figures (we are already at 16 figures in the main text), or simply removing this sentence.

L898: delayed and more rapid compared with data, or with other model runs?
Compared to other model simulations; this will be clarified.

L896–904: My guess is that this is due to the power-law (Weertman-type) basal friction parameterization used in the PSU model, which uses a relatively large exponent and thus assumes a relatively hard ice-sheet bed. A more-plastic bed rheology would likely be more appropriate for much of the Antarctic Ice Sheet, though the appropriate rheology and thus the appropriate friction parameterization will vary widely in space. The power-law exponent has not been varied as a parameter for any study with this model that I am aware of, but it has a large impact on time-evolving behavior in other models (Parizek et al., 2013; Gillet-Chaulet et al., 2016; Nias et al., 2018; Joughin et al., 2019; Brondex et al., 2019; Hillebrand et al., 2022; Schwans et al, 2023). It cannot be determine with a snap-shot inversion approach, but instead requires calibration in time-evolving simulations, or a transient inversion. I think using a more-plastic bed rheology might tend to lower the maximum thickness (leading to better exceedance scores) and also lead to steadier, less abrupt thinning (potentially improving float misfit scores if I understood the previous text correctly). Obviously you don't need to attempt this here, but it's worth mentioning that this key assumption has never been tested with this model.
Thank you for this context; we have not previously explored the idea of a plastic rheology vs power-law basal friction impact on our results. Our varied parameter 'CSHELF' represents that power law exponent, though only for modern ungrounded continental shelf (in other words, we are only changing the basal friction treatment for the outer shelf). It seems that using a plastic bed rheology would have a similar impact (lower LGM ice thickness and less abrupt thinning rates) as using a smaller power law exponent

(CSHELF=5).

L980: One explanation that occurs to me (although there are many many possible and complementary explanations) is that deposits in currently ice-free valleys alongside TAM outlet glaciers can be hundreds of meters lower in elevation than deposits of the same age that are on the glacier valley walls because the glacier margins extended several km into these valleys at the LGM (see, eg., the wide range of elevations of the mapped Britannia I limit at Lake Wellman in King et al., 2020). Those two populations would both be compared against very similar (or identical) modern-day surface elevations, so they record very different ice thickness changes despite recording the same event at almost the same location. Without some calculation that accounts for the very different elevation of these valley-floor deposits (e.g., using some estimated surface slope to project them back to the glacier centerline or nearest model grid-cell), it's unlikely that they will be used accurately in this analysis. Recorded rates of thinning also vary widely (up to maybe a factor of 2) between valley-floor and nearby valley-wall samples. An automated approach will likely always miss the difference between these types of samples. To be fair, however, most expert curation would has also missed that distinction.
This is a great example of why we developed our float scoring approach – allowing the model curve to 'float' vertically in elevation to best match the exposure age thinning profile – because the modern 'reference' ice elevation is difficult to constrain/align between models and data, as described. Using this float scoring approach, both valley-wall and glacier-margin thinning curves would be compared to their respective model thinning curves *independently of the modern ice surface elevation.*
The key part that we still need to get right, though, is making sure the model can resolve between a valley wall and a glacier margin – and we think this is why our high-resolution model domains do a better job of matching measured amplitudes of ice thickness change (for example Fig 12), although 2km is still coarse relative to the complex topography of the TAM. If the model is high enough resolution, in principle, it would capture both of these environments correctly.

There's also the issue that the small-scale meteorology of the glacier margins is a strong control on location of deposits, and is not going to be represented at all in the model. For instance, the presence of algae-hosting melt ponds at the LGM and even the deposition of erratics in valleys requires marginal ablation areas.
Yes, this small-scale but impactful meteorology (especially in the TAM) is still very poorly represented in models, even with our 2km domains.

L991: Resolution is just one of many important model choices that have a bearing on this, so it's not necessarily resolution that is limiting accuracy at this point. The remaining discrepancies are more likely due to all the other sources of uncertainty (model structure, unrepresented physics, parameters that likely need to vary spatially, bed topography, forcing, etc). I would guess that you're not going to see much further improvement at higher resolution at this point, although it

would be interesting to test that for a few of these ensemble members. The nested domains are also still driven by the low-resolution simulations at the boundaries, so they cannot completely decouple from the inaccurate low-resolution ice dynamics occurring outside the high-resolution region. This might not be a big issue at the interior sites, but in the TAM it is probably important, since the grounding-line position in the Ross Embayment is likely to be resolution dependent. **These important considerations are now added to the main text.**

The sign convention on time in the figures in the appendix is reversed relative to the main text. Please point this out in captions. *

References:

Balco, Greg. "A prototype transparent-middle-layer data management and analysis infrastructure for cosmogenic-nuclide exposure dating." *Geochronology* 2.2 (2020): 169-175.

Bintanja, Richard. "On the glaciological, meteorological, and climatological significance of Antarctic blue ice areas." *Reviews of Geophysics* 37.3 (1999): 337-359.

Brondex, J., Gillet-Chaulet, F., and Gagliardini, O.: Sensitivity of centennial mass loss projections of the Amundsen basin to the friction law, The Cryosphere, 13, 177–195, https://doi.org/10.5194/tc-13-177-2019, 2019.

DeConto, Robert M., et al. "The Paris Climate Agreement and future sea-level rise from Antarctica."*Nature* 593.7857 (2021): 83-89.

Gillet-Chaulet, F., Durand, G., Gagliardini, O., Mosbeux, C., Mouginot, J., Rémy, F., and Ritz, C.: Assimilation of surface velocities acquired between 1996 and 2010 to constrain the form of the basal friction law under Pine Island Glacier, Geophys. Res. Lett., 43, 10311–10321, https://doi.org/10.1002/2016GL069937, 2016.

Hillebrand, Trevor R., et al. "Holocene thinning of Darwin and Hatherton glaciers, Antarctica, and implications for grounding-line retreat in the Ross Sea." *The Cryosphere* 15.7 (2021): 3329-3354

Hillebrand, Trevor R., et al. "The contribution of Humboldt Glacier, northern Greenland, to sea-level rise through 2100 constrained by recent observations of speedup and retreat." *The Cryosphere* 16.11 (2022): 4679-4700.

Joughin, I., Smith, B. E., and Schoof, C. G.: Regularized Coulomb Friction Laws for Ice Sheet Sliding: Application to Pine Island Glacier, Antarctica, Geophys. Res. Lett., 46, 4764–4771, https://doi.org/10.1029/2019GL082526, 2019.

King, Courtney, et al. "Delayed maximum and recession of an East Antarctic outlet glacier." *Geology* 48.6 (2020): 630-634.

Nias, I. J., Cornford, S. L., and Payne, A. J.: New Mass-Conserving Bedrock Topography for Pine Island Glacier Impacts Simulated Decadal Rates of Mass Loss, Geophys. Res. Lett., 45, 3173–3181, https://doi.org/10.1002/2017GL076493, 2018.

Parizek, B. R., et al. "Dynamic (in) stability of Thwaites Glacier, West Antarctica." *Journal of Geophysical Research: Earth Surface* 118.2 (2013): 638-655.

Schwans, Emily, et al. "Model insights into bed control on retreat of Thwaites Glacier, West Antarctica." *Journal of Glaciology* 69.277 (2023): 1241-1259.