We thank the reviewer for their time and effort in providing feedback on our manuscript. Our responses to their comments are given below, in red. In the case where no direct response is given, we will implement the reviewer's suggestions directly into the revised manuscript without further changes.

General Comments:

In this discussion, the authors introduce a new methodology for comprehensively determining uncertainty of ice sheet model projections. The study focuses on modeling ice sheet change in the Amundsen Sea sector of Antarctica under two different future emission pathways through 2100 and then through 2250, using the ice sheet model Úa. The authors design a method to derive uncertainty in their projections by training a surrogate model and then using Bayesian calibration to down select the most reasonable parameter space for their set of historical simulations. They then sample the calibrated parameter distributions, running an ensemble of future projections, which are in turn used to train a new, time-dependent surrogate model to exhaustively cover the possible parameter space. Results allow the authors to quantify uncertainty for every year of their projection and attribute uncertainty to each of their various parameters. They conclude that uncertainty is dominated by parameters related to initialization of basal sliding and ice rigidity for both scenarios followed by parameters related to ice flow and basal melting. Finally, the authors extend a subset of simulations in time through 2250 and find that mass loss accelerates through the year 2200, returning to a sea-level contribution similar to today's by the end of the simulation, for both scenarios similarly. Results are compared against projections from previous Bayesian-based studies, concluding that the sea-level projections for this study are more conservative than others, especially considering that the simulations presented include a MICI mechanism for ice shelf collapse.

This manuscript is well written and of high quality. The work is novel, and the figures are well designed, easy to read, and support the stated conclusions. The experiments themselves required a significant amount of work and thoughtful effort toward their design. The authors take an organized approach to describing the complex workflow, including a helpful schematic figure and extensive appendices. For these reasons, I support the publication of this manuscript in The Cryosphere.

However, from a scientific point of view, I think the authors could expand upon their discussion of the results more. I realize that the main point of the paper is to introduce the novel workflow to quantify uncertainty in ice sheet model projections, but the authors do make a point to compare their projection results against other published projections for the region. They also conclude that their projections are more conservative than others, showing very little signs of instability (even though MICI processes were invoked during the simulations). Because of this, I believe that the authors should do more than just show the end states of their projection ensembles; it would greatly enhance the manuscript if they also discussed why the simulations are tracking on the low end of sea-level contribution. This includes expanding the discussion to investigate why the simulations do not match present-day sea-level contribution until 2200 and why the simulations for both scenarios increase so distinctly in the extended simulations (at year 2100). I also think it would improve the manuscript if the authors noted some implications for their stated results and conclusions. Finally, I find that in some cases, the method description lacks detail and is vaguely described; it would aid the reader to have them clarified and/or quantified before publication. More specific comments are noted below.

We will expand on the interpretation and discussion of our model results, as requested by all reviewers. We intentionally avoided adding too much information into the methods section since this is already very long, but where the reviewers have requested more details as noted in specific comments below, these will be added.

Specific comments:

Fig. 1: From my understanding, both the historical runs and the forward runs simulate the year 2021. Is that correct? If not, please clarify this throughout the text.

The Ua_{obs} simulations cover the period 1st Jan 1996 to 1st Jan 2021, whereas the Ua_{fwd} simulations begin on the 1st January 2021, so there is no overlap and this will be clarified in the text.

Table 1: I suggest that you also include the ranges for these parameters in the table. That would be very handy information for the reader to have for reference.

We agree that this would be helpful, although using the range only makes sense for parameters with a uniform distribution and is not easy to interpret for parameters with a Gaussian probability distribution, so we will consider how best to incorporate this information without confusing the reader.

Lines 78-79: … along "with" other sliding laws. Additionally, please quantify "some time". It would also be helpful to the reader if other key sliding laws were explicitly listed (i.e. Coulomb and Weertman).

We will add these changes to the revised manuscript.

Line 94: Since you discuss the bias-corrected aspect later in the paper, it would be helpful if you noted that the correction later described (and perhaps reference the section) for the reader.

We will add a reference to the relevant section to better guide the reader to this information.

Line 106: Please include what this threshold is in the text.

As also noted in our reply to reviewer 1, we use the capillary retention model as described in Janssens and Huybrechts (2000) and we will add a brief description of this in the revised manuscript.

Line 148: "which is at least two orders of magnitude larger than basal melting due to geothermal heat flux." Is there a reference that can be added here for a noted estimate of geothermal heat flux magnitude?

We will briefly expand on this comparison and add a reference.

Line 210: Please quantify what qualifies as "large" spatial gradients.

In fact the regularisation penalises any spatial gradient in the inverted field, but larger spatial gradients are penalised more. We will remove the word "large" here to make this less ambiguous.

Line 216: Please specify mass balance, surface+basal mass balance?

We will clarify this means surface and basal mass balance in the revised manuscript.

Line 219: "after which all model forcing terms evolve based on the ERA5 outputs for the period 1996-2021." This statement is somewhat confusing, because if I understand correctly, the ocean forcing does not rely on ERA5. Similarly, my understanding is that the other (atmospheric) historical forcing is not dictated by ERA5 either (but corrected CESM1?). Please rephrase this section for clarity. Also, ERA5 or the historical forcings have not yet been introduced at this point in the paper, so it is unclear what you mean by "ERA5 outputs". Please either introduce (reference) it here or make a reference to Section 2.2 to help guide the reader.

Thanks to the reviewer for spotting this mistake, line 219 should refer to CESM1, not ERA5, and this will be corrected in the revised manuscript.

Line 224: Please give a value for the initial coarse resolution.

The coarser resolution mesh is itself unstructured and so it does not have a specific resolution.

Lines 239-241: This statement is confusing to the reader because RCP4.5 is noted as running through 2080, but the obs simulations only run through 2021, is that correct? Also it not clear what "that most closely matches observed atmospheric conditions in the model domain" means. Please rephrase.

We will clarify both these points in the revised manuscript.

Line 243: Please add a reference for the CESM1 simulations used here, as well as the ensemble simulations. This part of your method discussing the inclusion of climate variability is quite vaguely described.

We will expand on this in the revised manuscript.

Line 252: Is the bias correction done monthly (for temp and precip)? Is the temperature correction spatially varying as well? Please add a more precise description of this method.

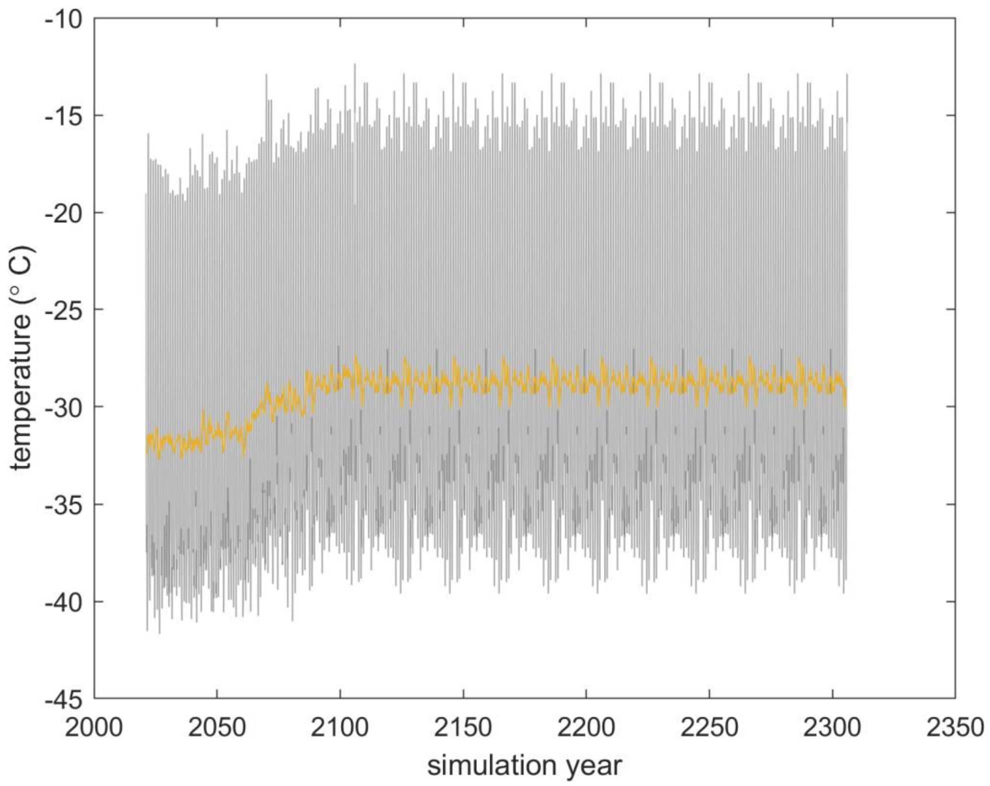
Bias correction in both cases is done with a spatially varying but constant in time term, so we assume that CESM1 has a systematic bias, whose severity depends on location, but which does not change as a function of e.g. present climate. We will add a better description of this in the revised manuscript.

Lines 257-258: Please clarify if the bias correction is done on all precipitation and temperature forcing, even the fwd simulations. In this case, if the historical bias correction for precipitation was used on the future forcing, then the larger precipitation values grow, the larger the correction (because a scaling factor is applied)? In this case, couldn't the surface mass balance future trend be altered by the precipitation scaling? Could you comment on whether or not that is a concern?

The bias correction is indeed applied in the same way for all simulations, however the term 'scaling factor' was poorly chosen, we use the same approach as Naughten et al. (2022) whereby a spatially varying field calculated as the difference between time averaged CESM1 precipitation/temperature and our 'preferred' dataset e.g. AntAWS/ERA5/MAR over the same period. This difference is considered to be the bias and the spatially varying bias field is added to the CESM1 field, so the size of the bias term will not be affected by changes in the temperature/precipitation fields with time.

Line 265: Please specify how this is done. Point by point on the forcing grid? Is it for temperature only, or also for the ocean modeling? I am curious about whether this affects your trend when you restart your simulations, as it is not clear why both emissions begin increasing in sea level contribution starting right at 2100 (Figure 9). The RCP8.5 in particular changes from a downward trend to an upward trend almost instantaneously. Have you diagnosed what causes the sudden retread based on your simulation ensemble? This is an example of a curious model response that could be explained and discussed in more detail in the paper.

For every point on the forcing grid, for both temperature, precipitation, and all MITgcm model forcing, a linear trend is calculated for the period 2080-2100. The forcing for all fields for the years 2100-2250 is then taken as a repeat of this final period with the calculated linear trend removed. An example time series of temperature from a point in the domain is given in the figure below (gray line is monthly temperature, yellow line is moving average temperature with a 12 month window).



Regarding what the reviewer highlights as sudden retreat in Fig. 9, firstly this figure shows sea level contribution rate in mm/yr, so it is not directly related to retreat and the change in the (average) rate that appears to happen at around the year 2100 would therefore be much less apparent if plotted as change in ice volume. That being said, there is a clear change in the trend, which is stronger for RCP8.5 than Paris2C. Attributing this change which is an average across all simulations with grounding lines in different positions is challenging but the most likely cause is simply as a result of changes in surface mass balance, since the trend of increased precipitation and warming is removed after 2100, this also explains why the change is more notable for RCP8.5 where these changes are stronger. As mentioned elsewhere, we are re-running a representative sample of simulations to enable a more thorough analysis of model behaviour and this change in trend will be one focus of that analysis.

Lines 276-277: "Thus, the simulation start dates of 1996 and 2021 were chosen to ensure that the velocity and surface elevation change data were approximately aligned in time, but the precise timing is not well defined." This sentence is awkward, and I am not sure what it means, please rephrase.

What we mean is that these dates were chosen to align as best as possible with all datasets that were used, but the actual timings of these datasets are (a) not precisely defined anyway and (b) do not match one another exactly, and so these chosen dates are a compromise that tries to minimise discrepancy in the acquisition time of each dataset and keep things as consistent as possible. We will rephrase this sentence in the revised manuscript.

Lines 311-312 and Fig. 2/3: My understanding is that you are focused on change of surface elevation and change in surface speed during your obs period. However, I have found that the wording to describe these terms throughout the manuscript is confusing. Sometimes you refer to just surface speed (i.e. Fig 2 is labeled surface elevation and velocity – a side note, in this case, perhaps velocity should be changed to speed?). The Fig 2 caption also specifies that you are plotting variable change: i.e., "surface ice speed (top row) and surface elevation change (bottom row)", but this sounds like you are considering only change with respect to the surface elevation, not surface speed. This is similarly confusing in Fig. 3, since the plot labels include a delta, but the caption is not clear that it is delta speed being shown. In general, speed seems to be referenced in many sections of the paper, but I do not find it explicitly stated in each instance that the benchmark of interest is “change” in speed. This just might be a question of wording and can be fixed easily. Please try to make this clear and consistent all throughout the manuscript.

This is indeed a question of poor wording. In all cases our ‘observations’ that we use to calculate model parameter likelihoods are change in surface elevation and change in surface ice speed. We will carefully go through the manuscript and ensure this is corrected or clarified in all cases.

Line 326: Please note what the resulting k values are in the main text for this analysis and Fig. 3.

We will add this information to the main text (note k=8 for surface elevation change and k=11 for change in surface ice speed).

Fig. 2: In my understanding, your calibration is based on the change of speed and the change of thickness over the observational period of interest. Do you think there are implications to choosing these diagnostics and how do you think it affects your results? For instance, you are testing for a linear change but change in this area is temporally variable and non-linear. Even the clear strong trend captured by the PC1’s (responsible for a significant portion of the changes) are in reality much more complex temporally. What are the repercussions of this assumption? Mentioning these caveats in the manuscript and discussing on how these constraints were decided upon (or computationally forced) would be a very interesting addition to the paper. This is especially the case because difficult decisions and concessions like these are likely why many others have not attempted an assessment of this magnitude.

This is a very valid point and we agree that it would be beneficial to add a discussion of this to the revised manuscript. While the data exists to extend our calibration to include temporal variability in the regional response, and this could help further constrain the model, this approach may currently be too ambitious in terms of how well the current generation of models can represent processes relevant over shorter timescales (e.g. calving, damage, hydrology, ocean processes etc.). While studies do exist that seek to match observed temporal variability over shorter timescales, they are generally focused on one outlet glacier or limited in other ways. Matching the differing response of each portion of the Amundsen Sea Sector over the entire satellite record would be a significant challenge without some modelling compromises such as enforcing the observed calving front positions rather than allowing them to evolve dynamically as we do in this study. We also arguably still lack sufficiently detailed knowledge of bedrock geometry that could greatly impact retreat rates over shorter timescales. That being said, the assumption we are making by only comparing speed and surface elevation at the start and end of each simulation is an important one, and extending this study to include more data for calibration would be an interesting and potentially valuable addition for future work.

Lines 347-348: With respect to capturing the trend and magnitude of these constraints spatially, do you think that PC2+ play an important factor in capturing the variability despite the lower dimensionality? Having done all the work to assess the presented method, how reasonable do you think this method of calibration is, considering mass loss in this area is so non-linear?

As noted in the paper, using just PC1 would only account for 64% of variability in ice speed change and 57% of variability in surface elevation change, so additional terms are certainly needed to help capture the complex response of this region. We refer to our answer above regarding non-linearity.

Lines 359-360: Please note or reference these physically plausible values, or reference where it is discussed (e.g. Appendix C).

We will add a reference to Appendix C here.

Line 413: Quantify “large”.

We will specify that our training set was of size N=2074 simulations

Fig 4: As you mention in Appendix B3, for instance, values for n have been questioned recently. Your calibrated ranges for m and n are interesting to see, and they are a nice result in themselves. It would be great to see some discussion on this result added to the text.

It is indeed interesting that for our calibration we find an optimal value for n centred around 4, rather than 3 that is more commonly used. We will add a brief discussion on this in the revised manuscript.

Line 424: Instead of using "~", please specify the exact number of simulations. Also, please note here that it is 2000 simulations in total (not for each of the scenarios as I understand it).

We will specify that our training set consists of 2074 simulations and that this is both scenarios together.

Fig 5: This is a really nice figure, with a significant amount of information portrayed. While it is clear that the sea-level projections are conservative, and this is amply noted in the manuscript, it would improve the text if you discussed the reasoning for why the trend of sea level contribution is lower than the past observed long-term trend, and much lower than the present-day trend. I would be curious to know what is happening dynamically to slow down the sea-level contribution, and an interpretation of what those results mean (i.e., line 532 suggests that by 2100 the rate of ice loss is ½ of the present-day rate, which is a quite surprising behavior – what is stabilizing the model projections?). For instance, it might be that the majority of your simulations are generally stuck on their current grounding line (but it is surprising that it would happen to be for all of them), or there is something about the initialization constraints that do not allow the transient fwd simulations to contribute as much sea-level as is observed at their start (present-day). Whatever is the cause, an analysis of the simulations could reveal a general conservative trend that many of ensemble members (ice sheet model runs) seem to be following.

We concede that the paper would benefit from more analysis to better understand why the model is behaving as it is, although attributing this behaviour to one thing for a complex model with many interacting components and thousands of simulations is presumably not going to be possible. As mentioned elsewhere, we could not save most model output fields for the thousands of simulations due to storage constraints, making more detailed analysis challenging. We are in the process of re-running a representative sample of simulations with detailed outputs and we will explore in more detail the reasons why the model behaves as it does using this sample for the revised manuscript.

Line 455: It would be interesting for you to comment in the paper if you think this happens because of the type of inversion method being used for this study? That is, there are larger degrees of freedom than if you just inverted for one parameter? Do you think this uncertainty would be as strong if only basal sliding was inverted for, for instance? Could you make a statement in the discussion about the implication for this and choices for model initialization impacting not only future projections but their uncertainty?

This is an interesting question, although we can only speculate on the implications of this choice for uncertainty. The Ua ice flow model inverts for both basal slipperiness and ice rate factor everywhere in the domain because, in the absence of far more detailed observations and a modelling framework to capture complex processes such as damage, hydrology etc. this is the only way to ensure that the ice sheet initial conditions (e.g. surface velocities) match what is observed. If we were to only invert for basal slipperiness, the model would not match observations as well, and in this respect uncertainty would be higher. We will add a remark on this in the revised manuscript.

Line 494: The Sun et al., 2014 is based on Bedmap2 uncertainties, perhaps you could add some additional references for more recent papers that support this hypothesis using BedMachine or similar.

We will add a reference to Wernecke (2022) and Castleman (2022) which are more recent examples that include Bedmachine.

Lines 536-537: As discussed above, please add some assessment of why this is the case and what the implications are for these results, including the sudden increase in sea-level contribution in the extended projections. In addition, do you know what causes what looks like another stabilization (and even possible downtrend in sea-level contribution in Paris 2C) around year 2200? Fig. 6 is very helpful to see that grounding lines do significantly retreat during the extended simulations, but have you seen any reasoning for why when analyzing the results?

As mentioned above, once a sample of simulations have been repeated with additional outputs this will greatly enhance our ability to dig deeper into the behaviour of our simulation ensemble and we will add more analysis related to this and other points raised elsewhere in the revised manuscript.

Line 541: Please make a statement about what it means that these two scenarios are so similar. (In some ways this does make sense with the attribution analysis presented, that model setup and initialization is more important than forcing). Does this mean that ice sheet modelers should take particular caution in initializing their models, especially as computational ability allows for a more complex method and therefore possibly more unknows?

Certainly, initialisation plays a key role in the model response to any forcing, as demonstrated in our analysis. In this region in particular, much of the observed grounding line retreat and acceleration of grounded ice may be a response to past perturbations that the glaciers are still adjusting to. Attempting to replicate these signals without necessarily including the original forcing that triggered them is challenging and although our inversion includes observed thinning rates which helps ensure that the ice sheet is initialised on the rate trajectory, the retreat history cannot be fully accounted for without having detailed observations going back further in the past. We will expand on this along the lines the reviewer suggests in the revised manuscript.

Fig 9: Could you add a line where the current (or historical) sea-level rate of change is for comparison?

We will add this to the revised manuscript.

Line 552: Could you make a statement about the implication of these results?

We will add this to the revised manuscript.

Line 553: Do you think that your historical being linear and calving/climate variability being partly stochastic drives this result in any way?

This is an interesting thought, and maybe partly explains the result. Regarding calving, while this may be stochastic in reality, the calving law that we use (with reasoning behind our choice explained in the paper) is not stochastic and in fact behaves very similarly no matter what else is happening within the model (as can be seen to some extent in Fig. 6). There would be substantial value in exploring uncertainty related to different calving laws, although arguably no calving law currently exists which can be easily implemented and captures the complex and stochastic nature of calving. In terms of internal climate variability, presumably a longer calibration period would increase the importance of this term significantly and 25 years is not really sufficient to fully capture the most interesting decadal variability observed in the region.

Fig E1: This is a nice, and very helpful, plot to include. It suggests that the pdf for your simulations should not be that different from what the surrogate model outputs. Is that correct? For the values of this plot, I think this is accumulated sea-level contribution at the end of each year. Please clarify what the values of the axes are in the caption, to prevent confusion.

The figure aims to demonstrate the fidelity of the LSTM surrogate to the Ua-fwd simulations and so by the same token they should have similar pdfs, yes. The reviewer is correct, the figure shows cumulative sea level contribution and this will be clarified in the revised manuscript.

Technical corrections:

Fig. 1: Please define RNN, GMSL, and LSTM for this figure. Some are defined in other parts of the text, but it would be helpful for the reader to have it spelled out here too.

Line 83: “popular” => I suggest using more formal wording here, like “used more frequently by the community”, or something similar that invokes a scientific backing.

Lines 89-91: “Note that other contributions to changes in precipitation (and hence accumulation) …” This sentence is difficult to understand. Please rephrase.

Line 100: Please make sure symbols are clearly defined (i.e. σ_M , T_{cesm})

Line 154: Please make sure symbols are defined (i.e. r , r_c)

Line 200: Bedmachine => BedMachine

Line 234: “this study” sounds like your study, but I think you mean the Naughten study? Please specify.

Line 330: “changed” => change

Fig. 3: Caption should have elevation as the top row and speed as the bottom row.

Fig 7: On the x axis: h_E => h_e

Appendix A2: Please update the title so it reflects that this is specifically for ocean-induced melt

Line 742: “an Long” => “a Long”

Line 752: This statement does not need to be approximate. Please include the exact number of simulations used.

All technical corrections will be implemented in the revised manuscript.

References:

Castleman, B. A., Schlegel, N.-J., Caron, L., Larour, E., and Khazendar, A.: Derivation of bedrock topography measurement requirements for the reduction of uncertainty in ice-sheet model projections of Thwaites Glacier, The Cryosphere, **16**, 761–778, https://doi.org/10.5194/tc-16-761-2022, 2022.

Janssens I, Huybrechts P. The treatment of meltwater retention in mass-balance parameterizations of the Greenland ice sheet. *Annals of Glaciology* **31**:133-140. doi:10.3189/172756400781819941

Naughten, K. A., Holland, P. R., Dutrieux, P., Kimura, S., Bett, D. T., & Jenkins, A. (2022). Simulated twentieth-century ocean warming in the Amundsen Sea, West Antarctica. *Geophysical Research Letters*, **49**, e2021GL094566. https://doi.org/10.1029/2021GL094566

Wernecke, A., Edwards, T. L., Holden, P. B., Edwards, N. R., & Cornford, S. L. (2022). Quantifying the impact of bedrock topography uncertainty in Pine Island Glacier projections for this century. *Geophysical Research Letters*, **49**, e2021GL096589. https://doi.org/10.1029/2021GL096589