The Cryosphere Discuss.,
https://doi.org/10.5194/tc-2019-240-AC2, 2020

The Cryosphere
Discussions

Interactive comment on “Scoring Antarctic surface mass balance in climate models to refine future projections” *by* Tessa Gorte et al.

Tessa Gorte et al.

tessa.gorte@colorado.edu

Received and published: 25 March 2020

Reviewer #2

We greatly appreciate the reviewer’s suggestions to add to the robustness of our study through the comparison across ensemble members and resolutions within a single model very insightful. These are both very good suggestions that will elevate the scientific quality of this paper. Additionally, we also thank the reviewer for the leave-one-out analysis suggestion. This is not an approach we had considered taking. We also appreciate the references the reviewer listed to help us add context to our study.

Overall, important aspects that are required include (among other things) utilizing the CMIP6 HighResMIP experiments to assess resolution-related aspects, incorporating multiple ensemble members to assess the role of internal variability and a more in-depth explanation, motivation and development (i.e. relative to other literature) of the scoring method. Indeed one possibility would be to re-formulate the manuscript with a focus on comparing scores across different resolutions in the CMIP6 HighResMIP experiments and less of a focus on projections.

These are good points, it is important to make sure we address spread within a model via ensemble members and resolution. To address the former, we will use the Community Earth System Model Large Ensemble (CESM-LENS) and score each of its 35 ensemble members. In doing so, we will be able to say whether a single ensemble member is representative of the whole model. If not, we will redo the analysis using ensemble means for the models. Additionally, we will do a similar analysis using low, middle, and high resolution simulations for the HadGEM-GC31 and ECMWF-IFS models. Here, though, our analysis will lend insight into whether resolution significantly impacts our results.

Comparing GCMs with reconstructions: A major issue with comparing standard resolution GCMs and observations/reconstructions, is that full GCMs are not able to reproduce the detail required in regions of high precipitation. Therefore a standard-resolution GCM that reproduces observed/reconstructed Antarctic-wide time-mean SMB is quite possibly doing so for the wrong reasons. This therefore may not be the most appropriate model for projections. The authors should utilize the HighResMIP dataset to determine the resolution dependence of participating models and the potential implications this might have on model selection. This is relevant to all 5 of the criteria used (mean SMB, SMB variability, SMB trends, modes of variability (EOF analysis) and variance explained by the modes). With regard to the EOF analysis, *from Figure 5 seems to suggest highly regionalized nature of patterns from the reconstructions. Indeed, an assessment of natural variability is again crucial here in identifying uncertainty in comparing observations and models.*

There is a lot of validity in saying that GCMs are, compared to RCMs, incredibly low resolution which, thus, makes it difficult for them to reproduce the detailed SMB response. We also feel that RCMs are much more accurate in capturing AIS SMB. However, due to their high resolution, RCMs are also relatively computationally expensive to run for long periods ($\sim$100s of years). Because one of the goals of this paper is to investigate the future of SMB over Antarctica, we use GCMs for their ability to simulate these long-term climate effects. Additionally, as RCMs are by definition regional, they need boundary forcings adding an additional layer of complexity to running the models for century-length timescales.
An additional reason that we decided to use GCMs is simply to figure out which GCMs perform best at capturing this phenomenon. There has been extensive work investigating SMB in RCMs as the reviewer points out, but relatively little looking at GCMs. To investigate the global coupled response to future SMB changes, one needs GCMs, not RCMs. As such, this work is meant to inform modelers who are concerned with global ramifications of changing AIS SMB.
We do not want to dismiss this point, though, that RCMs will almost always be more accurate at representing SMB and its drivers. We will be sure to add text to the introduction section that underscores our reasoning for investigating GCMs over RCMs. Additionally, we will include 2 models each with 3 resolutions from the HighResMIP experiment to help address the valid concerns about the impact of resolution on model score. We will also include in supplementary materials a scatter plot of model resolution versus overall score to further address this issue.

A lack of mechanistic explanation for why each of the 5 criteria are relevant for improving reliability of projections: Firstly the authors should outline the rationale for

inclusion of each of the criteria and how they may potentially improve reliability of projections. It is important to discuss this in the context of existing literatures. For example, Krinner et al. (2014) found that future change in SMB was more associated with thermodynamic, rather than dynamic, factors. Secondly the authors should consider the possibility of leave-one-out cross validation, whereby the real world is can be replaced by each member of the model ensemble in turn to see whether evaluation against that model can help improve predictions from that model. This can help to identify which criteria are most relevant in terms of future projections.

We understand the reviewers comments here regarding the criteria selection process. This process went through several iterations of internal review and revision. As the reviewer rightly points out, there is preexisting literature investigating underlying thermodynamic processes that drive AIS SMB. However, we feel that our paper is different from these earlier papers in that we are not trying to investigate a models ability to capture these drivers as much as reproduce the actual reconstructed SMB record. We do, with our EOF analysis, check whether models are recreating spatial SMB patterns of variability which, we feel, addresses the point as to whether the models are, on whole, doing a sufficient job of recreating the physical SMB drivers. We do, however, feel that the reviewer also makes a valid point that we could further justify the choices we made in selecting the criteria and cite more preexisting literature. We will also include text that denotes the separation of this work from the other related literature. We also appreciate the suggestion of the leave-one-out cross validation and will perform this analysis with our current criteria in the supplementary material.

The methodological framework for model weighting: In addition to the criteria selected, the rationale for the methodology on model weighting needs to be carefully introduced and motivated. Indeed it is common for a model weighting method to be developed initially in a separate paper and then applied to model output in subsequent papers. Specific suggestions are: Firstly the authors need to bring in more of the *previous substantial literature on model weighting. Agosta et al. (2015) use a Climate Prediction Index approach which, as I understand it, draws from probability theory and the probability that observations and models may agree (this goes back to Murphy et al., 2014). There are also detection and attribution approaches, which use past trends to scale future projections and should be mentioned. What is the advantage of the approach used in this discussion paper? Secondly, the authors should consider the implications of situations where the reconstruction uncertainty is small. In the extreme case where it approaches zero, in general models would be many multiples of this uncertainty range away from the reconstruction. How is/would this be handled in terms of relative weighting across different criteria? Thirdly, the method needs to be described more clearly and is in fact difficult to fully evaluate. The whole section needs to be improved and I have just identified one example starting on line 109. Specifically the text: "if a model time series was fully captured within 2× the reconstruction uncertainty, the model would receive a score of 2". I could not find a clear definition of "reconstruction uncertainty". This exact term is only referred to once in the preceding text on line 69. Is it the same as the "total uncertainty" mentioned on lines 72/73? If so, how does the spatial and temporal information map of total uncertainty map onto the AIS-integrated SMB? In the same paragraph it is not clear what is meant by "model time series fully captured"? Does this mean that even extreme years in the model time series are considered? My recommendation is to write out these score criteria as equations to make it easier for the reader to understand and assess them.*

Addressing the point that justification for the criteria needs to be introduced and/or expanded upon: we will make sure to further rationalize the inclusion of each criteria and add context from the literature.

In regards to the point wherein the reconstruction uncertainty approaches zero, if this is the case, then all the models would score highly on this criterion, that is correct. However, after doing the initial scoring, each criterion score spread is normalized to a scale ranging from 1 to 10. As such, all scoring criteria are weighted equally. We

realize that this needs to be expanded upon in the text, and so we shall be sure to include a more detailed description of this process to alleviate further confusion.

We agree with the reviewers recommendation for added clarity in the text regarding the scoring process. We will add Fig. **??** as well as several equations for relevant criteria to help illustrate the process which, ideally, will offer much more insight into the process.

The role of internal climate variability in trend and spatial EOF analysis: The potential role of internal climate variability in evaluating trends is not mentioned, but could be very important. This could be very important for 50-year trends and the spatial EOF patterns. The authors should test the possible role of internal variability by assessing climate models with multiple ensemble members of their historical runs.

To address a very valid comment by the reviewer earlier, we will be including the CESM-LENS experiment to take into account changes due to internal variability. In our analysis of this sub-experiment, we will make sure to address in greater detail the potential role of internal variability both with regard to shorter trends and EOF analysis.

Final model selection: The final selection of 4 CMIP5 models for projections should be compared and contrasted with related studies, Agosta et al., (2015) and Barthel et al. (2020). The reasons for, and implications of, differences should be discussed. What is the significance of the smaller spread across these four models. They come from only two model centers (GISS and MPI). Such close links calls into question the statistical significance of spread across models from just two groups. This could be small or large by chance.

We will add in discussion of the results found in both papers, stressing the differences between the methodology and and the influence of these differences on the final results. Both papers reached different final conclusions regarding which models they concluded captured AIS SMB best due, in large part, to disparities in the methodology. Agosta et al. (2015) focused their investigation into more atmospheric and oceanic dynamics (sea ice extent, sea surface temperature, sea surface pressure, precipitable water, 850 hPa temperature) while Barthel et al. (2020) ruled out both the GISS and MPI modeling groups due to their initial selection criteria. However, we will still explore how the common models compare overall from our analysis compared to their papers. We will also add in further analysis looking at the spreads for each modeling group and note whether the reduction in spread is more a result of modeling group (and, thus, model physics) or a more reflective spread due to uncertainty reduction with regard to AIS SMB.

Impacts of wider factors on projections (e.g. conditions over the Southern Ocean): Another major caveat with the SMB-focused model evaluation is that wider model biases that are known to be important for projections, such sea-surface conditions surrounding Antarctica and hemispheric-scale atmospheric circulate biases, could have an effect on projections (e.g. Krinner et al., 2014; Kittel et al., 2018). The authors do acknowledge this, but don't make implications of differences clear. Could it be that the results of this study should be interpreted alongside other studies?

We appreciate the reviewer's comment here that we should include more of this literature to provide context to our experiment. Acknowledging these biases is important for providing a complete picture to the reader as well as putting into context the limitations of this work. We will include in our text words to the effect of: One of the major caveats with this work is the relatively narrow scope of just looking at AIS SMB. Because we refined our criteria at the outset of our experiment to solely reflect model performance with regard to capturing SMB and didn't include outside factors like synoptic weather patterns, sea ice or sea surface conditions (Krinner et al. (201); Kittel et al. (2018)), there are potentially some wider model biases that we are missing that could affect SMB projections. In our analysis, we make the significant assumption

that the past ability to capture SMB correlates to higher skill in projecting AIS SMB into the future. However, model biases in some of the larger physical drivers – and how those biases change into the future – will significantly impact future AIS SMB trajectory.

Inter-annual variability in GCMs and regional models: On line 79 it is stated that "Global climate models tend to show higher skill at representing interannual variability compared to regional climate models (Medley and Thomas, 2019)". It is not clear to me why this should be since regional models derive their variability from global models. It is also then notable that all CMIP5/6 models over-estimate SMB variability by so much (line 197). An explanation needs to be provided for this, or at least a discussion of the point.

This is a misinterpretation on our part here. Medley & Thomas point out that global atmospheric reanalyses show higher skill than regional models in capturing interannual variability. We will remove this sentence and adjust the following sentence to stress our other main reasons for using global climate models for comparison. Additionally, we have done further analysis on the reconstruction interannual variability and found that the reconstruction process dampens the actual SMB interannual variability signal by a factor of about 1.7 compared to the original reanalysis P-E data. Our analysis shows that this is predominantly owing to under-sampling of highly variable ice cores in selection process. As a result, for this criterion, we will use the original reanalysis data and make a strong note about the interannual variability issues in the reconstruction. We will still, however, include the reconstruction temporal variability analysis in the supplementary material and note if and how changing the "observational" record for this single criterion impacts the final scoring result. With regards to global reanalysis models, though, Medley et al. (2013) did find that global reanalyses exhibited higher skill at reproducing interannual variability than the Regional Climate Model RACMO2. Further analysis revealed that the lack of upper atmosphere constraint allowed the weather to deviate too far from the driving reanalysis. Van de Berg & Medley (2016) determined that applying upper air relaxation within the RCM provided the necessary constraint and significantly improved the relationship between RCM and observations. Thus, RCMs that use upper air relaxation typically exhibit higher skill in reproducing the interannual variability, so often there is a range of skill depending on how much freedom the RCM is given to deviate from reanalysis forcing.

Interactive comment on The Cryosphere Discuss., https://doi.org/10.5194/tc-2019-240, 2019.

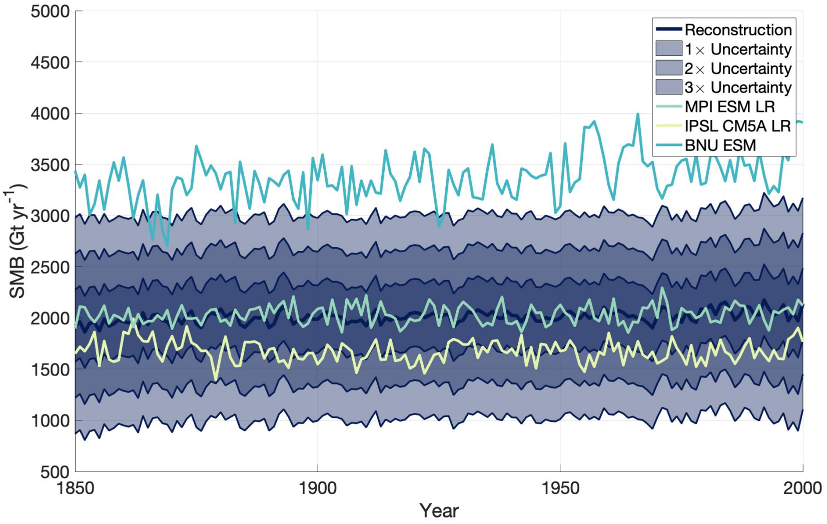

Fig. 1.