Review of Mottram et al., What is the Surface Mass Balance of Antarctica? An Inter-comparison of Regional Climate Model Estimates

Summary

The authors present an intercomparison exercise of five different regional climate model surface mass balance estimates, as well as the near surface climate, over Antarctica. The authors find a large spread in total SMB (1961 to 2519 Gt year-1), which largely stems from differences in West Antarctica and the Antarctic Peninsula.Variability is quite consistent between models, which is unsurprisingly since they are all forced by ERA-Interim, but the trends differ in sign and magnitude and are quite sensitive to the time period selected.

Also, not surprisingly, the nudged models simulate the near-surface climate better as they are not allowed to deviate as substantially fromERA-Interim as the un-nudged models. Finally, the authors discover that the biases are typically consistent between models. The paper presents a significant amount of work but still requires improvements.

We thank the reviewer for their careful reading and thoughtful comments, which have led to significant improvements in the manuscript.. We address the specific points in the text below.

First, the manuscript has numerous mistakes throughout and needs refinement of the language in several places (see Minor Comments).

This point has also been made by the other reviewers. The whole manuscript has been thoroughly proof-read and made simpler to read and more consistent in language and structure.

More importantly, there are several major issues with the analysis that need to be addressed to improve the scientific rigor of the paper.

Major Comments

1. Throughout the manuscript, it is not clear what time periods are being used. There Is the common model interval, climatological mean, etc. The authors need to be very clear throughout the paper because it often seems that different intervals are being confused in nomenclature. It's not clear to me why the common reference intervals are not the common period between all models: 1987 – 2015. Throughout the paper sometimes its 1980-2010, 1987-2015, and 1987-2018. I recommend using the same interval through to avoid confusion. If the authors have a reason to use different intervals, then please make it clear what interval is being used. It is additionally unclear why 1980-2010 is representative of the climatological period, please explain.

We agree it is a bit confusing that the models were run for slightly different periods and this also makes comparisons between them more complicated. Unfortunately we were constrained to use simulations that were already available for this analysis. We have

added a paragraph explaining the different periods used to the Methods section and explicitly mentioned all the way through the paper which periods are being used during the results and discussion when relevant. We have also done additional analysis that shows the results of the observational comparison to the different models does not substantially change when using the shorter period. This is added to the supplementary material in order to avoid further lengthening the paper.

2. Similarly, there is no discussion of significance for the statistics presented. There are claims within the text that certain models perform better than, but without significance levels, these claims lack strength and are more speculative.
Trends are discussed at both long (1987-2015) and short (decadal) time intervals, but the significance is never discussed. I would caution the authors' descriptions of trends, especially at short time scales, since it is very hard to observe a significant trend in SMB since its highly variable year to year.

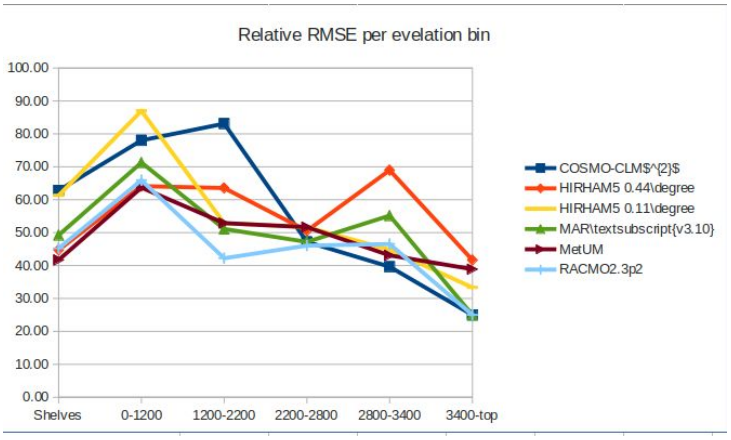
We absolutely agree that detecting a significant trend in SMB is almost impossible and in fact that is partly why we include Figures 7-9. However, given all reviewers comments we have clearly not described this well enough. We have added a paragraph making this point explicitly and setting the SMB in context.

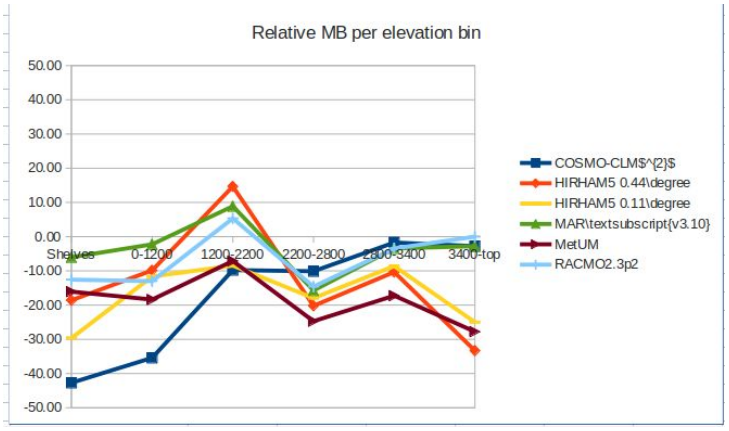
“Unlike previous studies, we detect no obvious strong trend in the modelled SMB in any of the models or in the driving ERA-Interim model. Shorter periods within the time series appear on first look to have quite strong trends, for example a steady declining trend is apparent through the 1990s and 2000s but appears to have reversed since 2014. Our results suggest that strong interannual and decadal variability makes the identification of meaningful trends over short periods very difficult. Distinguishing noise from signal will be challenging in coming decades and this also emphasises the importance of long time series of observations.”

Furthermore, because this is an intercomparison paper, it’s important for the authors to be very clear concerning the metrics of how they conclude one model outperforms the other. Is it RMSE? R2? Bias? And what is the threshold? Is an RMSE of 93 better than 97? What if one model performs differently at different elevation bands?
I did not find the argument compelling that the models tuned to specific Antarctic conditions outperformed the others because there was not a clear frame-work for comparison. The authors need to make clear the evaluation metrics and how they evaluate model performance, which will require more detailed statistical analysis throughout.
Model means are compared, but its not clear if the paper considers even a simple statistic of the standard error of the mean. The Student’s t-test can be used to evaluate whether the means are different. Please be transparent with the limitations of the analysis and provide meaningful significance tests on all of the comparisons,otherwise the conclusions are speculative rather than significant.

This is a very important point and in part one of the drivers for this paper. We do not attempt to rank the models because it is clear from our results that on different measures, (bias, RMSE etc) the different models perform quite differently for different variables (both meteorological) and there is also a spatial component as the reviewer points out with different biases apparent at different elevation bands and in different locations. This means that most likely different models should be used for different purposes. It is also an aim of this paper however, to give clarity on exactly how the models compare, for which reason we give extensive statistics in figure 1 and table 2, which we have also expanded to include the mean value of SMB observations for the elevation bands as requested by reviewer 1.

As the paper is already very long, we propose to add 2 new figures and associated table in the supplementary section showing how the relative RMSE and mean bias compares between models for the different elevation bands. These for instance show that at high elevation COSMO_CLM, MAR and RACMO better represent SMB but HIRHAM, MetUM and COSMO-CLM have a lower mean bias in the middle elevations and MAR, HIRHAM and RACMO have a lower mean bias in the lowest elevations. In addition we have added some extra discussion comparing the different statistical methods and their use in evaluating the models in the discussion of table 2.





3. All of the RCMs presented are forced by the ERA-Interim reanalysis product. I find it concerning that there is no discussion of the role of using a single reanalysis to force all of the RCMs. Thus, this is not a definitive evaluation of the full range of possibilities in SMB, but rather a range due to RCM differences alone. There should be more discussion about how there would be additional spread due to varying choice of forcing; specifically, what is the impact of comparing models that are all forced by the same reanalysis. I think the paper needs to tone done the claims about the work providing the "likely range of SMB" in the first sentence of section 4.1, as it is more the likely range of RCMs forced by ERA-Interim. Basically, this explores the range in RCM space, but not reanalysis forcing space.

The point of this study is to determine the RCM uncertainty space rather than different boundary conditions. We specifically excluded models that ran different reanalyses as we would like to determine how different models compare with each other. However, having said that, analysis by Agosta et al., 2019 used different reanalyses to force the same model and found that the results were quite similar. We have added two extra sentences and this reference stating this in the methods section.

"All models were forced on the lateral boundaries with the ERA-Interim climate reanalysis (Dee et al., 2011) but downscaling used different grids, over slightly different domains and at different resolutions with slightly different ice masks used in the different model versions (see A1 in the appendix). Simulations with MAR forced by different reanalyses (e.g. Agosta et al., 2019) found that results were rather similar to ERA-Interim. However, in order to exclude additional variability potentially introduced by using different boundary forcings, we chose to use a single common reanalysis only"

4. It's obviously quite a challenge to compare these models, which have differing levels of complexity. But it seems that the comparison would be better suited by comparing all the variables consistent between models (Precip-Evap-Subl). Otherwise, an inter-comparison doesn't shed much light on direct model to model differences. In fact, it appears that the authors could investigate whether these Antarctic specific physics actually provide improvement, which would be of great interest to the community. Therefore, the paper should do an ideal comparison of all 5 models with common variables (P – E – S) and evaluate performance. Then evaluate the models with extra physics (RACMO/MAR) to see if and how much model performance improves. Otherwise, it is difficult to untangle whether those additional processes provide any more information.

We have also addressed this point in our response to the second reviewer. Melt is likely to become more important in the future, but at the present day melt and runoff are only observed at a few very specific locations. However, while all of the models simulate melt, they have varying degrees of complexity to calculate refreezing so purely to simplify the comparison here we focus only on the precipitation,

evaporation and sublimation terms. Then, RACMO and MAR include sublimation from blowing snow subroutines, while it would be ideal to separate these out, the physical parameterisations have been developed and tuned with these processes in the model, so it is difficult to remove them without negatively and unfairly affecting the results- we have therefore also used the sublimation from snow schemes in calculating SMB. We note that both RACMO and MAR groups have published articles demonstrating the improvement from the enhanced snow schemes (Van Wessem et al and Agosta et al., 2019). We have added extra text to make this point in the description of the SMB in the methods section.
"As the RACMO and MAR models have been developed to include the wind blown snow sublimation terms, they cannot easily be removed without retuning the models, and for this reason we have opted to include these within the SMB calculation for these two models."

5. The manuscript needs to justify the use of SMB observations starting in 1950.There are regions of strong trends in snow accumulation that might end up biasing the comparison. If the issue relates mainly to reducing the number, the authors could present a comparison of only coincident SMB observations with the data, but then also provide the more liberal comparison as it currently exists in the text.

Unfortunately there are relatively few observations in Antarctica and including only those that were taken during the period of the simulations would make the model - observation comparison less robust. Including observations taken from 1950 increases the number of observations available to 923 comparisons from 469. More importantly, while the total number still sounds substantial, the main benefit is in fact in spatial representativeness. The 1987-2015 observations cover only a very small part of Antarctica. We discuss this problem in section 2.3.3 but we have expanded the discussion and we propose to include a new figure showing the spread of observations by date in the supplementary material. As we point out in the discussion and conclusions, the difficulty is also that the places where the models disagree most are also the areas with the sparsest observations of SMB.
Furthermore, observations starting before 1987 are often ice cores and are therefore the average of this long period (there is also a selection criterion which requires that they must last more than 5 years to be considered in the comparison. This obviously smoothes out the results of a strange year (for example due to sastrugi or scour) in addition to reducing measurement error. At the same time, we compare the average smb of the models over our entire period, so we also smooth out variability between simulated years. This means that using the observations before 1987, we likely have higher quality obs (lower measurement error and better spatial representativeness) and also slightly smoothed models. We have mitigated the problem of unrepresentativeness as much as possible by for example excluding observations before 1987 that cover too short a period (less than 5 years) in order to keep only observations representative of a mean climate. However, we are not immune to introducing biases because these

observations include biases arising from regional circulation trends that the models cannot represent, nonetheless the comparison is more robust if it represents a larger area. We have added these points in the expanded methods section describing the comparison with observations.

6. Finally, the paper needs to discuss the impacts of its findings. With the present day mass loss from Antarctica on the order of 100 Gt per year, this is quite concerning finding the differences in SMB from RCM choice alone are several hundred Gt per year.Please contextualize the findings in regard to how we can measure the mass balance of the ice sheets.

We have added a paragraph in response to reviewer 1's comments along these lines, where we relate the modelled SMB to the latest analysis of Antarctic mass budget from altimetry and GRACE observations:

"It is interesting to compare our results with those used in the IMBIE study of Antarctic mass budget (Shepherd et al., 2018). When taking into account the published uncertainties on the observational mass budget estimates from the input-output method, only the COSMO-CLM and MetUM estimates are outside the range defined by the IMBIE study based only on altimetry and GRACE data. However, as these two models, particularly MetUM, perform well in comparison to meteorological observations, the source of the mismatch is unclear and an area that requires significant future work. It may also indicate either that some of the components of SMB are poorly captured by the models or that there are compensating errors in the modelled SMB components and/or their spatial variability. Nevertheless it is therefore also important to consider the wide uncertainties in both observations and the likely biases in models discussed in this paper, in assessing the contribution to sea level rise from Antarctica"

Minor Comments
Several model names and versions are discussed before they are described, which makes it quite hard to follow. Please reorder the sections to ease.
For instance, section2.1 and the end of Section 1 mention several models and different version, but there is no description, so it's hard for the reader to follow. It would also be appropriate to cite the papers that refer to these model versions.

We have reordered and expanded this whole section to make it easier to read and to follow which models are under discussion and how they relate to each other and to give further details on the different schemes and parameterisations.

P1, Line 7: Is this for grounded ice only? Does it include islands and ice shelves?
This was for the whole ice sheet including ice shelves. We have added a paragraph in the introduction discussing how SMB is calculated and the differences between grounded ice sheet and ice shelves to clarify our ice mask definitions. The abstract has been completely rewritten to summarise the conclusions of the paper more effectively.

P1, Line 7-8: What do the values after the ± represent? The standard deviation of all the models?
The values represent the standard deviation of the annual ensemble mean including all models, but see comment above also.

P1, Line 10-11: Why is 1980-2010 chosen as the climatological period? Later in section 2.3.3, it appears that 1987-2015 is the common modeling period that is used to"compute a climatological mean" (P10, Line 12). Please rectify.
We realise that it is confusing that the models were run for slightly different periods and this also makes comparisons between them more complicated. We have added a paragraph explaining the different periods to the Methods section and clarified all the way through the paper which periods are being used during the results and discussion. The abstract has been rewritten to reflect this also.

P2, Line 1: change "compar" to "compare"
Fixed

P2, Line 11: remove either "potentially" or "potential" since its repetitive
Fixed

P2, Line 13: add "and" after "2002,"
Fixed

P3, Line 16: remove the comma after "published"
Fixed

P3, Line 23: remove "drive"
Fixed

P4, Line 1: please describe what a "reinitialized hindcast" is
A reinitialised hindcast is a model run in weather forecast mode that is reinitialised by observations every 48 hours. We have added a line to explain this.

P4, Lines 5-7: While this is true, it might have a limit. See Lenaerts et al., 2018, which shows that often the snow is not dumped in the proper place when moving from 27 km to 5.5 km. Please add a sentence clarifying this.

This is a good point and in fact one of the reasons we have undertaken this comparison. We have added a line mentioning this
"Lucas-Picher et al. (2012); Lenaerts et al. (2012b); Franco et al. (2012); van Wessem et al. (2018) among others have found that a higher spatial model resolution gives more physically plausible results, especially with respect to precipitation processes

in areas with steep terrain. However, there is also evidence that moving to high resolution (~5.5km) can lead to precipitation falling in the wrong place due to e.g. upslope effects (e.g. Lenaerts et al., 2018; Schmidt et al., 2017).”

P4, Line 8: add "is used" after "ensemble mean"
Fixed

P5, Line 6: change "developed in" to "developed for"
Fixed

P5, Line 18: the end of this sentence needs to be reworded
Edited to:
“Although the mesoscale version includes a multi-layer snowscheme (Walters et al., 2019), in these simulations we used a simplified single-layer scheme with for example, no refreezing (Cox et al., 1999). SMB was calculated based on output precipitation and sublimation and evaporation.”

P6, Line 4: Do you mean "processes" not "process"?
Fixed

P6, Line 22: change "includes no" to "does not include"
Fixed

Table 1: What does SMB scheme mean?
SMB scheme refers to whether or not the regional climate model has been modified to take into account atmosphere - ice sheet interactions, or if it is run in a standard mode without explicitly calculating SMB.

P7, Line 10: change "schemes" to "scheme"
Fixed

P10, Line 10: add "that are" after "2015 and 2018"
Fixed

P10, Lines 14-20 need clarification
We have added more explanation of the technique used to make the comparison between dataset and observation and added more detail on the dataset used.

P12, Lines 4-7: This sentence is very long and needs to be split in two.
Fixed

P12, Line13-14: Please reword the sentence as its confusing.
Fixed

P13, Line 6: Remove "In"
Fixed

Figure 3. Please add the statistics to these plots (RMSE, etc.). Also, its very difficult to distinguish the colors here. Maybe large dots would help.
The statistics in these plots are given in Table 2 for clarity, we also present the models separately in the supplement and (in response to a request by reviewer 1) we have expanded this analysis to include a relative RMSE plot that we intend to include in the supplementary material (see above)

Figure 4. This figure should be much bigger. It's very hard to see the colors. Also, in the caption there are "a", "b", etc., but they do not exist on the plots.
Figure 4 has been revised and enlarged to make it easier to read, we have removed the superfluous letters from the caption.

P18, Lines 9-11: are these values consistent with what is listed in the abstract?
We have revised the text of the manuscript to be more clear about which periods the SMB figures refer to. The abstract has been completely rewritten to simplify it and summarise the conclusions further

Figure 5. This needs to be in landscape orientation. The numbers are much too small to read.
We have edited this figure to make it larger and the labels clearer, however we prefer to keep this orientation as it makes it easier to read and interpret the figure when printed.

P20, Lines 5-6: What does "much clearer mean," please clarify
The topography in the regions noted in the text have a substantial influence on the modelled SMB, this allows physical features such as the Transantarctic mountains to be picked out in the SMB maps. We have updated the sentence to reflect this.

Figure 6. Again, these plots are too small, and the numbers are nearly impossible to read.
Figure 6 has been made bigger and restructured for ease of reading with larger font on the labels.

P22, Line 2: remove "below" There should be significance values associated with the trends. It looks like none would be statistically significant and thus are effectively no different than zero.
We fully agree about the lack of significance of the trends and have added a paragraph discussing this point as discussed above.

P24, Line 6: remove "very"
Fixed
P24, Line 9: change "bring" to "brings"
Fixed
P25, Line 7: Should the interval be 1987-2015?

Yes, here we discuss the overlapping period.

P27, Line 29-30: Do your results actually support "Models that have not undergone specific adjustments for Antarctica clearly represent the SMB in Antarctica more poorly".Look at the RMSE value in Table 2, it looks like sometimes they perform better. Please Clarify.

As discussed above, assessing how well the models perform is complex. The new figures discussed above help to clarify this somewhat but we have modified the text here to take into account the spatial and process variability.
“Models that have not undergone specific adjustments for Antarctica clearly represent the SMB in Antarctica more poorly than those that have been adjusted in some regions . However table 2 shows this is not unambiguous as in some elevation bands the unmodified models have lower bias and RMSE (see section 3.3).

P 28, Line 11-12: please give values in Gt of these processes to show that they are effectively negligible
Fixed

P 28, Line 20: add "fig." before "7"
Fixed

P28, Lines 20-21: the sentence needs to be improved.
Modified to: “The added value from a higher resolution model is that it better captures local topography and associated weather phenomena that in turn leads to more representative outputs.”

P29, Line 2: replace "mod-latitudes" with "mid-latitudes"
Fixed