The Cryosphere Discuss.,
https://doi.org/10.5194/tc-2019-240-AC1, 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 #1

We thank the reviewer for their insightful and thorough feedback. We found these comments incredibly thoughtful and helpful to ensuring that this paper is of the quality expected for the Cryosphere and the field at large. To address some of the reviewers most major comments, we are reforming our EOF introduction, discussion, and analysis, adding in Fig. 1 and eq. (1), and changing the way we are assessing the temporal variability criterion by switching to the original reanalysis data set here.

Major

"To score the time series magnitude, we assigned a score, x, for how many x-times the reconstruction uncertainty was required for the entire time series to be within the reconstruction uncertainty."
** I think you should reformulate this sentence in a more mathematical framework. What did you code? What is the minimum value of your score, 0 or 1?*
if I understand well, you did max(abs(Model - obs))/(reconstruction uncertainty)? So you scaled the maximum difference of model to obs with the reconstruction uncertainty? Why not using the RMSE scaled by the reconstruction uncertainty?

Addressing the first point, we agree that the wording here is tricky and we find it difficult to express this code in words alone. As such, we will add Fig. 1 (see below) to the supplementary material (with additional text) to help illustrate our point. We will add that "the minimum possible score, then, is one, for a model who represents SMB that fits entirely within $1\times$ the reconstruction uncertainty."

"involves finding what spatial SMB patterns explain the highest variance in the AIS integrated SMB time."
** are you sure this what EOF do? Is is not the variance of space-time SMB variability? time series (typo?)*

To the first point, EOFs map the spatial pattern of a variable associated with the highest temporal variance of another variable; in this case, we map the spatial pattern of sea level pressure associated to the highest variability in SMB integrated over the AIS. To the second point, the typo will be corrected.

"To avoid manually sorting the top three modes of variability for all 53 models, we generated difference maps between each of the top three reconstructed modes and each of the top three modes for each model:"
** why do you do this only for the top 3 modes of each model and not e.g. the top 10?*

We will add text to the effect of "the top three modes explain roughly 76% of the total variance explained. The fourth mode explains only about 6% of the total variance and all other modes explain <5% of the total variance. As such, we only include the top three modes in our analysis."

"We then sorted the top modes of variability for each model based on smallest difference"
** what do you call "the smallest difference"? Do you average absolute differences over the map? Do you compute a RMSE?*

We will include the following text to clarify this point: "For each grid point, we took the absolute value of the difference between the model and the reconstruction. We then summed those differences to generate a single number ("difference number") that represented the difference between the model and the reconstruction in terms of spatial variability. Mathematically, this looks like:

$$\text{difference number} = \sum_{lat} \sum_{lon} |\text{reconstruction}_{lat,lon} - \text{model}_{lat,lon}| \quad (1)$$

We did this for all nine combinations of model and reconstruction maps for the top three modes variability (model_1:reconstruction_1, model_1:reconstruction_2, model_1:reconstruction_3, model_2:reconstruction_1, model_2:reconstruction_2, etc.). For reconstruction mode 1 (reconstruction_1), then, we matched which model mode best represented this spatial variability by sorting the model modes based on the smallest difference number. We did this for each reconstruction mode (excluding previously matched model modes) to sort the modes based on the smallest difference.

"After compiling scores for all five of the aforementioned scoring criteria, we normalized each set of scores to be on a scale from one to ten to ensure that

each criterion was equally weighted."
** So, if I understand well, you divide each criteria by the max of the criteria? This scaling is extremely sensitive to outliars. You should consider scaling by the interquantile range or by the standard deviation of each of you criteria.*

This is a very good point. We will do a similar scaling by the interquartile range and compare the results. Should the results differ greatly to those done in the current fashion, we will change the paper results to reflect this new method and move the results obtained using the original method into the supplementary material and note how different methodology affects the final outcome.

"To refine the scope of what we predict for SMB in the future, we used a subset of models that had a final score in the top 10th percentile of CMIP5 and compared them to the entire scope of CMIP5"
** I am not sure it is a correct method. How much is your method sensitive to the number of models you keep? Why do you use this "10th percentile" criteria? I think that 4 models is too little to compute a robust statistic. Is it statistically correct to compare 4 members to 30 members? You should consider e.g. ensemble regression based on models' scores (Bracegirdle and Stephenson, 2012, doi: 10.1007/s00382-012-1330-3)*

Thanks for this suggestion. We did some analysis on how sensitive this result is to the choice of what we define as our 'top models'. To include this point in the text, we will add words to the effect of: We ran a Monte Carlo simulation in which four random CMIP5 models were selected 100,000 times. Those 100,000 sets of four random scores were compared to the four best scoring model scores using a two-sided t-test. From this, we found that, to a 95% confidence level, we can reject the null hypothesis that the four best scoring models are not statistically significantly different from any random four CMIP5 models.

Figure 1.
** when I see the spatial pattern of trends in 1B and 1C, I wonder why you use a criteria for SMB-integrated values instead of comparing spatial maps of trends? I think using spatial maps of trends would be more relevant.*

In our analysis, we made this separation by first analyzing the AIS-integrated trends and variability, and then focus on the spatial pattern of variability, and how the trend is spatially variable, on sub-ice sheet scales using EOF techniques. As Figure 5 shows, one of the dominant modes of variability in the reconstruction is reflective of the trend shown in Figure 1B, and criteria 4 and 5 score the ability of the models to simulate that pattern.

"Looking at multiple time "slices" allows us to investigate if models capture the reconstructed SMB trends for the whole time series compared to more recent decades. Here, we looked at three time slices: the entire over-lapping time series from 1850-2000, the last century from 1900-2000, and the last 50 years from 1950-2000."
** I understand that simulating correctly the trends for 1950-2000 may be useful because it quantifies if the global climate models are able to simulate correctly the response to anthropogenic forcing. However I don't think that scoring the trends over the century is useful for your purpose. Your uncertainty on century-scale trends is very small and I wonder if it is not underestimated. It seems difficult to estimate century-scale internal variability from a 200 year reconstruction in fact.*

We appreciate this comment here regarding the long-term variability of SMB. There is a difficult balance, we feel, in selecting the correct time scale for doing this trend analysis. As the reviewer points out, the last 50 years is useful for quantifying the anthropogenic forcing, but the interannual variability over this time period makes for a very large trend uncertainty. The century-length timescale loses this forced response

aspect, but the trend uncertainty is greatly reduced as the reviewer points out. The method for assessing our trend uncertainty is outlined in the text: we performed Monte Carlo simulation wherein we assumed a normal distribution where σ = reconstruction uncertainty of possible SMB values for each year. Then we created 10,000 potential SMB time series by choosing SMB values based on that normal distribution for each year and recalculated the trend for each of these time series. Our uncertainty, then, was the standard deviation of this range of trends done in the same method as published by Medley & Thomas in 2019. Basically, while the anthropogenic signal is concentrated in the latter half of the century, the longer time slices confirm the robustness of the trends as the period length increases. The variability overwhelms the signal at shorter period lengths, which results in large uncertainty bounds. By looking at several time slices, we ensure consistency between the model and reconstruction over different intervals. It is equally important to confirm that pre-1950, the trends are relatively small.

"All CMIP5 and CMIP6 models overestimate SMB variability. The CMIP5 and CMIP6 models range from overestimates of 144% to 261% and 151% to 217% of the reconstruction standard deviation, respectively" ** A strong warning here. I have doubts on the reliability of the reconstruction for interannual variability. How does the reconstruction interannual variability compare with the reanalyses variability for the common period? I suspect that the annual accumulation signal extracted from ice cores is dampened.*

We thank the reviewer for this comment. We performed further analysis on the reconstruction interannual variability and compared it to the original reanalysis interannual variability at the 53 ice core sites. Through this analysis, we found that the reconstruction does, certainly, underrepresent interannual variability compared to the reanalysis by a factor of about 1.7. The variability in the reconstruction can only be as large as the variability in the ice core records. Thus, variability can be

heightened or dampened depending on ice core sampling. Further analysis of a synthetic reconstruction that uses the reanalysis P-E time series (rather than the ice cores) suggest that indeed our sampling has biased variability low. Thus, we will evaluate the variability in CMIP5/6 models using the reanalysis data rather than the reconstruction. While reanalyses struggle with trends and magnitudes, Medley et al. (2013) showed that they sufficiently reproduce the interannually variability with high skill.

"This dipole corresponds to variability in precipitation generated by variations in the track and strength of the Amundsen Sea Low. The Amundsen Sea Low, which represents the pole of circulation variability in Antarctica (Turner et al., 2013), is marked by high precipitation around the coast of the Antarctic Peninsula (Grieger et al., 2016)."
** All this sentence is strange. It is more a discussion than a result.*

This is a very reasonable point. We will move the discussion of the underlying causes for the patterns seen in the EOF analysis to the discussion section of the paper.

"The Amundsen Sea Low, which represents the pole of circulation variability in Antarctica"?
** What is a pole of circulation variability?*

We will change the wording here to reflect a more accurate description of the Amundsen Sea Low to the effect of: The Amundsen Sea Low, a dominant synoptic phenomenon that drives a significant amount of the circulation variability in West Antarctica and on the Antarctic Peninsula...

"The second mode of variability represents high variability in West Antarc-

tica and the Antarctic Peninsula. This could be caused by the topography in these regions which can induce large amounts of snowfall."
** I am not sure that you interpret the EOFs correctly. The spatial pattern of an EOF associated to its time series explains to a certain amount of the space-time variability, but it does not mean that where the EOF spatial pattern is high there is a high variability.*
" This could be caused by the topography in these regions which can induce large amounts of snowfall." I don't understand why?

We thank the reviewer for catching this misrepresentation of EOF analysis. The above statement is incorrect in that, high values in the EOF map do not indicate higher variability but rather how much variability that region explains. We will remove the false statements and replace them with words to the effect of: high values on the EOF map indicate regions that explain large amounts of the variability in AIS SMB. Previous work by Scott Hosking et al. (2013) and Turner et al. (2012) (among others) have shown that variability in the Amundsen Sea Low is responsible for large amounts of precipitation variability in West Antarctica and on the Antarctic Peninsula. Because this region dominates the overall AIS precipitation signal (as East Antarctica sees little snowfall by comparison), a variable Amundsen Sea Low signal, here, would explain the EOF pattern reflected in mode 2 of the reconstruction.

"By comparison, one of the better scoring models for the EOF map criterion, CMCC CM, also shows a dipole between the Antarctic Peninsula and the Ross Sea region for the top mode as well as strong variance signal around the Antarctic Peninsula for mode 2 and a quadrupolar pattern for mode 3."
** When looking at Fig. 5, EOF modes from the two climate models do not ressemble the reconstruction EOF modes, even for the best performing model (row B). Maybe showing the patterns with the same sign as for the reconstruction modes will help (multiply by -1 the climate model patterns). But still, they will remain very different.*

E.g. in row B there is no high spot near Davis for EOF 3, and there is a large dipole in WAIS. Are you sure of your computation? If yes, are you sure your analysis is relevant?
What are the biases of the best scoring models for the large scale circulation fields (e.g. sea level pressure over Southern Ocean) over the last 40 years?

To the reviewer's first point, we can multiply the model EOFs by -1 to make the comparison easier, but, generally, we think that the main point here is not that the models match perfectly with the reconstruction EOF, but rather that it's more about the general regional patterns than local phenomena. No model will perfectly recreate the the regional specifics of the EOFs, nonetheless those on a more local scale, due to the fact that no model is fully capable of perfectly recreating real world physical parameters. To the reviewer's second point, while we find this question interesting, we feel it is beyond the scope of this work which focuses on determining SMB performance based on a select set of scoring criteria related to the Antarctic Ice Sheet proper.

Fig 9 and associated text : ** The climate sensitivity for SMB must be shown in % K-1, because SMB varies exponentially with temperature. You should revise the end of section 4 with regard to climate sensitivities computed in % K-1.*
Given the issues on the scoring and the relevance of selecting four models, the new version of the manuscript might give different results.

To the first point, we will change everythig to K^{-1} for consistency. To the second point, yes, given the sensitivity of the models to the scoring criteria, changes to these criteria could easily result in a different conclusion as to the top four scoring models. We will make sure that any changes to the scoring regimes that result in changes in the top four scoring models are duly noted in the text.

Minor

"Integrated over the grounded Antarctic ice sheet (AIS), the blowing snow and runoff terms are negligibly small (Lenaerts et al., 2012a)."
** Drifting snow sublimation is still not well modeled and evaluated. You should reformulate, e.g. something like "we neglect blowing snow and runoff and estimate SMB as precipitation minus sublimation"*

We will change this sentence to "We neglect blowing snow and runoff and estimate SMB as precipitation minus sublimation."

"Over longer time scales"
** Which ones?*

We have added "Over longer ($\sim$100-1000 year) time scales..."

"The strong regional variability suggests an important impact of variations in synoptic scale patterns around the AIS (Fyke et al. (2017); Marshall et al. (2017))."
** It is known that synoptic scale patters drive the accumulation variability, reformulate, e.g. "Synoptic-scale variability induces a strong regional variability of the SMB"*

We will change the sentence to "Synoptic-scale variability induces a strong regional variability of the SMB."

"Additionally, as the atmosphere has been warming over large parts of the AIS and can hold more moisture per the Clausius-Clapeyron relation, SMB is expected to show an overall increase"
** Previdi and Polvani (2016, https://iopscience.iop.org/article/10.1088/1748-9326/11/9/094001) state that "the forced SMB increase due to global warming in recent decades is unlikely to be detectable as a result of large natural SMB variability".*

Your sentence is unclear and potentially wrong for the last decades. Modify and add references.

We thank the reviewer for catching this clunky language here. The point we were trying to make relates to future SMB rather than that of the past. We will rewrite this sentence to reflect this more accurately to the effect of: Additionally, as the atmosphere is projected to warm both globally and especially in the polar regions, the atmosphere is expected to be able to hold more moisture per the Clausius-Clapeyron relation. As such, SMB is expected to show an overall increase. In recent decades, this forced SMB response is undetectable due to the significant natural SMB variability (Previdi & Polvani (2016)). Teasing apart the forced response from natural SMB variability requires longer SMB time series – on the order of centuries. In 2017, Thomas et al. found no significant SMB trend over the last 1000 years. In 2019, however, Medley & Thomas found that, over the past 200 years, there is a statistically significant SMB increase that can be derived from ice core measurements.

"but many of those models tend to overestimate annual precipitation values due to poor representation of coastal topography"
** Are you sure it is because of the poor representation of coastal topography?*

We will add that it is likely due to poor representation of coastal topography as previous studies have shown this to be a significant factor in how precipitation is represented of the AIS. We will also add the reference: Genthon et al. (2009) doi:https://doi.org/10.3189/172756409787769681

"This allows the atmospheric moisture to penetrate too far inland and leads to excessive precipitation on much of the grounded AIS, while underestimating precipitation nearby the coasts (Lenaerts et al. (2012b))."
** I did not read again this article, but it is about "Modeling drifting snow in Antarctica*

with a regional climate model: 1. Methods and model evaluation", so I am not sure it is the right paper to cite here? Do you have other references to show that resolution is the most important factor for modelling Antarctic precipitation?

We will add references including Palerme et al. (2019) and (2017) here that better reflect recent studies of Antarctic precipitation patterns in climate models – including, specifically, CMIP5.

"Barthel et al. (2019) investigated the Ice Sheet Model Intercomparison Project version 6 to determine a recommendation of which models to use for ice sheet model forcings based on best captured current Antarctic climate relative to observations and their ability to project certain metrics into the future"
** It's "Ice Sheet Model Intercomparison Project *for CMIP6*" and not "version 6" (in fact it's version 1).*
Barthel et al. (2019) evaluate the global climat models based on their ability to capture the large scale circulation around ice sheets compared to reanalyses. It is not "very similar" to your study because the "observation" they use is well evaluated (reanalyses large scale fields after 1979) and they don't use this criteria to constrain future projections.

Addressing the first point: we will fix this typo as follows: Barthel et al. (2019) investigated the Ice Sheet Model Intercomparison Project for CMIP6 to determine... Addressing the second point: we will rephrase this sentence to the effect of: The object of this paper is similar in that Barthel et al. (2019) use scoring criteria to refine model selection specifically for ice sheet model forcing. Their work differs in that their criteria look more at the large-scale circulation patterns around ice sheets and the data set to which they compare models consists of large-scale fields reanalysis fields. Additionally, they don't then use this subselection of models to constrain future projections.

"To improve upon model estimates, several groups have combined ice core data with models to create spatio-temporally robust SMB data sets (Monaghan et al. (2006), Thomas et al. (2017), Medley and Thomas (2019))." ** this sentence should be in the Method section*

We will move this sentence to the methods section.

"In this work, we leverage the availability of that new avenue for climate model evaluation of AIS SMB, and compare the suite of CMIP5 and CMIP6 climate models to that new SMB reconstruction." ** repetition of the sentence P2 L50-52, merge the two.*

We will merge these two sentences to avoid repetition.

"they weighted each ice core spatially to generate the 200-year data set"
** give the period*

We will change this sentence to: they weighted each ice core spatially to generate the 200-year (1800-2000) data set

"they calculated spatial sampling uncertainty is based on the RMSE"
** "they calculated spatial sampling uncertainty based on the RMSE"*

We will correct this typo.

"Global climate models tend to show higher skill at representing interannual variability compared to regional climate models (Medley and Thomas, 2019)."

** it is not what is said in Medley and Thomas, 2019. They say "Because of their aforementioned ability to reproduce the interannual variability[17], which strengthens the weighting scheme, we used *global atmospheric reanalyses* over regional climate models.". So this statement is for *reanalyses* compared to RCM only, and is based on [17] Medley, B. et al. Airborne-radar and ice-core observations of annual snow accumulation over Thwaites Glacier, West Antarctica confirm the spatiotemporal variability of global and regional atmospheric models. Geophys. Res. Lett. 40, 3649–3654 (2013).*

We thank the reviewer for catching this error here. This is absolutely correct and we will remove this sentence and adjust the following sentence to stress our other main reasons for using global climate models for comparison including that we want to compare future model output to the end of the 21st century for which GCMs are necessary and that this work is meant to guide the selection of GCMs for ice sheet modelers to investigate the global impacts of changing ice sheets.

"To get a comprehensive look at how well global climate models capture SMB, we compared the suite of CMIP5 models to the reconstruction."
** and CMIP6?*

We will change this sentence to: To get a comprehensive look at how well global climate models capture SMB, we compared the suites of CMIP5 and CMIP6 models to the reconstruction.

P4 L90-95
** I am not sure the detail of conversion of kg m-2 s-1 in Gt yr-1 is useful. Just saying that it is computed on the original GCM grid is enough.*

We will remove the details regarding unit conversion for succinctness.

P4 L99-100
remove parentheses

We will remove the parentheses.

P4 L107: "the magnitude of the SMB time series"
** do you mean "the SMB mean value"? If yes it seems clearer for me to replace "magnitude" by "mean value" everywhere.*

We will change magnitude to mean value throughout the document and add a sentence explaining explicitly what is meant by "mean value."

"To achieve this goal, we analyzed trends from 1850-2000, 1900-2000, and 1850-2000." ** typo*
how do you combine the 3 periods?

There is a typo here: the last time period should be 1950-2000. Beyond that, we are not sure that we understand the reviewer's question.

"To score the time series variability, we detrended and normalized each time series to separate the SMB trend from its absolute magnitude using:"
** I don't understand "to separate the SMB trend from its absolute magnitude"*

This is a typo and will be changed to: to separate the SMB variability from its absolute magnitude... We will also add a couple of sentences clarifying what this means to the effect of: if a model should greatly underestimate the mean value, for example, the variability about that mean value will also likely be underestimated. To ensure that we are not double-counting the impact of SMB mean value, we calculate

the variability about the normalized time series.

"To do so, we performed an empirical orthogonal function (EOF) analysis"
** on annual data over 1850-2005(?)*

We will add this information to this sentence to the effect of: to do so, we performed an empirical orthogonal function (EOF) analysis on annual data over 1850-2000.

"By breaking this criterion down into two main factors, we were able to determine the models' abilities to accurately capture the modes of variability as well as how much variance each mode explained."
** what are the two main factors you are talking about?*

We will expand this sentence to include: By breaking this criterion down into two main factors, spatial variability and variance explained, ...

P5 L169 "All four of best scoring models are captured within the reconstructed uncertainty for the entire 150 year time series."
** After reading further I understood that the best scoring models are for the combination of criteria. I think you should begin your result section by presenting the best scoring models (currently presented P10 and in the Figures' legends)*

We will add a paragraph at the beginning of the results section explaining the selection of the top four scoring models and listing the models. We will try to make abundantly clear in the text that the result of the top scoring models that appear throughout Figures 2-5 were added in retroactively to show how well these models do for each criterion in comparison to the rest of the ensembles.

"We weighted all scores from the five scoring criteria equally on a scale from 1 to 10 with lower scores indicating better performance. The final score, then, is the sum of all the individual scores, which is renormalized on a scale of 1 to 10 with lower scores still indicating better performance."
** repetition of P5 L141-143*

We will remove this repetition.

" The reconstructed AIS SMB averaged from 1801-2000 shows higher SMB values around the coastal areas, particularly in the Antarctic Peninsula and West Antarctic regions (Fig. 1A)."
** This is really the most basic feature of Antarctic SMB, this sentence is not useful*

We will remove this sentence to keep the section concise.

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

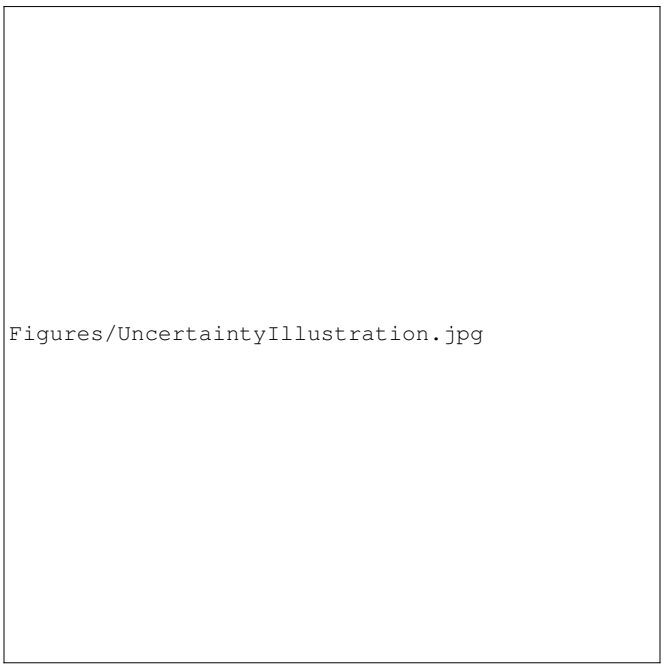

Fig. 1. Time series of the reconstructed AIS-integrated SMB time series (purple) with 1×, 21×, and 31× the uncertainty in dark purple, medium purple, and light purple, respectively. Three model AIS-integrated SMB time series, MPI ESM LR (green), IPSL CM5A LR (yellow), and BNU ESM (teal) have been plotted as well to demonstrate different model scoring. MPI ESM LR is entirely captured within 1× the reconstruction uncertainty and, thus, receives a score of 1. IPSL CM5A LR is entire captured within 2× the uncertainty so its score for this criterion is 2. BNU ESM is fully captured within 7× the uncertainty.

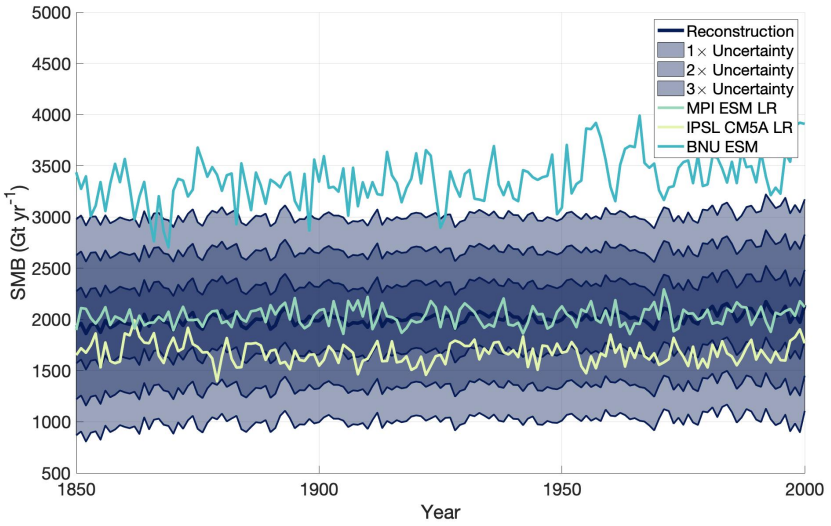
Reconstruction
1× Uncertainty
2× Uncertainty
3× Uncertainty
MPI ESM LR
IPSL CM5A LR
BNU ESM
SMB (Gt yr^{-1})
Year
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
1850
1900
1950
2000

Fig. 2.